Proceedings Paper:
Ruddle, R.A. and Naqvi, S. (2026) An Evaluation of AI-Based Grading of Multiple Choice Assessments. In: Schlippe, T., Cheng, E.C.K. and Wang, T., (eds.) Artificial Intelligence in Education Technologies: New Development and Innovative Practices. 6th International Conference on Artificial Intelligence in Education Technology, 29-31 Jul 2025, Munich, Germany. Lecture Notes on Data Engineering and Communications Technologies, 279. Springer Nature, pp. 291-307. ISBN: 978-981-95-4422-6. ISSN: 2367-4512. EISSN: 2367-4520.

https://doi.org/10.1007/978-981-95-4423-3_19

An evaluation of AI-based grading of multiple choice assessments

Roy A. Ruddle[1*] and Shabbar Naqvi[1]

[1*]School of Computer Science, University of Leeds, Woodhouse Lane, Leeds, LS2 9JT, West Yorkshire, United Kingdom.

*Corresponding author(s). E-mail(s): r.a.ruddle@leeds.ac.uk;
Contributing authors: s.naqvi@leeds.ac.uk;

Abstract

Multiple choice questions (MCQs) are widely used to assess students. Motivated by issues with accuracy and reliability that were found during university exams, we conducted a controlled user experiment with 53 participants and a commercial MCQ system that used an AI engine for grading. Each participant filled in three paper answer sheets to a prescribed pattern, one with a black pen, and the others with heavy and light pencil shading. The pattern contained 100 questions (an equal number with one, two, three, four and five correct answers). The sheets were digitized using two scanners, with each set of scans graded separately and producing a similar pattern of results. In the pen condition, the AI engine did not make any grading errors and was uncertain for 0.8% of answers (those needed to be graded by hand). However, the AI engine made grading errors for 0.25% of the heavy pencil answers and 4.9% of the light pencil answers, and was uncertain for many more answers. The results show that AI grading was only reliable when participants used a pen, which raises concerns about the guidance some organizations provide for students to use a pencil. From an explainable AI perspective, conducting rigorous user evaluations would improve transparency about AI products for end-user stakeholders, help AI developers understand the limitations of their models and identify checks and balances that should be incorporated.

Keywords: Multiple choice assessment, User evaluation, Explainable AI

1 Introduction

Multiple choice questions (MCQs) are widely used in both formative and summative assessments, requiring respondents to select one or more correct answers from a list of options. The initial concept of MCQs is credited to psychologist Edward Thorndike [1], while Frederick J. Kelly pioneered their large-scale application in 1914 with the Kansas Silent Reading Test [2].

Today's multiple choice assessments may be answered on paper or a computing device. Both have advantages, with paper-based MCQs being easy to administer in the setting of an invigilated exam and do not require specialist technology. However, marking paper MCQs by hand is time-consuming and prone to human error, and it is well-known that automated marking (e.g., using optical character recognition) is not error-free and should be accompanied by some manual checks. The authors' university is one that uses paper MCQs for some invigilated exams, and administers those exams with a widely used AI-based product (Gradescope).

The present research was motivated by issues with accuracy and reliability that were found during some such exams. The issues ranged from bubbles sometimes being missed when Gradescope indicated that it was "uncertain", to crossing out not always being detected (an 'X' was, but horizontal and diagonal lines drawn through a bubble were sometimes not) and answers being graded incorrectly. The issues were resolved by making time-consuming checks by hand, and there were some indications that issues became more prevalent with multiple answer questions (i.e., when two or more options had to be selected for an answer to be correct). That led to the present research, which had two main objectives: (1) understand how different writing implements and question parameters (the number of correct choices) affect the accuracy and reliability of Gradescope AI grading, and (2) provide guidelines for the use of Gradescope bubble sheets in exams.

A multiple choice assessment system is complex because it involves three distinct actors. First, the students who take an assessment are individuals, so there is inevitably variation in how they indicate their choices for each answer (e.g., in terms of the writing implement used and their neatness when shading bubbles). Second is the mechanism used to convert paper answer sheets into a digital form (e.g., by scanning). Third is the grading software.

We conducted a user experiment with 53 participants to investigate AI-based grading. In total, the participants answered 15,900 questions with three implements (pen, and dark and light pencil shading), the answers were converted into PDF documents with two different scanners, and each scanner's PDFs were graded separately with Gradescope's bubble sheet software. As far as we are aware, this is the largest study to date that has systematically evaluated AI-based multiple choice software. The primary contributions are: (1) establishing the circumstances under which the AI grading was, and was not, reliable, (2) providing a method that may be used to rigorously evaluate such software, and (3) using an explainable AI framework to discuss the

implications of the findings from the perspective of AI developers and end-user stakeholders.

2 Related work

This section is divided into two parts. The first provides an overview of MCQ systems and software. The second provides a background to explainable AI.

2.1 MCQ systems and software

MCQs are widely utilized to assess students and provide formative feedback during learning. Online MCQs are a preferred choice for distance learning tests, and are used in platforms like Moodle, Blackboard, Kahoot! and Socrative, as well as being adopted by certification agencies such as Coursera and Udemy [3]. A key advantage of online MCQs is that students select their answers within the MCQ software so processing is straightforward and can be fully automated. However, when such MCQs are used in a distance learning setting then it is not possible to be certain that a student's answers are wholly their own, and usage of online MCQs in an invigilated setting (e.g., an exam hall) requires the MCQ software to be firewalled so that external references, etc. may not be accessed. Paper-based MCQs do not require any such technical controls and, therefore, are straightforward to use in an invigilated setting. However, processing the paper answer sheets to identify the answers that each student selected is not trivial. That led to the scope of the present research, which was to investigate AI-based software for grading paper MCQs.

