European Journal of Education Studies
ISSN: 2501 - 1111
ISSN-L: 2501 - 1111
Available online at: www.oapub.org/edu

DOI: 10.46827/ejes.v12i4.6007 Volume 12 | Issue 4 | 2025

APPLICATION OF AI IN GRADING NEW QUESTION TYPES IN THE 2025 VIETNAMESE HIGH SCHOOL GRADUATION EXAM

Tran Quang Hieu[i],
Nguyen Thị Phuong,
Tran Thi Thu
Thai Nguyen University of Education,
Vietnam

Abstract:
The 2025 Vietnamese National High School Graduation Exam introduces a new three-format structure (multiple choice, True/False, short answer), while still relying on traditional OMR answer sheets that often lead to mechanical errors. This study proposes an AI-assisted grading system combining a direct-answer sheet format with Mathpix OCR handwriting recognition to address these limitations. Using a five-stage simplified Borg & Gall model—regulatory analysis, system development, expert validation, dual-phase trials, and pilot testing on 93 student papers—the system achieved 98.5% character recognition accuracy and high user satisfaction (Likert ≥ 4.0/5) in reducing errors and supporting all question types. Findings show that AI-enhanced direct-answer formats maintain OMR-level reliability while improving test validity. A proposed implementation roadmap includes expanding handwriting datasets, standardizing 300 dpi devices, province-wide teacher training, and integration into the official grading process.

Keywords: AI-based grading, handwriting recognition, high school examination, direct-answer format, assessment validity

1. Introduction

In 2025, Vietnam's National High School Graduation Examination will be administered for the first time under the 2018 General Education Curriculum. Earlier, in late December 2023, the Ministry of Education and Training (MOET) released an illustrative sample paper outlining the new exam format to be adopted from 2025. This sample, reflecting the revised structure, was piloted with nearly 5,000 students in five localities—Hanoi, Hai Phong, Ninh Binh, Gia Lai, and Thai Nguyen—and the resulting data are being used to build the item bank and generate official exam papers for 2025 onward. In October 2024, MOET issued 18 reference papers for the 2025 examination; the

[i] Correspondence: email hieutq@tnue.edu.vn

objective-test sections are now divided into three parts: multiple choice, True/False, and short-answer items. As for the answer sheet, MOET had already unveiled a draft template in December 2023 that, for the first time, provides dedicated areas for recording True/False selections and short answers.

However, retaining the traditional multiple-choice bubble sheet (black-and-white circles) exposes numerous shortcomings. A survey of 668 high-school students and 93 preservice teachers in Thai Nguyen revealed that 66.7 % had mistakenly shaded an answer at least once, while 9.6 % reported doing so frequently. Detailed analysis showed the two "middle" options (B and C) were most error-prone, with mis-shading of option C reaching 57.5 %. For short-answer multiple-choice items, learners complained that filling dozens of contiguous bubbles was time-consuming and prone to mistakes. A host of shading/erasing errors—unrelated to content knowledge—nevertheless directly affect scores.

Although several Vietnamese handwriting-OCR systems have been developed in Python (e.g., CRNN and other deep-learning models), they are difficult to integrate into Windows-based test environments and are not optimized for the brief letter-number combinations required by the new answer sheet. This technical gap necessitates a two-pronged solution: (1) an answer-sheet design that lets candidates write their responses directly (A–D, T/F, or 1–4, 0/1) instead of shading bubbles, and (2) AI software capable of recognizing these entries rapidly and accurately for automatic scoring. Both experts and students concur that such a solution would eliminate mechanical shading tasks and allow candidates to focus on cognitive reasoning.

Against this backdrop, the present study aims to develop an AI-based scoring system that integrates Mathpix OCR with C# to: design an answer-sheet template aligned with the 2025 exam structure and reduce recording errors; recognize concise handwritten letters and digits with an accuracy of ≥ 98 %; shorten grading time and generate instantaneous reports for proctors; and evaluate the system through three iterative trials ($10 \rightarrow 30 \rightarrow 100$ students) to refine the algorithm and demonstrate practical feasibility in schools.

This paper is expected to provide a reference model for provincial education departments and high schools during the transition to the new exam format, while offering a scientific basis for standardizing automated scoring procedures amid Vietnam's educational digital transformation.

2. Literature Review

Vietnam's National High School Graduation Examination has undergone significant changes in recent years, mirroring broader educational reforms in the country (Hien & Toai, 2024). The exam has shifted from a unified system to a "dual" model that serves both graduation and university-admission purposes (Hien & Toai, 2024). Beginning with the 2025 session, each objective-test subject will comprise three sections—multiple choice, True/False, and short answer (Ministry of Education and Training, 2024). Students will

record their answers on a bubble sheet divided into three corresponding zones, after which the sheets will be scored using Optical Mark Recognition (OMR) technology.

