International Journal of Evaluation and Research in Education (IJERE)
Vol. 15, No. 4, August 2026, pp. 3215∼3227
ISSN: 2252-8822, DOI: 10.11591/ijere.v15i4.38456

Image-native automated scoring of handwritten mathematical responses: reliability evidence and teacher–AI collaboration

JiEun Janet Song[1], Young-Seok Oh[2], Dong Joong Kim[3]
[1]Department of Curriculum and Instruction, Korea University, Seoul, Republic of Korea
[2]Division of Artificial Intelligence and Data Science, Korea Cyber University, Seoul, Republic of Korea
[3]Department of Mathematics Education, Korea University, Seoul, Republic of Korea

Article Info

Article history:

Received Dec 31, 2025
Revised Jun 26, 2026
Accepted Jul 6, 2026

Keywords:

Automated scoring
Handwritten mathematics
Image-native grading
Multimodal large language model
Reliability
Rubric-based scoring
Teacher–AI collaboration

ABSTRACT

This study examines the reliability of an image-native multimodal AI system for automated scoring of handwritten responses to Advanced Placement (AP) Calculus free-response items without requiring optical character recognition (OCR) preprocessing. Using inter-rater agreement indices and test–retest reliability analyses, we found substantial to almost perfect agreement between artificial intelligence (AI)-generated scores and calibrated human ratings, as well as almost perfect stability across repeated scoring sessions. These results suggest that the observed reliability of the AI scoring system warrants further investigation of validity-related evidence and inferences. As a practical implication for assessment practice, we propose a human-in-the-loop teacher–AI collaborative (TAC) framework in which automated scoring operates under teacher oversight. Taken together, these findings provide initial evidence of reliability supporting the responsible use of AI-based scoring as a measurement instrument in high-stakes educational assessment.

Corresponding Author:

Dong Joong Kim
Department of Mathematics Education, Korea University
145 Anam-ro, Seongbuk-gu, Seoul 02841, Republic of Korea
Email: dongjoongkim@korea.ac.kr

1. INTRODUCTION

A central concern in educational evaluation is whether assessment scores provide a valid and reliable basis for interpreting student performance [1], [2]. In large-scale assessment contexts, this concern becomes especially important because score interpretations often inform consequential decisions about student achievement, placement, and instructional decision making. For this reason, any scoring system—whether human or automated—must be examined not only as a scoring mechanism, but also as part of a broader measurement framework. From this perspective, the use of artificial intelligence (AI) in assessment raises a psychometric question as much as a technological one: whether AI-generated scores demonstrate sufficient reliability to support defensible interpretations of student performance.

In recent years, standardized testing environments have increasingly transitioned to digital formats. In 2024, the College Board announced an accelerated transition to digital administration of Advanced Placement (AP) exams, and by the May 2025 administration many AP exams were delivered through either fully digital or hybrid digital formats [3], [4]. AP Calculus, including both AB and BC, continues to employ a hybrid digital

administration format in which students view free-response items in Bluebook, handwrite their responses in paper booklets, and the booklets are subsequently returned for scoring [5], [6]. College Board materials note that, unlike multiple-choice sections that are computer-scored, free-response sections are scored by AP teachers and college faculty [7]. This format preserves handwritten mathematical expression within a broader digital testing environment, but leaves scoring workflows only partially aligned with the digital transition. From this perspective, image-native automated scoring is relevant not because students should enter mathematical work digitally, but because it may help to connect digital exam delivery with more scalable scoring of handwritten responses while preserving human oversight of consequential score interpretations.

Automated scoring systems have a long history of development [8], yet extending these approaches to mathematics presents distinct challenges. Traditional optical character recognition (OCR)-based pipelines have struggled to interpret the spatial and symbolic complexities of mathematical notation [9], [10], and the recognition errors they introduce may produce construct-irrelevant variance that threatens the validity of score-based inferences [11]. Recent advances in multimodal large language models (LLMs) offer a potential methodological shift [12], [13]. Models capable of processing both textual and visual inputs can bypass the OCR preprocessing step [14]. In particular, OpenAI's o1 reasoning model, with enhanced mathematical reasoning through extended chain-of-thought processing [15], offers a promising approach for image-native automated scoring of handwritten mathematical responses.

However, empirical evidence on the reliability of image-native scoring systems for mathematics assessment remains limited. Although recent studies have explored LLM-based grading of handwritten mathematics [16]–[18], most still rely on OCR as an intermediate step and have reported limited grading accuracy for handwritten science, technology, engineering, and mathematics (STEM) responses [19]. As the broader use of LLMs in education expands [20], researchers have also emphasized the importance of explainable AI (XAI) frameworks to ensure that algorithmic decisions remain transparent and interpretable [21].