Paper MCQs are filled in by using a pen or a dark pencil [4], with the precise instructions differing from one organization to another. Some organizations that use Gradescope bubble sheets state that students may fill in the answer sheet with a "pencil or pen" [5] or a "pencil or pen with dark ink" [6]. Others specifically recommend that students use a "2B pencil" [7]. That type of pencil is slightly softer than a number 2 pencil in the USA (softer pencils generally create darker marks) [8]. However, that recommendation is contradicted by organizations which state "Use a dark pencil. You do not need to use a number 2 pencil" [9] or "Students may use any writing utensil as long as it is dark and the bubbled cells are distinct" (in answer to a frequently asked question (FAQ) titled "Do students need to use a number 2 pencil?" [10]).

Due to the processing that is involved, one cannot guarantee that every answer on a paper MCQ will be marked correctly. Gradescope's solution is to only provide a grade if the AI-assisted Grading engine is certain, and one of the above organizations states that that should be the case for more than 95% of answers on "Correctly completed papers" [7]. The implicit assumption is that if an answer was graded then that was done correctly, but the implication is that up to 5% of answers may need to be manually marked, which could be time consuming.

In addition, the onus seems to be placed on students to complete papers "correctly" [7]. Other organizations warn students to "Fill in bubbles fully"

and "Completely erase mistake answers and stray marks" [9] or "fully erase the unwanted bubble (pencil) or cross it out with an X (pen) to make it clear that it is not the answer you selected" [6].

Some research has investigated factors involved in the second component of an MCQ system – scanning or otherwise digitizing the paper answer sheets. One study that used a total of 5785 questions found that the accuracy of custom-developed OpenCV marking software improved slightly as the scanner resolution decreased (99.97% bubbles graded correctly at 200 DPI, compared with 99.94% at 300 DPI and 99.88% at 600 DPI [11]). Other research investigated the effect of lighting conditions on answer sheet acquisition with a smart phone [12], and lighting and camera resolution for smart phone [13] and Raspberry Pi systems [14].

Other research focused on the third component – AI methods themselves. One study used K-means clustering to take account of the way individual students shaded their answer sheet and improve the reading of feint and rubbed-out answers, resulting in a small but significant increase in accuracy from 98.30 to 99.98% [15]. Another study used 10,200 MCQ answers to investigate convolutional neural networks for detecting whether MCQ answers were crossed out or empty, but only achieved an average accuracy of 95.96% [16].

2.2 Explainable AI (XAI)

With the growing usage and adoption of AI, governments are considering how it should be overseen or regulated. In the United Kingdom, "appropriate transparency and explainability" is one of five principles that underpins the Government's approach to AI applications, and even when AI models are too complex to be truly explainable (the Black Box Challenge) then testing and verifying a model's output should be prioritized [17]. To make AI more transparent and explainable to a general audience, the British Computer Society (the organization that accredits computer science degree programs in the UK and some other countries) considers key measures are: (a) to identify reasonably foreseeable exceptional circumstances that may affect the operation of AI systems, and (b) evidence that organizations have properly explored and mitigated against reasonably foreseeable unintended consequences of those systems [18].

Guidance produced by the Information Commissioner's Office and Alan Turing Institute identifies six types of explanation in AI, including fairness (individuals are treated equitably), and safety and performance (accuracy, reliability and robustness) [19]. Transparency, the opposite of an AI system operating as a black box, is increasingly being demanded by end-user stakeholders [20]. Those stakeholders include lecturers/teachers using AI systems to mark assessments, students who bear the consequences of decisions made by those AI systems, and universities and examination bodies whose reputation depends on those decisions being fair and correct.

AI developers have different requirements from XAI processes. Key XAI stages for developers are understand how an AI model works, diagnosing limitations, and then refining and optimizing the model [21]. The first stage includes explaining how the model works in general (global explanation) and for specific input/output instances (local explanation), the second stage includes investigating class separation (e.g., answer graded as correct vs. wrong), decision boundaries, outliers and trustworthiness (whether a model will act as intended when facing a given input), and the third stage includes robustness, confidence and tuning parameters [19–22].

3 Experiment

Prior to the exams where the accuracy and reliability issues were found (see Introduction), one of our school's senior members of staff had attended a training session provided by Gradescope. They explained that the bubble sheet AI system used a thresholding technique to detect which bubbles a student had selected for each answer, no particular type of writing implement was required and the paper bubble sheets could be imported via normal office scanners.

We designed a user experiment to investigate the effect of writing implement and scanner on Gradescope's AI-Assisted Grading of multiple choice bubble sheet tests. Each participant filled in three bubble sheets to a prescribed pattern. For one they shaded bubbles with a black pen, and for the other two they used heavy and light pencil shading, respectively.

The completed bubble sheets were digitized using two different scanners, and each set of scans was loaded into a separate Gradescope test for grading. Inevitably, participants sometimes made mistakes copying the prescribed pattern. That was handled by manually checking each paper bubble sheet and incorporating the "wrong" answers into the software we wrote to analyze the results (for details, see Section 3.1.3).