OMR is widely employed for the automatic grading of multiple-choice assessments and surveys (Agarwal *et al.*, 2023; Tinh & Minh, 2024). Typical systems can process hundreds of documents per hour with high accuracy, often achieving error rates below 1.5 % (Agarwal *et al.*, 2023). Nonetheless, OMR requires costly, dedicated hardware and offers limited flexibility (de Elias *et al.*, 2021). Even with technological improvements, challenges persist: detecting blank responses or multiple marks when layouts deviate from strict templates (Tabassum & Rahman, 2024); time-consuming, expensive workflows that demand trained personnel (Jain *et al.*, 2022); and sensitivity to print quality and paper thickness (Agarwal *et al.*, 2023).

Bubble-type answer sheets themselves raise additional concerns. Common scoring errors include under-scoring, over-scoring, and incomplete erasure (Muangprathub *et al.*, 2018). Using separate answer sheets—rather than writing directly on the test—heightens grading errors, especially among younger students (Muller *et al.*, 1972). Consequently, the bubble-sheet format carries various technical pitfalls, and students risk losing points for mechanical reasons unrelated to their cognitive performance.

Notably, in the short-answer section, a student must shade up to four bubbles to respond to a single question (Figure 1). This is four times the marking effort of a multiple-choice or True/False item, increasing the risk of mis-shading by a factor of four and making any correction cumbersome if the student wishes to change an answer.

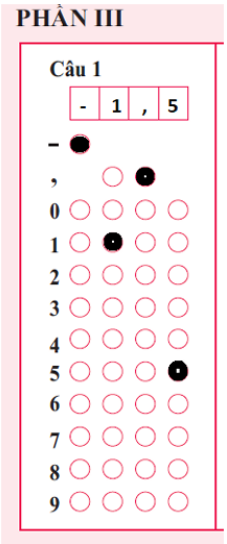


Figure 1: How to shade a short-answer question

Recent studies have concentrated on developing computer-vision systems for the automatic scoring of multiple-choice tests, aiming to reduce the time and costs associated with manual marking. These systems generally employ image-processing techniques and machine-learning algorithms to recognize answer marks on scanned or photographed sheets (Ascencio *et al.*, n.d.). A range of approaches has been explored, including OpenCV libraries (Rodrigo *et al.*, 2016), Tesseract OCR combined with YOLOv8 (Mahmud *et al.*,

2024) and convolutional neural networks (Afifi & Hussain, 2019). Nevertheless, these methods still focus on multiple-choice or True/False items and have not yet provided an effective solution to reduce student errors in short-answer multiple-choice questions.

The use of AI in exam marking is gaining traction globally, with applications in automated essay scoring, adaptive testing, and assessment analysis (Gardner *et al.*, 2021). Research shows AI can effectively grade exams, often outperforming human markers (Scarfe *et al.*, 2024). However, concerns persist regarding AI's reliability, explainability, and potential bias in high-stakes assessments (Aloisi, 2023). For handwritten math responses, GPT-4's accuracy is still too low for real-world applications (Caraeni *et al.*, 2024). Comparisons between AI and human graders suggest that while AI excels in scalability, humans are better at interpreting complex answers and evaluating creativity (Ragolane *et al.*, 2024). The effectiveness of AI grading systems often depends on well-structured rubrics and prompts (Ragolane *et al.*, 2024). Overall, these studies suggest that a hybrid model combining AI and human grading may be the most effective approach. Overall, these studies suggest that a hybrid model combining AI and human grading may be the most effective approach. Thus, if performance goals are set within AI's current capabilities, fully automated exam scoring remains entirely feasible.

In summary, while OMR has effectively served traditional multiple-choice examinations, the three-part structure of Vietnam's 2025 National High School Graduation Exam reveals clear limitations for short-answer items. Existing AI grading systems still center on essays or bubbled responses and lack an optimal solution for recognizing short handwritten character strings on answer sheets. Therefore, this study proposes a direct-write answer sheet coupled with a classroom-grade HTR-AI algorithm to eliminate shading errors, enhance accuracy, and shorten grading time - paving the way for practical implementation in Vietnamese high schools.

3. Material and Methods

3.1 Research Objectives

This study aims to develop, implement, and validate an AI-driven scoring system—combining a direct-write answer sheet template with a handwriting-recognition algorithm—to eliminate the technical errors of traditional bubble sheets and fully accommodate the three new question formats (multiple choice, True/False, and short answer) in Vietnam's 2025 National High School Graduation Exam.