To our knowledge, no previous study has systematically examined whether image-native AI scoring produces stable scores across repeated evaluations of the same response. This study makes two related contributions. First, it provides initial empirical evidence for considering image-native AI scoring in educational assessment and for framing the validity questions that follow from its use. Second, it introduces a teacher–AI collaborative (TAC) framework in which AI functions as a first-pass scoring instrument while educators retain authority over final score interpretation. Accordingly, this study examines the reliability of an image-native automated scoring system based on OpenAI's o1 multimodal reasoning model for handwritten AP Calculus free-response items and considers its implications for assessment practice.

- To what extent do AI-generated scores using direct image input demonstrate acceptable inter-rater agreement with human-assigned scores on AP Calculus AB/BC free-response items? (RQ1)
- To what extent does the image-native AI scoring system demonstrate score stability (test–retest reliability) across repeated evaluations of the same student response images? (RQ2)
- How does measurement error in AI-generated scores relative to human ratings vary across student proficiency levels? (RQ3)

The development of automated scoring systems traces back to Page's Project Essay Grade [8] in 1966, which demonstrated the potential for computer-based evaluation of student writing. Subsequent decades brought substantial advances, including operational systems such as ETS's e-rater, which combines statistical modeling with natural language processing to evaluate essay quality [22]. The criterion online writing service further extended these capabilities by integrating automated scoring with diagnostic feedback to support formative assessment [23]. More recent studies have applied deep learning approaches to automated essay scoring, achieving levels of agreement with human raters in many contexts [24], [25].

However, automated scoring is fundamentally an assessment validity issue rather than merely a technical problem. Bennett and Bejar [11] emphasized that score agreement alone is insufficient evidence of validity; the scoring process must represent the intended construct and support valid score interpretations. This perspective is particularly important for mathematics assessment, where the construct extends beyond final answers to include reasoning processes, problem-solving strategies, and mathematical communication [26].

As a prerequisite for any validity argument, reliability must first be established in evaluating automated scoring systems [1]. In educational assessment, reliability is typically examined along two dimensions. The first, inter-rater reliability, refers to the degree to which different raters—in this context,

AI and human scorers—produce consistent scores for the same response. It is commonly assessed using indices such as Cohen's κ, quadratic weighted κ (QWK), and exact match (EM) rate [24], [25]. According to the benchmarks proposed by Landis and Koch [27], κ values of .61–.80 indicate substantial agreement, whereas values above .81 indicate almost perfect agreement. The second dimension, test–retest reliability, refers to the consistency of scores when the same response is evaluated across repeated scoring trials. This dimension is particularly important in AI-based scoring because generative models may produce stochastic outputs. As emphasized in prior research on open-ended mathematics tasks [26], stable scores across repeated evaluations are essential for defensible score interpretation. Without reliability evidence, subsequent validity claims lack a sound empirical foundation.

Mathematical responses present distinct challenges compared with essay scoring [28]. Student work often combines natural language, symbolic notation, diagrams, and graphs in spatially complex arrangements that resist linear text processing [28]. Surveys of mathematical expression recognition emphasize that effective systems must address two-dimensional layout, symbol ambiguity, and structural parsing, which are not easily handled by standard text recognition alone [9].

Benchmark competitions such as competition on recognition of online handwritten mathematical expressions (CROHME) have repeatedly shown that handwritten mathematical expression recognition remains challenging despite advances in deep learning [29]. Traditional pipelines that convert images to symbolic text through OCR before scoring face compounding errors where recognition mistakes propagate into scoring decisions [10]. From an educational measurement perspective, such OCR-induced errors introduce construct-irrelevant variance—measurement noise attributable to the recognition process rather than to the student's mathematical proficiency—thereby threatening the validity of score-based inferences derived from automated scoring systems. This limitation has motivated the search for alternative approaches that can evaluate mathematical work more holistically.

Recent advances in multimodal LLMs have expanded the methodological possibilities for educational applications [12], [13]. Models such as GPT-4V can process images together with text, making it possible to analyze handwritten work without first converting it through OCR [14]. OpenAI's technical report on GPT-4 describes both the multimodal capabilities and the safety considerations associated with vision-enabled deployment [30].

The o1 series is particularly relevant in this context because the task involves more than visual recognition alone. Scoring handwritten mathematics also requires judgments about logical structure, the sufficiency of the evidence provided, and whether a solution actually satisfies the rubric [15]. Recent studies have begun to explore LLM-based grading in educational settings. Prior work has examined handwritten physics and mathematics exams [16], [18], [19], while other studies have investigated written assignments in other disciplines [17]. Related research on chain-of-thought prompting suggests that explicit reasoning traces may improve scoring transparency and consistency [31].