3.1 Method

3.1.1 Participants

A total of 53 people participated. The participants were final year undergraduate (N = 5), masters (N = 40) and PhD students (N = 8). The participants took approximately 45 minutes to complete the experiment, gave informed consent and were paid £10 for their participation. The study was approved by the University Ethics Committee.

3.1.2 Materials

A 100-question bubble sheet multiple choice test was created in Gradescope. There were 20 questions where the correct answer required one bubble to be selected, 20 requiring two bubbles, 20 requiring three bubbles, 20 requiring four bubbles, and 20 requiring all five bubbles to be selected (see Figure 1). The correct bubbles (e.g., A, B, C, D or E for a one correct answer question)

varied systematically to produce test that was easy to copy but not biased to any particular letter.

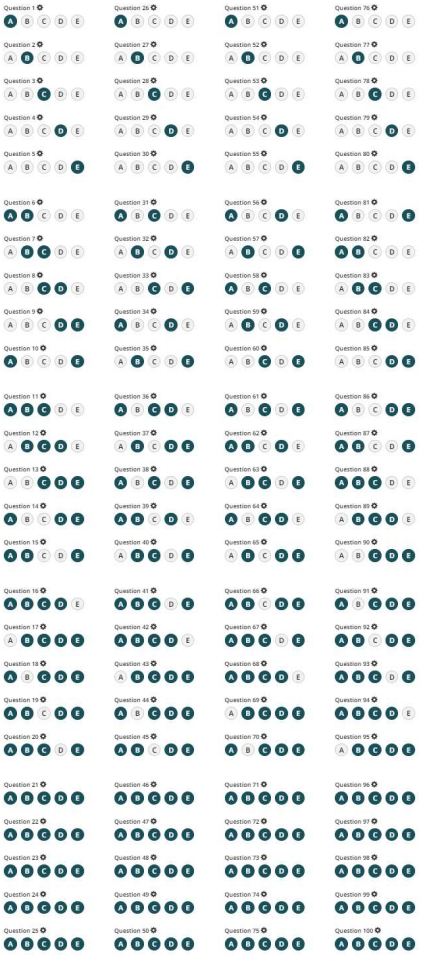

Fig. 1 The bubble sheet that was used in the experiment. There were 100 questions - 20 each where one, two, three, four and five bubbles had to be shaded for the answer to be correct (e.g., A for question 1, and ABCDE for question 100).

3.1.3 Procedure

Each participant was given one bubble sheet showing the answers and three blank bubble sheets. For one they were instructed to shade bubbles with a black pen, for another they were instructed to use heavy pencil shading so the letter inside a bubble was obscured, and for the third sheet they were instructed to shade lightly with a pencil so the letters inside bubbles remained visible.

The completed bubble sheets were scanned on two scanners (a Konica Minolta Bizhub c458 and a Bizhub c650i) using the same settings (300 dpi and color) to produce PDF documents that were loaded into separate Gradescope tests for grading. The University's scanners are all provided by the same manufacturer, but using two scanners of different models ensured that the results were not specific a single scanner.

Data analysis was performed in three steps. First, Gradescope's automatic, AI-Assisted Grading was run and then the "Download Responses" option was selected to output a comma separated values (CSV) file for each test. That CSV shows whether Gradescope was Certain or Uncertain with the marking of each answer. When it was Certain, a mark of 0 (wrong) or 1 (correct) was provided (every question was worth 1 mark in the evaluation) together with a list of the bubbles that corresponded to the correct answer, and a list of the bubbles that Gradescope thought the participant had shaded.

Second, the paper bubble sheets were manually checked by hand to identify any answers where a participant had crossed out a bubble or filled in the incorrect bubbles. Third, those "wrong" answers were incorporated into the software we wrote to analyze the CSVs, so we assessed the accuracy of Gradescope's marking not students' answers.

3.2 Results

Gradescope's AI-Assisted Grading flags each answer as certain or uncertain. The AI engine only provides a grade (0 or 1 mark, in the present study) for an answer if is certain. An answer is flagged as uncertain if the AI engine "needs more information to be more confident about what a student has selected as their answer" (e.g., a partly shaded bubble or the student changed their mind) and needs to be reviewed by a human [23].

We classified the AI engine's grading of each answer as correct, error or uncertain. A correct answer was when the AI engine correctly identified all of the bubbles that a student filled in (e.g., just A for question 1, but A and B for question 6; see Figure 1). An error was when the AI engine did not correctly identify the bubbles that were filled in. Uncertain answers were those where the AI engine did not provide a grade.

The amount of correct, error and uncertain grading with each scanner is shown in Table 1. The AI engine did not make any errors in the pen condition. However, in both the dark and light pencil shading conditions the AI engine did make errors (more with light shading). In addition, the AI engine was uncertain about more answers with than dark shading than a pen, and light shading than dark shading. Although there were differences between the two scanners (more errors and uncertain marking with the c650i scanner), the general pattern of the results was the same.

The percentage of errors increased overall with the number of correct answers in a question (see Figure 2). In absolute terms, the number of errors doubled from one to five correct answers (light shading) and from two to five correct answers (dark shading).