3.2 Research Methodology

Using a streamlined Borg and Gall methodology (Borg & Gall, 1984) condensed into five phases: (a) analysis; (b) development; (c) validation; (d) testing; and (e) implementation.

3.3 Research Procedure

The Borg and Gall methodology, originally comprising ten stages, was streamlined into five stages. The condensed sequence is: (a) analysis; (b) development; (c) validation;

(d) testing; and (e) implementation. Reducing the framework to five steps keeps the process efficient and focused—particularly important when creating AI-based solutions—though it may limit feedback loops in the early stages. Nevertheless, the benefits of resource optimization, simplified design, and faster prototyping outweigh any drawbacks, making the five-stage framework ideal for this study.

During the analysis phase, national graduation-exam regulations and real-world student use of answer sheets in high schools were reviewed to pinpoint challenges and needs. In the development phase, software prototypes and answer-sheet templates were created to address the identified issues. The validation stage gathered expert feedback on both the template and the software, after which the system was refined accordingly. Subsequent testing involved a small cohort of students to gauge practical applicability; their recommendations further improved the software and template.

Finally, in the implementation phase, the system was piloted in two actual classrooms, where its effectiveness and practicality were assessed through accuracy rates in recognizing student responses.

4. Results

This study seeks to design a handwritten answer-sheet template and to develop AI-based software for grading that template. Both deliverables are produced through a streamlined Borg and Gall model comprising five phases: analysis, development, validation, testing, and implementation.

4.1. Analysis Phase

The analysis phase aimed to map current research trends in applying AI technologies to education, especially in assessment. A bibliometric scan revealed that AI-based exam grading remains under-represented, with only 42 Scopus-indexed publications between 2006 and 2025.

During this phase, a questionnaire survey and follow-up interviews were conducted to identify students' difficulties with traditional bubble sheets. Aggregated results indicate that shading answers is a systemic "bottleneck." Overall, 65.8 % of teachers reported that students "very often" or "sometimes" struggle with shading, whereas only 34.2 % deemed the issue rare or nonexistent—an almost 2:1 ratio confirming the problem's pervasiveness across test sessions.

Three primary obstacles emerged: correcting mis-shaded answers (65.8 %), marking the wrong bubble (57.9 %), and the time required for short-answer items (39.5 %). All three are purely mechanical and unrelated to academic ability. Correction forces students to erase or cross out the old mark before shading anew, a multi-step process prone to smudging, scanner errors, and time pressure. Mis-shading occurs when eyes and hands constantly shift between the test booklet and answer sheet; one-row misalignment renders the entire solution meaningless.

The survey also pinpointed specific error hotspots: option B and the True/False choice each accounted for 34.2 % of mistakes, followed by C (28.9 %) and D (23.7 %); option A was relatively "safe" at 18.4 %. This pattern suggests that bubbles near the matrix center or with a different format (True/False) cause spatial confusion, implying a need to redesign the grid with wider spacing and clearer symbols.

For short-answer items, the risk skyrockets: 76.3 % of teachers report that students "easily make shading errors," a rate three times higher than concerns about corrections (28.9 %) or time consumption (23.7 %). When the response is a number or symbol, students must execute a three-step conversion—identify the value, locate the corresponding bubble, and shade it precisely—so even a minor slip can generate a "phantom" answer. The nearly 50-percentage-point gap between this top difficulty and the next issues highlights the bottleneck nature of this technical error.

Overall, the quantitative data show that the primary obstacle lies in realizing the answer on the sheet, not in thinking it through. Traditional bubble sheets demand micro-mechanical precision without any built-in error-checking. Consequently, any attempt to improve multiple-choice test quality must advance on three fronts: (i) Redesigning the bubble grid with a layout that minimizes confusion and provides convenient correction zones; (ii) Training students in shading and self-checking skills through guided, repeated practice; (iii) Gradually shifting to a simpler answer-recording method to eliminate intermediate, error-prone steps.

On this basis, we propose the following solution trajectory:

Issue	**Proposed Solution**
Shading four-option questions and True/False answers is highly prone to bubble misalignment.	Students write their chosen option—A, B, C, D, Đ (Đúng = True), or S (Sai = False) directly in the blank boxes on the answer sheet; grade with AI
Shading short-answer responses is time-consuming and error-prone because many bubbles must be filled; correcting mistakes also takes considerable time.	Write answers directly; grade with AI

4.2. Development Phase

The development phase aims to propose a new answer-sheet template and an AI-based scoring program. It comprises three specific steps:

a) Drafting the answer sheet and scoring software design – identifying the core elements required to address both technical issues and real-world needs.
b) Gathering relevant data to refine the design – collecting regulatory documents that specify the 2025 exam structure and surveying current AI handwriting-recognition technologies. This step also involves selecting a programming language that best suits the chosen AI approach.
c) Building the answer-sheet template and developing the AI-powered scoring software.