The emergence of multimodal and reasoning-oriented models has not eliminated the need for psychometric scrutiny, but it has made image-native scoring a more plausible object of investigation. Earlier automated scoring pipelines for handwritten work typically relied on OCR or handcrafted feature extraction. By contrast, GPT-4 and GPT-4V made it more realistic to investigate whether handwritten mathematical responses could be scored directly as images [14]. The later introduction of the o1 series further strengthened this possibility for tasks whose evaluation depends on more than surface-level recognition [15]. That technological shift does not establish validity on its own, but it does make image-native scoring a more credible object of psychometric investigation.

Rather than positioning AI as a replacement for human judgment, emerging frameworks emphasize collaborative approaches where AI assists teachers while humans retain oversight authority [20], [21]. Myint *et al.* [32] proposed an AI-instructor collaborative grading approach that incorporates instructor-refined marking schemes and emphasizes explainability and fairness. This perspective also aligns with broader concerns that trustworthy multimodal AI systems require transparency, fairness, and meaningful human oversight, particularly in consequential decision contexts [33]. Such collaborative frameworks are particularly important given ongoing concerns about AI reliability, fairness, and validity in educational assessment [34], [35]. While AI systems may provide efficiency gains, maintaining teacher involvement ensures that scoring decisions remain grounded in pedagogical expertise and can be adapted to individual student circumstances that automated systems might not fully capture.

2. RESEARCH METHOD

2.1. Research design

This study employed a quantitative reliability design to evaluate the performance of an image-native automated scoring system for handwritten mathematical responses. The design compared AI-generated scores with official College Board scores across multiple response samples and repeated grading trials. Figure 1 illustrates the overall approach, which combines a public data collection pipeline with a direct image-native scoring architecture. To establish reliability, we evaluated inter-rater agreement between AI-generated scores and official College Board scores, as well as test-retest consistency across repeated grading trials.

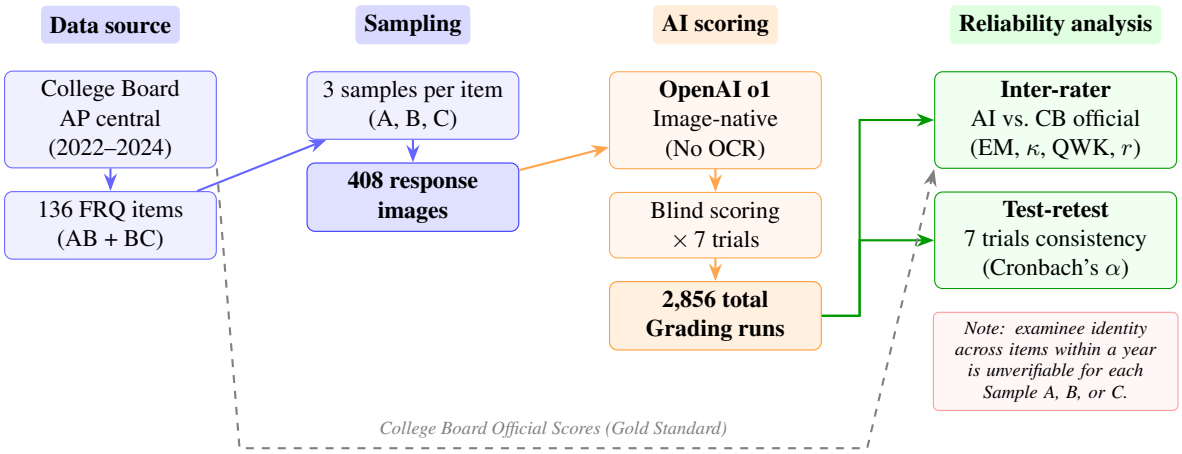

Figure 1. Overview of the image-native AI scoring design for 408 responses evaluated across seven trials

2.2. System architecture and experimental setup

Two system components were developed: an interactive user interface for individual grading sessions and an automated batch grading system for large-scale reliability evaluation.

2.2.1. Automated batch grading system

To conduct the 2,856 grading trials, we implemented an automated batch grading system that operated independently of a user interface. This system executed blind scoring, meaning that the AI grader had no access to human-assigned scores during the grading process. The design ensured that scoring decisions were based solely on rubric criteria, preventing score anchoring or confirmation bias. The system accessed a structured repository containing: i) student response images organized by examination year, course, and question number, and ii) official scoring rubrics specifying point allocation criteria for each item. Importantly, the system did not receive official worked solutions or model answers. Instead, the AI evaluated responses exclusively against the rubric criteria, mirroring the rubric-based evaluation process used by human AP readers.