Table 1 The number of answers for which the AI engine was correct, made an error or was uncertain with each scanner.

	Bizhub c458 scanner		
Condition	Correct	Error	Uncertain
Pen	5257	0	43
Heavy pencil	4937	13	350
Light pencil	3239	259	1802
	Bizhub c650i scanner		
Condition	Correct	Error	Uncertain
Pen	5239	0	61
Heavy pencil	4741	19	540
Light pencil	2695	352	2253

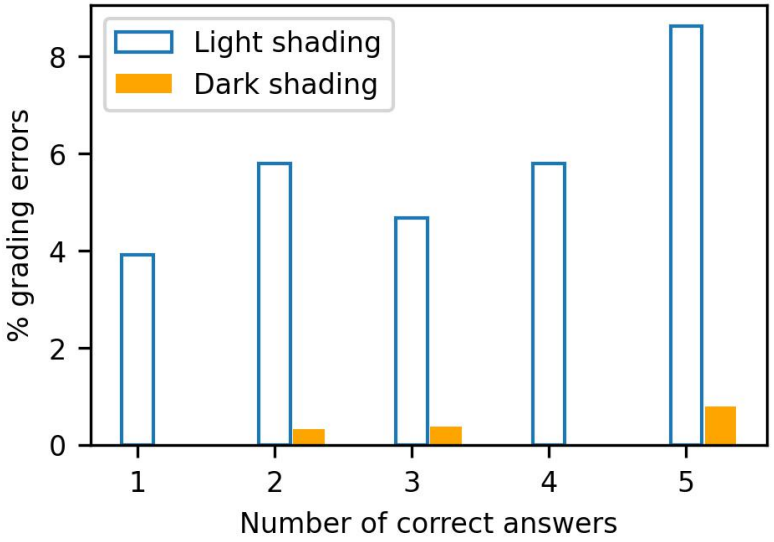

Fig. 2 The percentage of grading errors (averaged across the scanners) plotted against the number of correct answers in a question. There were no errors in the pen condition.

The percentage of answers for which the AI engine was uncertain also increased with the number of correct answers in a question (see Figure 3). The overall magnitude of that increase from one to five correct answers differed between the conditions, and was a factor of eight for light shading, 179 for dark shading and 58 for the pen condition.

The rest of this section is divided into three parts. First we report more detail about grading errors. Next we analyze answers where participants had

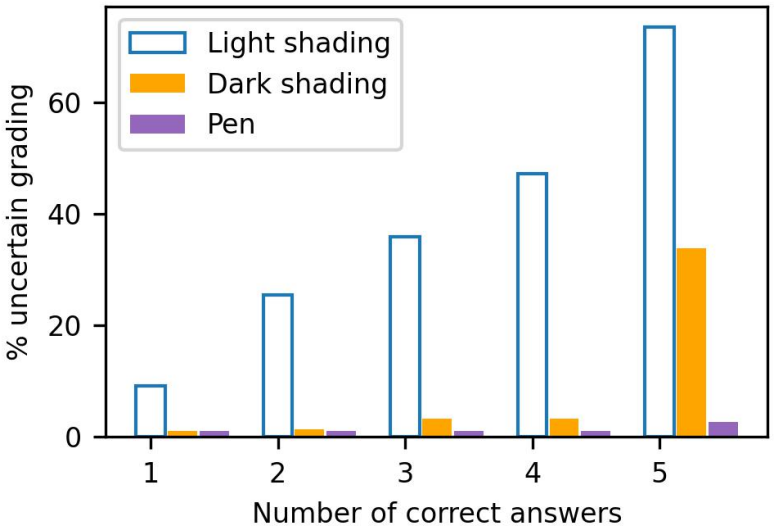

Fig. 3 The percentage of answers for which the AI engine was uncertain (averaged across the scanners) plotted against the number of correct answers in a question.

crossed or rubbed out a bubble, and then we report results about grading uncertainty.

3.2.1 Grading errors

As noted above, the AI engine made grading errors in the dark and light shading conditions. In the dark shading condition, errors occurred for 13 answers from three participants with the Bizhub c458, and for 19 answers from 10 participants with the Bizhub c650i (see Table 1). The AI engine made errors with both scanners for seven of those answers, and in each case there was no difference between the scanners for the bubbles that the AI engine thought were filled in. The errors always involved the AI engine missing bubbles that a participant had shaded, rather than including bubbles the participant had not shaded. As Table 2 shows, the AI engine missed up to three bubbles when making an error.

For every dark shading condition error the bubbles that the participant had shaded were obvious on the paper bubble sheets. Nine of the errors occurred when the AI engine thought that the participant's answer was blank (see Figure 4a). For other errors participants shaded all the bubbles similarly but in the scans some bubbles were noticeably lighter than others (see Figure 4b and Figure 4c). In the fourth example the participant's shading was quite messy, but the scanner exaggerated the difference in shading intensity and that may have caused the error (see Figure 4d). As can be seen in the examples, participants did not always shade the bubbles sufficiently heavily to obscure the letter that was inside, but errors occurred whether or not the letter was obscured.

Table 2 The number of questions for which the AI engine made marking errors when participants used dark shading, subdivided by the number of bubbles in the correct answer, the number of bubbles that the engine missed, and the scanners that were involved.