Based on the official 2025 National High School Graduation Exam format (Ministry of Education and Training, 2025), we focus on five subjects—Mathematics, Physics, Chemistry, Biology, and Geography—because each contains a short-answer multiple-choice section.

For these subjects, the maximum numbers of questions are: Part I – 18 items, Part II – 4 items, and Part III – 6 items. Accordingly, we propose the following answer-sheet layout:

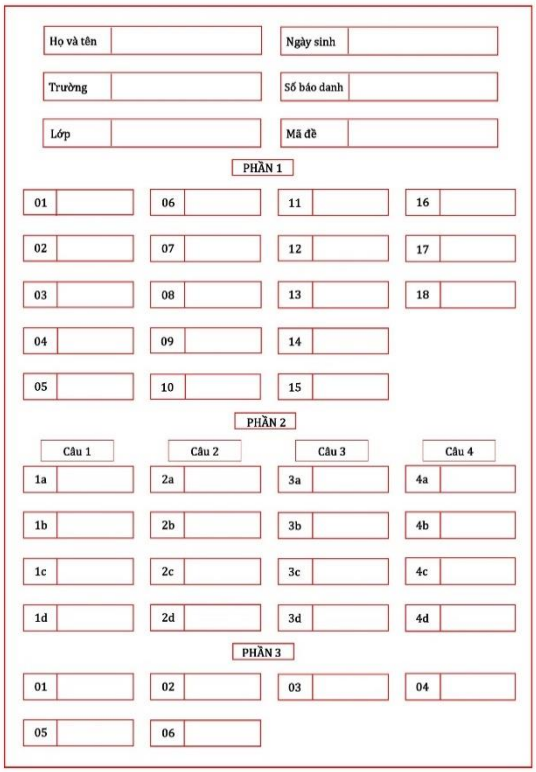


Figure 2: AI-graded answer-sheet template

4.3 Steps for Using the Software

Step 1: Data preparation

- To grade the exams, prepare the following items:
  - A folder containing the scanned exam files
  - A folder containing the answer-key files
  - A folder for the annotated images generated after grading
  - An Excel answer-key file in the following format:

+ Question type classification: mark an "x" or "X" in the corresponding column.
+ Tolerance: for short-answer items, an allowable-error option is provided.
+ Answer key: use an Excel answer-key file formatted as shown in Figure 3.

Câu	Phân loại			Đáp án	Sai số cho phép (±)
	Nhiều lựa chọn	Đúng sai	Trả lời ngắn		
1	x			A	
2	x			B	
3	x			C	
4	x			D	
5	x			A	
6	x			B	
7	x			C	
8	x			D	
9	x			A	
10	x			B	
11	x			C	
12	x			D	
13	x			A	
14	x			B	
15	x			C	
16	x			D	
17	x			A	
18	x			B	
1a		x		Đ	
1b		x		S	
1c		x		Đ	
1d		x		S	
2a		x		Đ	
2b		x		S	
2c		x		Đ	
2d		x		S	
3a		x		Đ	
3b		x		S	
3c		x		Đ	
3d		x		S	
4a		x		Đ	
4b		x		S	
4c		x		Đ	
4d		x		S	
1			x	3.7	0.2
2			x	3.7	0.2
3			x	10.8	0.2
4			x	10.9	0.2
5			x	11.9	0.2
6			x	11.9	0.2

Figure 3: Bảng mẫu đáp án Excel

Step 2: Grade the exams

- First, enter the paths of the data folders into the software.

Specify three folders:

- The folder containing the scanned exam files
- The folder containing the answer-key files
- The folder where graded images will be saved

- Next, click "CHẤM BÀI THI" as shown in Figure 4.