For each trial, the system generated a prompt containing the response image and the corresponding rubric and submitted it to the OpenAI application programming interface (API) using the o1 model (resolving to the o1-2024-12-17 snapshot). All grading trials were conducted on March 12–13, 2025. The API returned structured outputs including: i) criterion-level scoring decisions, ii) the total score derived from those decisions, and iii) explanatory scoring commentary. All outputs were automatically parsed and exported to CSV format for subsequent statistical analysis. Figure 2 illustrates this blind scoring pipeline used in this study.

2.2.2. Scope of the present study

The present study focuses exclusively on score-based reliability metrics derived from the automated batch grading outputs. The criterion-level commentary generated by the system provides rich qualitative data for validity analysis—specifically, whether the AI's scoring rationales accurately reflect the rubric criteria and mathematical concepts. However, validity auditing of AI reasoning processes lies beyond the scope of the present study and remains a next step built on this reliability evidence.

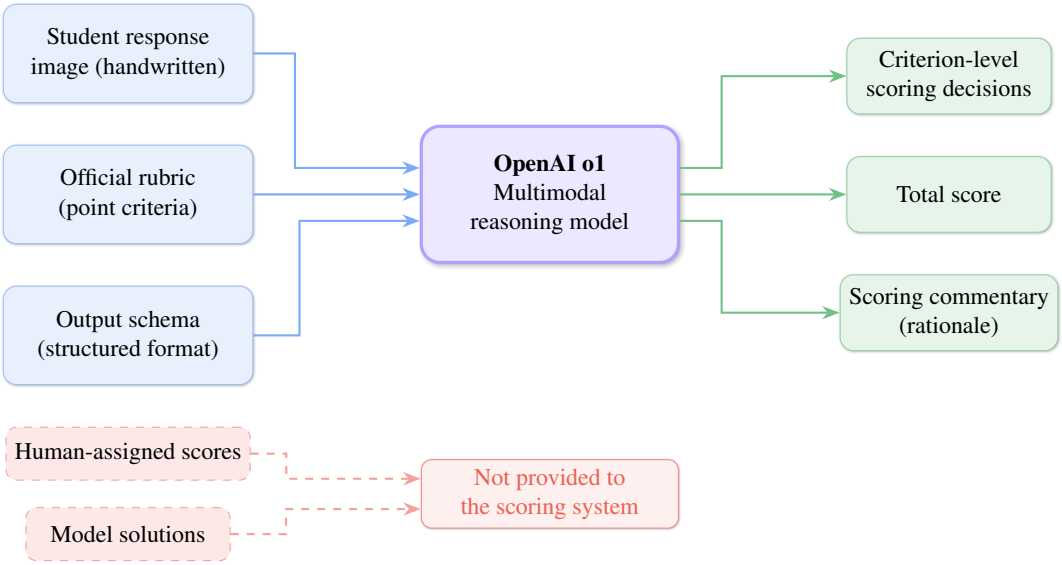

Figure 2. Blind automated scoring architecture

2.3. Prompt structure and output enforcement

A key design goal was to maximize scoring transparency and reduce stochastic variability across repeated grading. The prompt includes: i) the student's handwritten response and ii) the official rubric specifying point allocation criteria, followed by iii) a required output structure. This structure forces the model to report criterion-level decisions (e.g., whether each point is earned) and to compute the final total score from those decisions. The design was informed by the finding that prompting strategies can elicit more explicit reasoning behaviors [31], while still constraining outputs to rubric-based justifications suitable for teacher review.

2.4. Population and sampling

The target population for this study comprises student responses to AP Calculus AB/BC FRQs. The sample was drawn from publicly available materials provided on the College Board's AP central exam pages for AP Calculus AB and AP Calculus BC [36], [37]. These materials include official exam questions, scoring guidelines, sample responses, and scoring distributions for recent administrations.

2.4.1. Human reference scores

A key methodological consideration in this study is the source of human reference scores used for AI–human agreement analysis. Rather than relying on researcher-assigned scores, which could introduce investigator bias, we used the official scores published by the College Board in its annual scoring guidelines and commentary for each sample response. These scores are assigned by trained AP readers—experienced educators who undergo standardized calibration procedures to ensure consistent application of scoring rubrics across large-scale examinations. For each of the 408 sample response images, the AI-generated score was compared with the corresponding official score published for that item. This approach was adopted for two reasons. First, official College Board scores provide an externally calibrated reference standard based on the systematic training of AP readers. Second, using publicly available scoring materials enhances the transparency and reproducibility of the study while minimizing potential investigator bias that might arise if the responses were rescored by the researchers.