Number of bubbles in the correct answer	Number of bubbles that a participant shaded but Gradescope missed	Number of questions Only c458	Only c650i	Both scanners	Total
2	1	2	1	1	4
	2	-	-	1	1
3	1	1	-	-	1
	3	1	-	3	4
5	1	2	10	1	13
	2	-	1	1	2
Total		6	12	7	25

With one exception, the above errors affected one or two answers per participant. The exception was where the AI engine thought a participant had left several answers completely blank and had missed one or two bubbles in other answers (see Figure 5; NB: one of the blanks is the example shown in Figure 4a).

In the light shading condition, errors occurred for the majority of participants – 32 participants with both scanners, another three participants only with the Bizhub c458, and another 12 participants only with the Bizhub c650i. For each error, the bubbles that a participant shaded were compared with the bubbles that the AI engine thought were shaded. Irrespective of the number of correct answers in a question, the AI engine thought that the participant's answer was blank for 70% of the Bizhub c458 and 68% of the Bizhub c650i errors (i.e., the engine missed all 5 shaded bubbles for a 5-bubble question, all 4 shaded bubbles for a 4-bubble question, etc).

The bubble sheet that produced the most errors (89 and 94 with the Bizhub c458 and c650i, respectively) was shaded particularly lightly, making the bubbles somewhat difficult to see on the paper bubble sheet. However, on all the other bubble sheets the shaded bubbles were obvious and there were still up to 53 grading errors (see Figure 6).

3.2.2 Crossed-out or rubbed-out bubbles

As Gradescope's documentation notes, one thing that leads the AI engine to be uncertain is when a participant changes their mind about the answer to a question. Therefore, all the paper bubble sheets were inspected to identify questions where an answer had been changed by crossing or rubbing out bubbles.

In the pen condition, eight participants crossed out bubbles in a total of 11 answers. Ten of those answers were indeed flagged as uncertain by the AI engine (see Figure 7a), and further checks showed that the engine had correctly

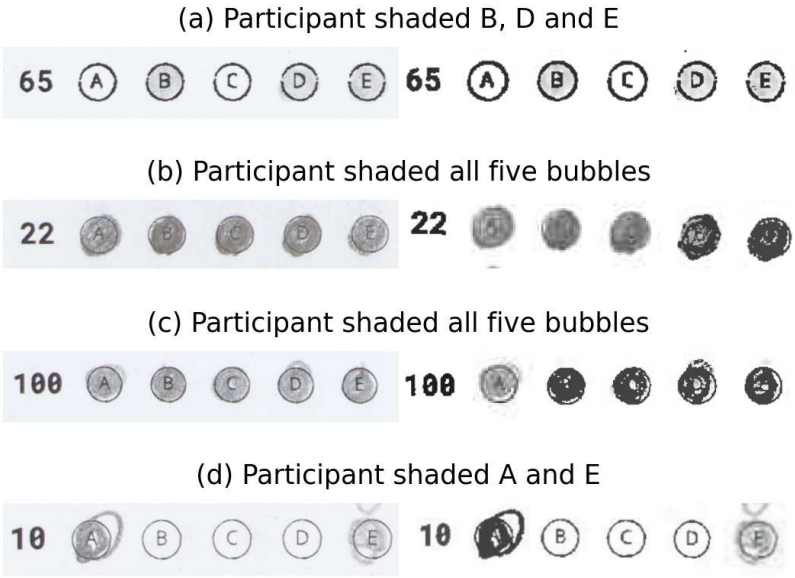

Fig. 4 Examples of grading errors in the dark shading condition from four participants. In (a) the AI engine thought that the participant had not shaded any bubbles. In (b) and (c) the AI engine thought that only BCDE were shaded. In (d) the AI engine thought that only A was shaded. The left column shows photos of participants' paper bubble sheets and the right column shows scans of those bubble sheets.

identified which bubbles were not crossed out, so in Gradescope's reviewing interface a user just had to confirm that the engine's identification was correct. By contrast, the engine flagged the 11th answer as certain (see Figure 7b) but did award the correct grade (the crossed-out bubble was ignored). The above results were not affected by the scanner.

Four participants rubbed out bubbles in a total of seven questions, and that was always in the light shading condition. The AI engine graded six of the questions correctly automatically with the Bizhub c458 scanner, and the seventh was flagged as uncertain. All seven questions were graded correctly automatically with the Bizhub c650i.

3.2.3 Uncertain grading

If the AI engine flagged an answer as uncertain then it needed to be reviewed by a human. For that reviewing, Gradescope's interface shows a processed image of the student's answer, automatically selects the bubbles that the AI engine thought were filled in, allows a user to manually correct that and then confirm the grade.