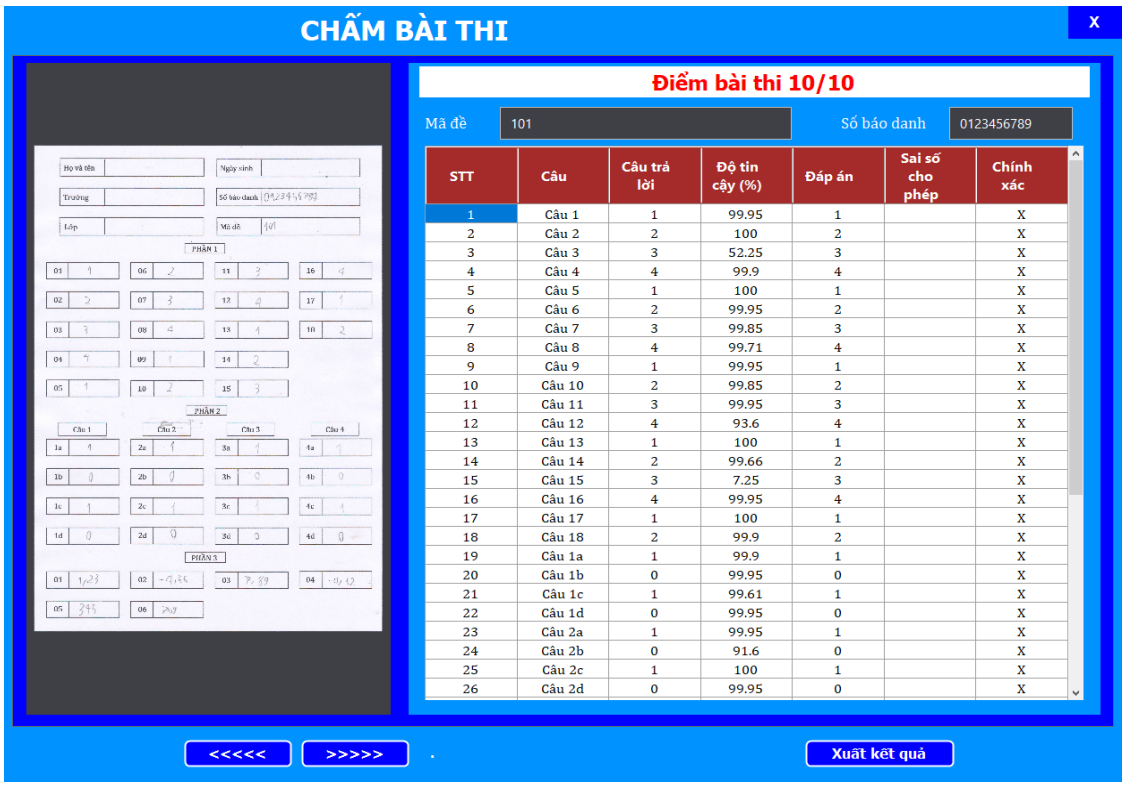
Figure 4: Information-entry and grading interface

The software will prompt you to set the correct page orientation; each option represents a specific rotation, as illustrated in Figure 5.

Figure 5: Page-rotation selection interface

The subsequent steps run automatically: the software detects and grades each candidate's paper without further input. It displays the grading results for every script, allowing exam staff to monitor progress in real time.

An example of the automatic grading output for a single exam is shown in Figure 6.



STT	Câu	Câu trả lời	Độ tin cậy (%)	Đáp án	Sai số cho phép	Chính xác
1	Câu 1	1	99.95	1		X
2	Câu 2	2	100	2		X
3	Câu 3	3	52.25	3		X
4	Câu 4	4	99.9	4		X
5	Câu 5	1	100	1		X
6	Câu 6	2	99.95	2		X
7	Câu 7	3	99.85	3		X
8	Câu 8	4	99.71	4		X
9	Câu 9	1	99.95	1		X
10	Câu 10	2	99.85	2		X
11	Câu 11	3	99.95	3		X
12	Câu 12	4	93.6	4		X
13	Câu 13	1	100	1		X
14	Câu 14	2	99.66	2		X
15	Câu 15	3	7.25	3		X
16	Câu 16	4	99.95	4		X
17	Câu 17	1	100	1		X
18	Câu 18	2	99.9	2		X
19	Câu 1a	1	99.9	1		X
20	Câu 1b	0	99.95	0		X
21	Câu 1c	1	99.61	1		X
22	Câu 1d	0	99.95	0		X
23	Câu 2a	1	99.95	1		X
24	Câu 2b	0	91.6	0		X
25	Câu 2c	1	100	1		X
26	Câu 2d	0	99.95	0		X

Figure 6: Grading-progress interface

The software displays:

- **Exam image:** with zoom controls for enlarging or shrinking the view.
- **Next/Previous buttons:** enabling staff to cycle through graded images in sequence.
- **Recognition-results panel,** featuring the columns:
  - **Candidate's answer (câu trả lời):** the response detected by the AI.
  - **Confidence score (độ tin cậy):** the AI's confidence level in the recognition result (0 % – 100 %).

Điểm bài thi 10/10

Mã đề 101 Số báo danh 0123456789

STT	Câu	Câu trả lời	Độ tin cậy (%)	Đáp án	Sai số cho phép	Chính xác
1	Câu 1	1	99.95	1		X
2	Câu 2	2	100	2		X
3	Câu 3	3	52.25	3		X
4	Câu 4	4	99.9	4		X
5	Câu 5	1	100	1		X
6	Câu 6	2	99.95	2		X
7	Câu 7	3	99.85	3		X
8	Câu 8	4	99.71	4		X
9	Câu 9	1	99.95	1		X
10	Câu 10	2	99.85	2		X

Figure 7: Grading-results interface with AI confidence scores

Step 4: Review recognition errors and make corrections

- At this stage, the software highlights any scripts containing errors—such as blank responses or invalid characters (anything other than A, B, C, D, Đ, or S).
- Relying on the actual student paper and the AI confidence score, the grader can edit the detected answer in the software and re-grade the item.