2.4.2. Sample selection criteria

No exclusion criteria were applied. All publicly available sample responses from the 2022–2024 AP Calculus AB and BC examinations were included in the study. Because the College Board publishes a fixed set of three sample responses per item, no additional researcher selection or filtering was required. This complete inclusion approach was adopted to maximize reproducibility and to avoid researcher-introduced selection bias, ensuring that the dataset reflects the full range of publicly available materials without subjective screening.

2.4.3. Sampling method

For each FRQ item, the College Board publishes three sample student responses in its annual scoring commentary, labeled Sample A, Sample B, and Sample C. These samples are selected to represent high-, medium-, and low-performing responses, respectively, based on the student's overall examination performance across FRQ items. Because examinee identities are not disclosed, it cannot be determined whether samples labeled A, B, or C across different items originate from the same student. Each published response was therefore treated as an independent observation in this study. We adopted a complete enumeration approach. For every FRQ item across the 2022–2024 administrations of AP Calculus AB and BC, all three published sample responses (A, B, and C) were collected. This procedure yielded the full set of publicly available sample responses without additional researcher selection, thereby ensuring reproducibility.

2.4.4. Sample size

The final dataset included 136 AP Calculus AB/BC FRQ items from the 2022–2024 examinations. With three sample responses per item (A, B, and C), this yielded 408 handwritten responses (136×3). Each response was scored seven times, resulting in 2,856 AI grading runs (408×7). Table 1 summarizes the dataset composition by year, course, and experimental runs.

Table 1. Summary of AP Calculus AB/BC FRQ dataset and experimental runs (2022–2024) [36], [37]

Year	Course	FRQ items	Samples per item	Repetitions	Total trials
2022	AB	24	3	7	504
2023	AB	24	3	7	504
2024	AB	21	3	7	441
2022	BC	23	3	7	483
2023	BC	22	3	7	462
2024	BC	22	3	7	462
Total	–	136	408	–	2,856

2.5. Student proficiency grouping

To evaluate whether grading reliability varies by student proficiency, responses were analyzed using the College Board's own sample designations. Published responses are categorized into three performance tiers—Sample A, Sample B, and Sample C—based on total examination scores, as in Table 2. In this study, Group A corresponds to Sample A responses, Group B to Sample B responses, and Group C to Sample C responses.

Table 2. Student performance group definitions based on College Board sample designations and total examination score (out of 54)

Group	Sample designation	Score range	Description
A	Sample A	High	High-performing students
B	Sample B	Medium	Medium-performing students
C	Sample C	Low	Low-performing students

2.6. Statistical analysis

Inter-rater reliability between AI-generated scores and official College Board scores was evaluated using four complementary metrics: i) EM rate, representing the proportion of responses receiving identical scores; ii) adjacent agreement rate, capturing agreement within ± 1 point; iii) Cohen's κ (chance-corrected agreement); and iv) QWK, which weights disagreements by their magnitude. The selection of these indices follows the standards for educational and psychological testing, which emphasize reliability evidence as a prerequisite for valid score interpretations in educational assessment [1]. Test-retest reliability across the seven repeated trials was assessed using the intraclass correlation coefficient (ICC), with Cronbach's alpha reported as a supplementary index of score consistency across trials. Group differences in agreement metrics were analyzed using one-way analysis of variance (ANOVA) with Tukey's honestly significant difference (HSD) post hoc tests, and chi-square tests were used to examine group differences in EM rates.

Agreement analyses were conducted at the sub-item level using the median AI score across seven trials for each of the 408 responses, while repeated-score consistency analyses used all seven trial-specific scores. Although AP FRQs vary in total point values, all analyses were performed using the official sub-item score scales defined in the College Board rubrics. AP Calculus FRQs consist of six questions (maximum of 9 points

each; total =54) with each question divided into three to four sub-items scored from 0 to 5 points. Reliability analyses were conducted at the sub-item level because total scores can obscure compensatory scoring errors, whereas sub-item analysis provides a more precise evaluation of scoring agreement in rubric-based assessment.

3. RESULTS AND DISCUSSION

3.1. Inter-rater reliability: AI–human agreement

Using the median AI score across seven trials for each of the 408 responses, the image-native AI scoring system demonstrated strong agreement with official human reference scores. As summarized in Table 3, chance-corrected agreement reached levels conventionally interpreted as substantial to almost perfect according to the Landis and Koch [27] benchmarks, indicating that AI-generated scores closely approximated expert human judgment under rubric-based scoring conditions. The confusion matrix in Figure 3, further supports this pattern, with exact agreement in the majority of responses and nearly all remaining discrepancies limited to one-point differences.