All 104 uncertain pen condition answers were reviewed using Gradescope's interface. That showed that with the Bizhub c458 scanner the AI engine always correctly identified the bubbles that participants had selected for those

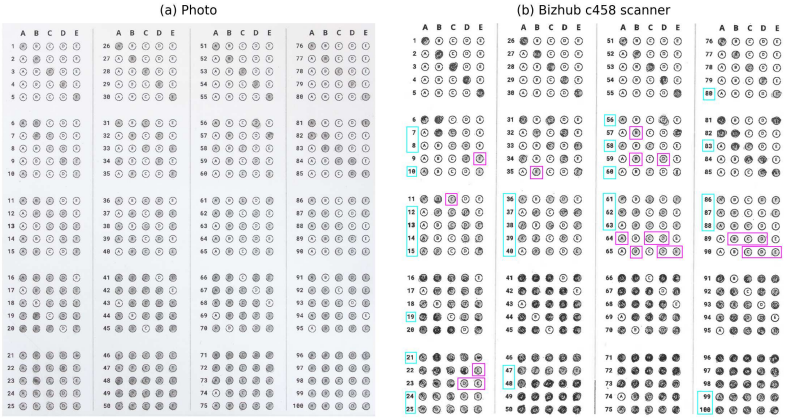

Fig. 5 The dark shading bubble sheet that produced the most grading errors, showing: (a) a photo of the bubble sheet, where is the evenness of the shading can be seen, and (b) the Bizhub c458 scan. The magenta rectangles indicate bubbles that the AI engine thought the participant had not selected and therefore made grading errors. The cyan rectangles indicate questions for which the AI engine was uncertain.

answers. However, the engine's recommendation was wrong for two of the 61 uncertain answers with the Bizhub c650i scanner (see Figure 8). Although the recommendation is just that and a user has to make the final decision about which bubbles had been selected, the user may wonder why a bubble that appears to be shaded has not been included in the recommendation, and therefore be influenced by the AI engine.

3.3 Discussion

Accuracy, reliability and robustness are key aspects of explainability for AI systems [19]. When participants used a pen in the experiment, the AI engine made no grading errors. However, the AI engine did make errors in both pencil conditions. The dark shading accuracy rate (99.7%) was high, but anything less than 100% would be a concern for end-user stakeholders such as an education organization.

Indicators of reliability and robustness would be an AI engine being capable of correctly grading all bubble sheet answers that were clear and unambiguous to a human. The occurrence of some answers where the AI engine made an incorrect recommendation for uncertain answers, even in the pen condition, is an indicator of a lack of reliability. The large amount of uncertain grading (e.g., 9% of dark shading answers) indicates that the AI engine was not robust. Although Gradescope's reviewing interface groups uncertain answer by question, it is clear that each recommendation must be individually checked by hand and the quantity of work involved reduces efficiency.

The results also demonstrate how useful experiments adopting the type of method we used can be to AI developers. On a positive note, the AI engine correctly identified the cases where participants had changed their mind (i.e.,

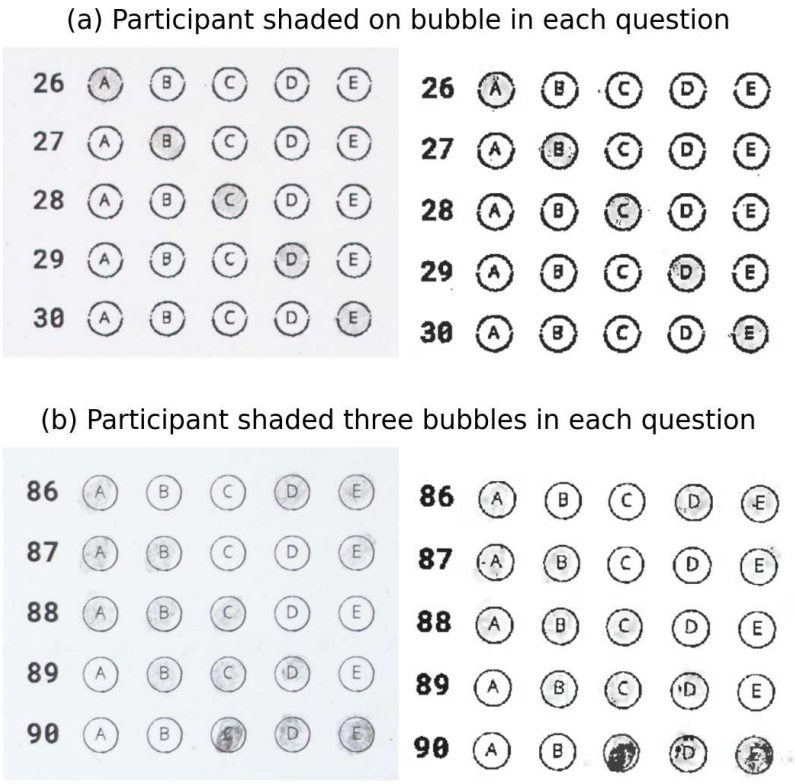

Fig. 6 Examples of grading errors in the light shading condition from two participants. In (a) the AI engine thought that the participant had left all five answers blank. In (b) the AI engine thought that the participant had left four answers blank (Q86 – 89) and was uncertain about Q90. The left column shows photos of participants' paper bubble sheets and the right column shows scans of those bubble sheets.

rubbed or crossed out bubbles), though it should be noted that the experiment was not designed to systematically investigate that. In addition, there was an interesting decision boundary [22] example which would deserve deeper investigation to understand why the AI engine was certain for one of the crossed-out answers but uncertain for all 10 of the others.

Figure 8 highlights other decision boundary examples. In Figure 8c why did the AI engine think the participant had selected bubbles A and B but not bubble E (in all three the participant's shading extends outside of the bubble perimeter), and why did the AI engine think bubble E in Figure 8b was selected but not bubble E in Figure 8d? These emphasize the benefits of testing AI systems with input from many users, to span the inherent variability in the way that people use writing implements, and following up imperfect outputs with deep investigations to improve and tune model parameters.