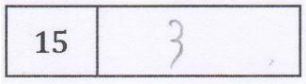
Example: Ranked by confidence, items 15 and 3b each have confidence levels below 50 %.

STT	Câu	Câu trả lời	Độ tin cậy (%)	Đáp án	Sai số cho phép	Chính xác
15	Câu 15	3	7.25	3		X
28	Câu 3b	0	8.96	0		X
3	Câu 3	3	52.25	3		X
32	Câu 4b	0	73.39	0		X
24	Câu 2b	0	91.6	0		X
12	Câu 12	4	93.6	4		X
37	Câu 3	7.89	97.03	7.89	0.2	X
35	Câu 1	1.23	98.29	1.23	0.2	X
36	Câu 2	-4.56	99.27	-4.56	0.2	X
21	Câu 1c	1	99.61	1		X

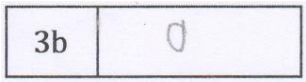
Figure 8: An example of the confidence scores for a single exam paper

The grader can zoom in on the two items above to double-check them:

- **Item 15** – AI reads **"3"**: correct.



- **Item 3b** – AI reads **"0"**: correct.



After standardising the data, the grader can edit the entries directly in the table and click Save to store the changes.

Step 5: Export exam results to Excel
When grading is finished, the grader can export all results to an Excel file for further processing.

4.3 Validation Phase

To ensure the software met the desired quality before testing and implementation, an expert-review method was employed. The consulted experts were high-school teachers from northern provinces of Vietnam—Bac Ninh, Bac Giang, Thai Nguyen, Ha Giang, Hanoi, Cao Bang, Bac Kan, and Quang Ninh. Data were gathered from 21 March 2025 to 31 March 2025, yielding 38 valid questionnaires, which fulfilled the requirements for the subsequent data-analysis stage (Table 1).

Table 1: Survey participants

Teaching experience	**Teaching experience**		
	< 5 years	**5 – 10 years**	**> 10 years**
Number of teachers	8	8	22

The survey items are presented in Table 2.

Table 2: Table of expert-survey questions on the answer-sheet template and AI-based grading software

Item No	**Evaluation Statement**	**Likert Scale (1 – 5*)**
Q1	Switching from bubble-shading to direct writing helps students reduce errors and save time while taking the test.	1 2 3 4 5
Q2	The current direct-write answer sheet accurately captures all three new question types (multiple choice, True/False, short answer) in the 2025 National Graduation Exam.	1 2 3 4 5
Q3	The layout of the response boxes (size, spacing, symbols) is easy to read and suits high-school students' handwriting habits.	1 2 3 4 5
Q4	Teachers can effectively instruct students to use the direct-write sheet in ≤ 10 minutes before the exam.	1 2 3 4 5

Q5	The fully automated AI process for scanning/photographing and grading the sheets is straightforward for teachers and exam staff.	1 2 3 4 5

4.4 Summary of Findings

- **High consensus** (mean ≥ 4.0) emerged for the two core statements:
  - *Q1* (direct writing reduces mistakes) and
  - *Q2* (the new sheet covers all three question formats). Even teachers with > 10 years' experience—typically cautious about change—rated Q1 at 4.14 and Q2 at 3.95, acknowledging the practical benefits.
- **Operational concerns** surfaced in *Q3–Q5*:
  - *Q3* averaged 3.24, indicating the box layout needs refinement, especially for handwriting clarity and student eyesight (veteran teachers = 3.05; mid-career = 3.38).
  - *Q4* and *Q5* scored 3.71 and 3.53, respectively: most teachers believe they can give quick instructions and run the software, yet 25–30 % remained "neutral/difficult."
  - A generational gap is evident: teachers with < 5 years' experience scored ≥ 4.75, whereas those with > 10 years hovered around 3.1 – 3.3, underscoring the need for training and infrastructure standardisation during large-scale rollout.

Overall, the rating matrix confirms strong pedagogical feasibility while pinpointing two priorities:

1) **Improve the answer-sheet layout** for greater readability and writing ease.
2) **Design a concise, hands-on training package** for senior teachers.

Although the template and software were deemed suitable, enhancements were made based on teacher feedback: larger response boxes to ease marking and an AI confidence-percentage display so graders can review low-certainty items. With these revisions, the system is now improved and ready for limited-scale trials.