Table 3. Overall AI–human agreement and repeated-score consistency indices

Statistical metric	Value	interpretation
EM rate	83%	Exact AI-human score agreement in most responses
Adjacent agreement ($\pm$1 point)	>95%	Most discrepancies were minor
Cohen's Kappa (κ)	.77	Substantial chance-corrected agreement
QWK	.88	Strong agreement with heavier penalty for large discrepancies
Pearson correlation (r)	.88	Strong linear association between scores
Intraclass correlation (ICC(3,1))	.88	High absolute agreement
Cronbach's alpha (α)	.99	Almost perfect consistency across repeated AI grading trials

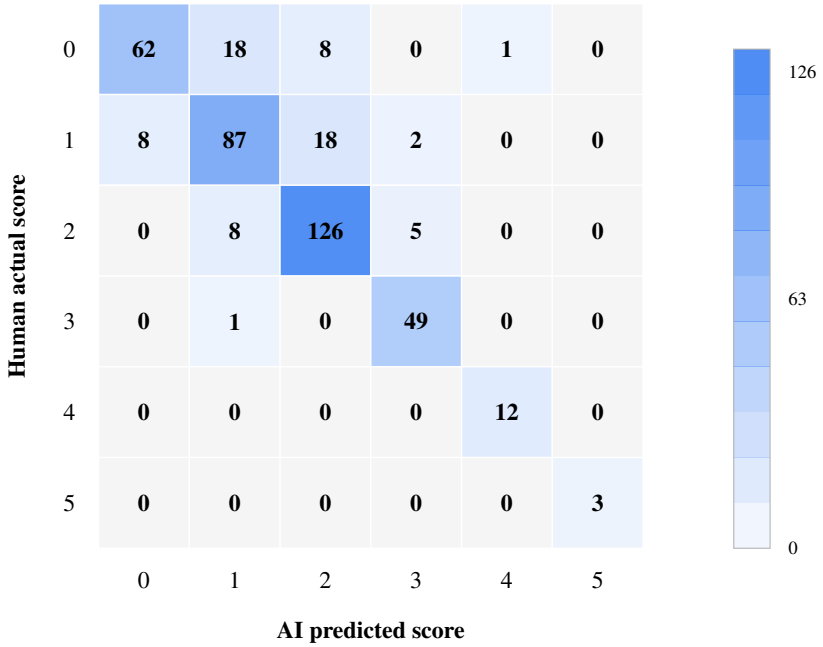

Figure 3. Confusion matrix of median AI scores versus official reference scores (0–5 scale). Diagonal cells represent exact agreement (339/408 = 83%)

3.2. Test-retest reliability: stability across repeated trials

Despite the probabilistic nature of generative language models, repeated scoring trials showed almost perfect stability, indicating minimal stochastic variability in the AI scoring process. The image-native AI scoring system therefore demonstrated strong test–retest reliability across repeated evaluations of the same responses. Table 3 summarizes the high level of agreement and score consistency observed across trials.

3.3. Performance differences by student proficiency

Although overall agreement and score stability were strong, agreement varied across student proficiency levels. Agreement metrics were therefore analyzed by proficiency tier. As shown in Figure 4, AI–human agreement differed systematically by group. Group A (high-performing) demonstrated almost perfect agreement across indices, whereas Groups B and C showed lower but still substantial agreement. Agreement declined modestly at lower proficiency levels, consistent with greater scoring ambiguity near performance thresholds.

A one-way ANOVA on absolute error scores revealed a significant effect of student proficiency group ($F(2, 405) = 17.21$, $p < .001$, $\eta^2 = .08$). Tukey HSD post-hoc tests indicated that Group A differed significantly from both Groups B and C ($p < .001$), whereas Groups B and C did not differ significantly ($p \approx .99$). A chi-square test on EM rates also confirmed significant group differences ($\chi^2(2) = 37.99$, $p < .001$). These findings suggest that the AI system performs more consistently when student responses are well-structured and aligned with expected solution approaches. Responses with incomplete reasoning, unconventional methods, or substantial errors present greater challenges for automated scoring, reflecting the inherent difficulty of interpreting ambiguous mathematical work [9], [29].