The grading error examples revealed ways in which all three components in the overall system (i.e., the student, scanner and AI engine) seem to interact.

Fig. 7 Examples of crossed-out bubbles: (a) flagged as uncertain by the AI engine, and (b) marked correctly automatically.

Fig. 8 Uncertain marking in the pen condition where Gradescope made: (a) wrong recommendations, and (b) correct recommendations. Each example is from a different participant.

The uneven density of bubbles in some scans (see the righthand column of Figure 4b, 4c and 4d) was most noticeable in the dark shading conditions. It was similar unevenness in the original exams that motivated the pen and pencil shading conditions in the experiment. We hypothesize that the AI engine compounds the unevenness by using a global threshold approach in its model, which causes the AI engine to decide a participant has left bubbles blank when in fact they were quite clearly selected. That provides some insights into the likely logic of the model [19] and raises questions about its suitability [22].

A way of making an AI system more transparent and explainable to a general audience is to "identify reasonably foreseeable exceptional circumstances" that affect the system's operation [18]. Most MCQs use a positive marking system (marks are gained for correct answers and there is no penalty for wrong answers). That means it is better for a student to guess a question's answer

than leave it blank. It follows that blank answers should be considered exceptional. In the experiment, the AI engine thought that 14 participants had left 1 – 94 answers blank on a bubble sheet. For 13 participants that occurred in the light shading condition, but for one participant it was the dark shading condition (5 and 4 blank answers with the Bizhub c458 and c650i, respectively; see Figure 5).

4 Conclusion

MCQ assessments have been widely used for decades. Some organizations that use Gradescope bubble sheets recommend that students use a dark pencil to fill in the assessment [7, 9], and other organizations recommend a pencil or pen [5, 6, 10]. In our evaluation there were no grading errors in the pen condition and only 0.8% of answers needed to be graded by hand. However, in a heavy pencil condition the AI engine did make grading errors (0.25% of answers; an average of 1 mark lost for every fourth student) and a significant quantity of answers needed hand-grading (6.6%).

Clear strengths of the present research were the large number of participants, the total of 15,900 MCQ answers that participants provided and the manual checks that were performed during Step 2 of the data analysis process. The greatest limitation was that only scanners from one manufacturer were used. All scanners include image processing, so Gradescope's AI engine may have worked differently with the output from other scanners.

Referring back to the contributions of this work, the first is that the AI grading was only reliable when participants used a pen to fill in the bubble sheets. A pencil was not reliable, makes it more likely that students' MCQs would be graded incorrectly and, in addition, the assessment would be less efficient from an organization's perspective because lecturers would have to spend more time grading answers by hand. Two guidelines that the authors' school now uses are: (1) students are instructed to use a black pen when answering bubble sheet MCQs, and (2) when reviewing uncertain marks, lecturers need to check each answer rather than thinking they can accept the default recommendations from Gradescope. Other policies that educational institutions could adopt from the broader implications of our findings are: (a) independently test AI systems rather than having blind trust in a commercial product's reliability, (b) define a number of marks threshold below which an MCQ answer sheet must be hand-checked (e.g., to mitigate against answers from being falsely categorized as blank), (c) perform spot checks on any MCQ answer sheets that were shaded untidily or with a pencil, and (d) record the number and type of any grading issues that were found, to provide evidence about a system's reliability and whether the above hand-checks continue to be needed.

The second contribution is defining a method that is effective for rigorously evaluating AI-based MCQ systems. The results evidence that claim and ours is the first study that we are aware of to systematically evaluate the grading accuracy and reliability of such systems.

The third contribution is using an explainable AI framework to discuss the implications of the findings. Fairness, reliability and transparency are three central tenets of XAI, from the perspective of end-user stakeholders such as universities, schools and other education organizations that use MCQ assessments. A first step in transparency would be for the companies that sell AI-based products (such as the MCQ system we evaluated) to provide factual information about the type of AI model a product uses and how it was tested, with quantifiable evidence about performance, limitations and the envelope of data inputs within which the product achieves a given level of reliability. The information could be presented in a standardized form (e.g., drawing on the Insurance Product Information Document (IPID) approach that has been mandated in the European Union and UK since 2018 for consumer insurance contracts). Such information would allow an organization to decide whether the risks associated with anything less than perfect accuracy were acceptable, and mitigations that could be adopted to reduce the risks.

Software such as Gradescope is cloud-based, which means that end users have no control over when it is updated. A consequence is grades may unexpectedly change if MCQs are remarked after a model update, so an additional onus on suppliers to be clear about which version of a model was used.

Finally, the results show why user evaluations such as the one we conducted are important for AI developers. At a global level, the large number of answers that the AI engine said were blank highlights the need for additional checks and balances – it is possible that a student may have one answer blank by accident, but many blanks were implausible. At a local level, the various examples of decision boundary issues revealed scenarios that an improved model should address. To help ensure trustworthy and reliable products, AI developers should ask themselves three questions:

1. Have you tested your system with real users (not just the development team)?
2. If the system is not 100% accurate then do you understand why and have you identified ways of resolving the issues?
3. If known issues remain in a released product then have you warned customers?

Acknowledgments. This research was supported by the Engineering and Physical Sciences Research Council (EP/X029689/1).