4.5 Limited Trial Phase

The aim of the limited trial was to gather preliminary feedback on the newly validated software and answer-sheet template. The experiment involved 10 Grade-12 students at Song Cong High School, Thai Nguyen, Vietnam. All participants belonged to the natural-science track and were already familiar with traditional bubble sheets, ensuring that the trial would yield meaningful evaluations.

4.5.1 Trial Procedure

- Students filled in the direct-write sheet according to a predefined answer key. The key was balanced as follows: each option A, B, C, D appeared **five times**; each True/False choice **eight times**; and the short-answer section contained **three negative** and **three positive** values, covering all digits **0–9**.

- Students were instructed to write neatly (no careless handwriting); all responses were uppercase letters (A, B, C, D, Đ, S).

4.5.2 Trial Results

- **Recognition accuracy: 92 %**.
- **Reason:** the AI was prone to confusion between certain letter pairs, notably **D vs. Đ** and **B vs. Đ**.
- Students suggested requiring **pencil** input to ensure mistakes could be fully erased.

4.6 Deployment Phase

The deployment phase assessed the system's usability on a broader scale. Tests were conducted with **93 Grade-12 students** from classes **12A1** and **12A2** at Song Cong High School. After refining the software and workflow, the same trial procedure was applied, with two additional requirements: responses had to be written in **pencil** and students had to bring an eraser.

- **Final recognition accuracy: 98.5 %**.

5. Discussion

The study demonstrates that replacing OMR bubble sheets with a direct-write sheet coupled with an AI algorithm is highly feasible both technically and pedagogically. The Mathpix OCR-based system achieved **98.5 %** character-recognition accuracy on **93 real exam papers**, matching or surpassing the **97–98 %** accuracy of dedicated OMR scanners while eliminating the need for specialised hardware. A mean Likert score of **≥ 4.0/5** from **38 teachers** confirmed two key points: (i) the new sheet reduces mechanical errors, and (ii) it fully captures the three-question formats of the 2025 exam.

Existing YOLOv8-based scoring engines still depend on detecting filled bubbles; their whole-sheet error rate is ≈ 1 % and they do not yet support short-answer items. Our system handles all three formats in the 2025 blueprint, outperforming prototypes that recognise only True/False or single digits. To our knowledge, this is the first dataset and deployment optimised for a large-scale national exam in Vietnamese.

For students, direct writing reduces "mechanical" score losses and improves test validity. Teachers affirmed they could instruct students in **≤ 10 minutes**. Nonetheless, limitations include: data collected only in Thai Nguyen; different handwriting styles, paper quality, or lighting elsewhere may affect accuracy; performance drops with very slanted or faint handwriting; the system relies on the fee-based Mathpix API and 300 dpi images, which may be a barrier in under-resourced areas; and the teacher sample (n = 38) is not fully nationwide.

7. Conclusion

This study designed a **direct-write answer sheet** compatible with the three question formats of Vietnam's **2025 National High School Graduation Exam** and developed **AI grading software** integrating Mathpix OCR in a C# environment. Across two test rounds, the system achieved **98.5 %** character-recognition accuracy and reduced grading time relative to traditional methods. A survey of **38 teachers** showed high agreement (average Likert ≥ 4.0) on the sheet's ability to cut errors and cover all item types, and confirmed that the software workflow is accessible after brief instruction.

Three limitations were identified: (i) accuracy still depends on handwriting style and scan quality; (ii) API costs and 300 dpi hardware requirements may hinder adoption in disadvantaged regions; (iii) the trial sample was confined to Thai Nguyen and should be geographically broadened.

Based on these findings, we propose a three-stage rollout:

1) **Expand the handwriting dataset** and release open-source code.
2) **Collaborate with provincial education departments** to standardise scanners and run province-wide training for teachers and students.
3) **Conduct security and reliability reviews** and integrate the system into official grading workflows.

Future work includes adding modules for **mathematical expression recognition** and automatic scoring of **constructed-response short answers**, moving toward full automation of standardised exam assessment and advancing Vietnam's educational digital transformation.

Conflict of Interest Statement
The authors declare no conflicts of interest.

About the Author(s)
Dr. Tran Quang Hieu is a lecturer at the Faculty of Physics, Thai Nguyen University of Education, Vietnam. He holds a Ph.D. in Physics and is currently involved in teaching and conducting research in the field of physics education and related interdisciplinary approaches. His academic interests include innovative teaching methodologies, integration of technology in science education, and the development of critical thinking skills in students.
ORCID: https://orcid.org/0009-0008-5618-6517
Email: hieutq@tnue.edu.vn

References

Afifi, M., & Hussain, K. F. (2019). The achievement of higher flexibility in multiple-choice-based tests using image classification techniques. *International Journal on Document Analysis and Recognition (IJDAR)*, *22*(2), 127–142. https://doi.org/10.1007/s10032-019-00322-3

Agarwal, S., Varun, M., & Prabakeran, S. (2023). OMR reader info scanner. *2023 9th International Conference on Advanced Computing and Communication Systems (ICACCS)*, *1*, 205–209.