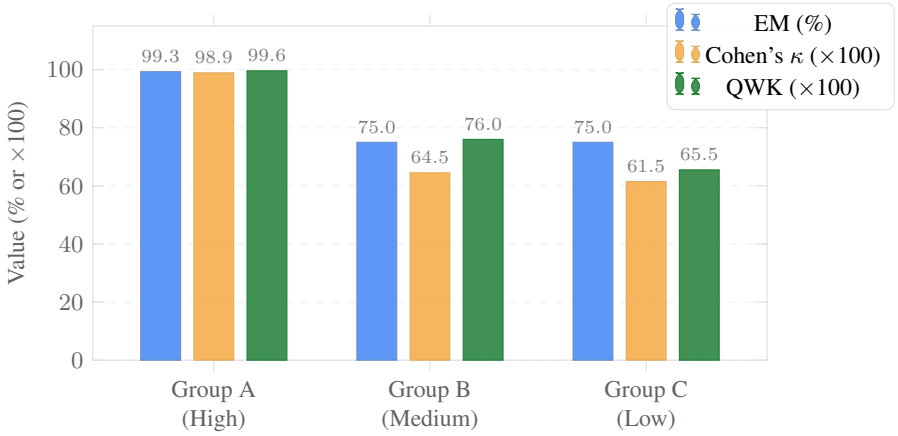

Figure 4. AI–human scoring agreement by student proficiency group

3.4. Qualitative error analysis

To identify the sources of AI–human scoring discrepancies, we conducted a qualitative analysis of responses where AI scores diverged from official College Board scores. This analysis focused on Group B (medium-performing) and Group C (low-performing) responses, where disagreement rates were highest. Two recurring error patterns emerged as the primary sources of these scoring discrepancies.

3.4.1. Outlier identification

Across all 2,856 grading trials, five cases showed large discrepancies (3–4 points) between AI and human scores in at least one trial. One of these—a Group B response that matched the official score in six of seven trials—represented an isolated stochastic deviation rather than a systematic scoring error and was excluded from further qualitative analysis. The remaining four cases involved three distinct items, as one item appeared in both the AB and BC examinations. All four occurred in low-performing responses (Group C) that received a human-assigned score of 0 but were awarded points by the AI system in multiple trials.

As shown in Table 4, Cases 1 and 2 involved the same item (2022 AP Question 1-d), which appeared in both the AB and BC examinations. Case 1 showed persistent over-scoring, with the AI awarding 4 points in 6 of 7 trials despite the official score of 0. Case 2 displayed an alternating pattern (0, 4, 0, 4, 0, 4, 0), suggesting ambiguity in the AI's interpretation. Cases 3 and 4 showed intermittent large deviations, with the AI occasionally awarding 2–3 points.

3.4.2. Pattern 1: informal cancellation marks (scribble-outs)

In Case 3 (2023 BC Question 6-c), the student's response contained extraneous markings, including scratch work that the student attempted to cancel using rough strikethroughs and zigzag cross-outs rather than

clean erasure. Human AP readers recognized these markings as cancellation indicators and excluded them from evaluation, scoring only the intended final answer as a 0. In contrast, the AI system interpreted these cross-out marks as valid mathematical content, occasionally assigning inflated scores. This pattern appeared more frequently among lower-performing students, who often revised their work without fully erasing earlier attempts.

Table 4. AI scores for outlier cases. Seven grading trials for Group C responses with official scores of 0

Case	Test	Item	1st	2nd	3rd	4th	5th	6th	7th	CB score
1	2022 AB	1-d	0	4	4	4	4	4	4	0
2	2022 BC	1-d	0	4	0	4	0	4	0	0
3	2023 BC	6-c	0	0	3	0	0	2	0	0
4	2024 BC	2-c	3	0	0	0	0	0	0	0

3.4.3. Pattern 2: notation ambiguity and non-linear spatial layout

For Cases 1, 2, and 4, qualitative analysis revealed a second error pattern involving ambiguous notation and non-linear spatial organization. These responses contained non-standard notation that human graders interpreted as incorrect but that the AI system occasionally interpreted more favorably. The two-dimensional layout appeared to hinder the AI's ability to reconstruct the logical structure of the response. Human graders therefore assigned zero points despite the presence of isolated valid symbols, whereas the AI system sometimes credited such fragments without fully evaluating their role in the overall reasoning.

3.4.4. Implications for error remediation

The identified error patterns suggest several directions for remediation. First, prompts can be revised to encourage holistic evaluation rather than crediting isolated notation fragments. In follow-up prompt tests, adding explicit holistic evaluation instructions restored correct 0-point scoring for Cases 1, 2, and 4 without reducing agreement on well-formed responses. Second, improved image preprocessing may help filter visual noise such as scribble-outs and overwritten work before scoring. These findings also suggest that some errors arise from the visual and structural characteristics of handwritten responses rather than from mathematical reasoning alone. Problematic responses often contained partially erased work, overwritten notation, or intermediate steps that remained visually salient despite not representing the intended solution. Such features appear to increase the likelihood that the model interprets residual notation as evidence for partial credit.

Future improvements in vision-language models may reduce some of these limitations, although this remains an empirical question [15], [16]. As these models become better at distinguishing intended mathematical content from visual artifacts, the need for external preprocessing may diminish. Enhanced reasoning capabilities may also improve the evaluation of incomplete or unconventional responses by helping the model distinguish responses that genuinely merit partial credit from those that merely resemble mathematically relevant notation.