Data availability statement. Data supporting this study are available on a by-request basis by contacting the School of Computer Science, University of Leeds, UK.

References

[1] Medawela, R.S.H.B., Ratnayake, D.R.D.L., Abeyasinghe, W.A.M.U.L., Jayasinghe, R.D., Marambe, K.N.: Effectiveness of “fill in the blanks” over

multiple choice questions in assessing final year dental undergraduates. Educación Médica **19**(2), 72–76 (2018)

[2] Zenderland, L.: Measuring Minds: Henry Herbert Goddard and the Origins of American Intelligence Testing. Cambridge University Press, Cambridge, UK (2001)

[3] Thathsarani, H., Ariyananda, D.K., Jayakody, C., Manoharan, K., Munasinghe, A., Rathnayake, N.: How successful the online assessment techniques in distance learning have been, in contributing to academic achievements of management undergraduates? Education and Information Technologies **28**(11), 14091–14115 (2023)

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

[5] University of Maryland: Gradescope Bubble Sheets: An Alternative to Scantron. Retrieved 3 January 2025 from https://itsupport.umd.edu/itsupport?id=kb_article_view&sysparm_article=KB0017747 (2025)

[6] University of Guelf: Student tips for Gradescope bubble sheets. Retrieved 3 January 2025 from https://support.opened.uoguelph.ca/students/gradescope/bubble-sheet-tips (2025)

[7] University of Queensland: Create a Gradescope Bubble Sheet Assignment (Original). Retrieved 3 January 2025 from https://elearning.uq.edu.au/staff-guides-original/turnitin/create-gradescope-bubble-sheet-assignment-original?p=1#1 (2025)

[8] Pencils.com: What is a No. 2 Pencil? Retrieved 11 December 2024 from https://pencils.com/pages/no-2-pencil?srsltid=AfmBOoqnMPoduonYrpN5I1MWLfcbUESIsRyKT9NMhM6kadKUHSnU8PpO (2024)

[9] University of Connecticut: Filling out a Gradescope bubble sheet. Retrieved 3 January 2025 from https://kb.uconn.edu/space/TL/27132559382/Filling+out+a+Gradescope+Bubble+Sheet (2024)

[10] University of Iowa: Gradescope Bubble Sheets. Retrieved 3 January 2025 from https://teach.its.uiowa.edu/gradescope/gradescope-bubble-sheets (2025)

[11] Özcan, F.T., Eldem, A.: A new application for reading optical form with standard scanner by using image processing techniques. MANAS Journal of Engineering **11**(1), 64–73 (2023)

[12] Hafeez, Q., Aslam, W., Aziz, R., Aldehim, G.: An enhanced fault tolerance algorithm for optical mark recognition using smartphone cameras. IEEE Access (2024)

[13] Largo, L.D., Guillermo, J., Jancinal, A.R., Wata, M.: Bubble sheet multiple choice mobile checker with test grader using optical mark recognition (OMR) algorithm. In: 2022 5th International Conference on Electronics and Electrical Engineering Technology (EEET), pp. 27–33 (2022). IEEE

[14] Wata, M.G., Villaverde, J.F.: Bubble-sheet assessment checker with test grader using computer vision through raspberry Pi. In: 2023 IEEE 6th International Conference on Computer and Communication Engineering Technology (CCET), pp. 137–142 (2023). IEEE

[15] Sancar, Y., Yavuz, U., Karabey Aksakalli, I.: Personal mark density-based high-performance optical mark recognition (OMR) system using K-means clustering algorithm. Multimedia Tools and Applications, 1–33 (2024)

[16] Mondal, S., De, P., Malakar, S., Sarkar, R.: OMRNet: A lightweight deep learning model for optical mark recognition. Multimedia Tools and Applications **83**(5), 14011–14045 (2024)

[17] House of Commons Science, Innovation and Technology Committee: Governance of artificial intelligence (AI). Retrieved 11 December 2024 from https://committees.parliament.uk/publications/45145/documents/223578/default/ (2024)

[18] British Computer Society: BCS submission on the Governance of Artificial Intelligence. Retrieved 11 December 2024 from https://www.bcs.org/articles-opinion-and-research/bcs-submission-on-the-governance-of-artificial-intelligence/ (2022)

[19] Leslie, D.: Explaining decisions made with AI. Available at SSRN 4033308 (2020)

[20] Arrieta, A.B., Díaz-Rodríguez, N., Del Ser, J., Bennetot, A., Tabik, S., Barbado, A., García, S., Gil-López, S., Molina, D., Benjamins, R., *et al.*: Explainable artificial intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. Information Fusion **58**, 82–115 (2020)

[21] Spinner, T., Schlegel, U., Schäfer, H., El-Assady, M.: explAIner: A visual analytics framework for interactive and explainable machine learning. IEEE Transactions on Visualization and Computer Graphics **26**(1), 1064–1074 (2019)

[22] Chatzimparmpas, A., Martins, R.M., Jusufi, I., Kucher, K., Rossi, F.,

Kerren, A.: The state of the art in enhancing trust in machine learning models with the use of visualizations. Computer Graphics Forum **39**(3), 713–756 (2020)

[23] Gradescope: Grading a bubble sheet assignment. Retrieved 11 December 2024 from https://guides.gradescope.com/hc/en-us/articles/22065675043341-Grading-a-Bubble-Sheet-Assignment (2024)