Aloisi, C. (2023). The future of standardised assessment: Validity and trust in algorithm s for assessment and scoring. *European Journal of Education*, *58*(1), 98–110. https://doi.org/10.1111/ejed.12542

Ascencio, H. E., Pena, C. F., Vasquez, K. R., Cardona, M., & Gutierrez, S. (n.d.). *Automatic Multiple Choice Test Grader using Computer Vision*. https://doi.org/10.1109/mhtc52069.2021.9419920

Borg, W. R., & Gall, M. D. (1984). Educational research: An introduction. *British Journal of Educational Studies*, *32*(3).

Caraeni, A., Scarlatos, A., & Lan, A. (2024). Evaluating GPT-4 at Grading Handwritten Solutions in Math Exams. *ArXiv.Org*. https://doi.org/10.48550/ARXIV.2411.05231

de Elias, E. M., Tasinaffo, P. M., & Hirata Jr, R. (2021). Optical mark recognition: Advances, difficulties, and limitations. *SN Computer Science*, *2*(5), 367.

Gardner, J., O'Leary, M., & Yuan, L. (2021). Artificial intelligence in educational assessment: 'Breakthrough? Or buncombe and ballyhoo?' *Journal of Computer Assisted Learning*, *37*(5), 1207–1216. https://doi.org/10.1111/jcal.12577

Hien, N. T., & Toai, D. B. (2024). The Study of the Development and Changes in Vietnam's University Entrance Examination System Over Forty-Five Years. *World Journal of Educational Research*, *11*(6), p.80. https://doi.org/10.22158/wjer.v11n6p80

Jain, V., Malik, S., & Bhatia, V. (2022). Robust image processing-based real-time optical mark recognition system. *2022 IEEE 6th Conference on Information and Communication Technology (CICT)*, 1–5.

Mahmud, S., Biswas, K., Alam, A., Al Mamun Rudro, R., Anannya, N. J., Mouri, I. J., & Nur, K. (2024). *Automatic Multiple Choice Question Evaluation Using Tesseract OCR and YOLOv8*. 246–252. https://doi.org/10.1109/cai59869.2024.00054

Ministry of Education and Training. (2024). *Quyết định số 764/QĐ-BGDĐT: Quy định về cấu trúc định dạng đề thi Kỳ thi tốt nghiệp THPT từ năm 2025*.

Ministry of Education and Training. (2025). *Công văn số 1239/BGDĐT-QLCL: Hướng dẫn một số nội dung tổ chức Kỳ thi tốt nghiệp THPT năm 2025*.

Muangprathub, J., Shichim, O., Jaroensuk, Y., & Kajornkasirat, S. (2018). Automatic Grading of Scanned Multiple Choice Answer Sheets. *International Journal of Engineering & Technology*, *7*(2.23), 175. https://doi.org/10.14419/ijet.v7i2.23.11910

Muller, D., Calhoun, E., & Orling, R. (1972). Test Reliability as A Function of Answer Sheet Mode. *Journal of Educational Measurement*, *9*(4), 321–324. https://doi.org/10.1111/j.1745-3984.1972.tb00964.x

Ragolane, M., Patel, S., & Salikram, P. (2024). AI Versus Human Graders: Assessing the Role of Large Language Models i n Higher Education. *Asian Journal of Education and Social Studies*, *50*(10), 244–263. https://doi.org/10.9734/ajess/2024/v50i101616

Rodrigo, T. C., F., Z., Rogério, N., & J., A. Q.-G. (2016). *An application for automatic multiple-choice test grading on Android*.

Scarfe, P., Watcham, K., Clarke, A., & Roesch, E. (2024). A real-world test of artificial intelligence infiltration of a university's examination system: A "Turing Test" case study. *PLOS ONE*, *19*(6), e0305354. https://doi.org/10.1371/journal.pone.0305354

Tabassum, K., & Rahman, Z. (2024). Optical Mark Recognition with Object Detection and Clustering. *2024 6th International Conference on Electrical Engineering and Information & Communication Technology (ICEEICT)*, 352–357.

Tinh, P. D., & Minh, T. Q. (2024). Automated Paper-based Multiple Choice Scoring Framework using Fast Object Detection Algorithm. *International Journal of Advanced Computer Science & Applications*, *15*(1).