Another notable finding is the directional asymmetry in AI–human score discrepancies across proficiency levels. Agreement was almost perfect for Group A, whereas Groups B and C showed a greater tendency for the AI system to assign scores above the human reference. This pattern is consistent with a possible leniency effect, particularly for responses containing incomplete reasoning, ambiguous notation, or unconventional solution strategies. However, this interpretation should be treated cautiously because score range restriction may have reduced the visibility of over-scoring among high-performing responses near the rubric ceiling. The present findings therefore do not determine whether the observed asymmetry reflects a lower-proficiency-specific issue or a broader scoring tendency masked by ceiling effects.

3.5. Interpretation of reliability evidence

The central issue in this study is whether AI-generated scores are reliable enough to support educational interpretation, not merely whether they appear plausible on the surface. In classical measurement theory, reliability is a necessary, though not sufficient, condition for validity [11]. The results indicate substantial to almost perfect agreement with expert human raters and high score stability across repeated trials, providing an empirical foundation for further validation. At the same time, these findings should be interpreted as reliability evidence rather than as proof of construct validity or operational readiness.

In this study, image-native AI scoring is considered not merely as a technical innovation but as part of the measurement process itself. In operational assessment contexts, any scoring mechanism—human or automated—must demonstrate adequate consistency, transparency, and interpretability [11]. Removing OCR preprocessing may reduce a source of construct-irrelevant variance introduced by transcription errors [9], [10], and may therefore strengthen score consistency for handwritten responses. Treating the system as a scoring instrument rather than as a productivity tool makes it possible to evaluate it using established psychometric criteria, including inter-rater reliability, score stability, and error patterns across performance levels. The reliability metrics reported here compare favorably with prior work on automated scoring of mathematical responses. Kortemeyer *et al.* [16] reported substantially lower agreement using GPT-4-based workflows for handwritten physics exam grading, whereas Liu *et al.* [18] achieved comparable agreement but relied on OCR preprocessing that introduced additional error sources. In high-stakes assessment contexts, this distinction is consequential because scores may influence placement, certification, and instructional decisions.

3.6. Implications for teacher–AI collaboration and policy

The proposed TAC framework is intended not as a detailed workflow but as a way of conceptualizing oversight in AI-assisted scoring. In this framework, AI serves as a first-pass scoring instrument, while educators retain responsibility for final score interpretation through targeted review of discrepant or low-confidence cases. This arrangement preserves professional judgment while leveraging algorithmic consistency to reduce part of the scoring workload [33]. From a measurement perspective, its value lies in treating AI as a supervised component within the scoring process rather than as a substitute for human raters. At the instructional level, the same approach may reduce teachers' workload while maintaining their authority over score review and interpretation. The key issue is therefore not whether human judgment remains involved, but how it is integrated into an AI-assisted scoring process [34], [35].

At present, TAC should be regarded as a conceptual framework rather than a tested implementation model. Questions regarding teacher acceptance, flagging thresholds, workflow efficiency, and the allocation of review responsibility remain empirical issues for future study. For now, image-native AI scoring is best viewed as a first-pass support mechanism within a human-supervised quality-control process rather than as a replacement for expert raters.

The key question is not whether AI scoring is technically feasible, but under what conditions its educational use can be justified. In classroom assessment, consistent AI-assisted scoring may help reduce grading workload while preserving defensible score interpretation through supervised frameworks such as TAC. For examination boards, image-native scoring may provide a more scalable approach for evaluating handwritten mathematical reasoning while avoiding some transcription errors associated with OCR-based workflows. At the policy level, operational use of AI scoring should depend on clear psychometric evidence. Reliability should be treated as a baseline requirement rather than an assumed consequence of technological innovation.

3.7. Limitations and future directions

Several limitations should be acknowledged. First, the study relied on publicly available AP sample responses selected for instructional purposes, which may not fully represent the diversity of handwriting styles, response patterns, and error types found in live examination settings. The absence of examinee-level information also limits inferences about within-student consistency across responses. Second, the reported reliability estimates are contingent on the stability of the underlying AI model version. Because model updates may alter scoring behavior even under the same rubric and prompting structure, future research should examine score comparability across model versions, particularly in high-stakes contexts.

Third, the almost perfect agreement observed for high-performing responses should be interpreted cautiously because score range restriction may reduce the visibility of scoring errors. In addition, treating all one-point discrepancies as equivalent may obscure differences in severity across sub-items with different point values. Finally, this study provides reliability evidence rather than full validation. Future research should therefore examine construct alignment, response-process evidence, fairness across student groups, and the consequences of score use in practice.

4. CONCLUSION

This study examined whether an image-native AI scoring system based on a multimodal reasoning model can function as a reliable measurement instrument for handwritten mathematical responses. Across

inter-rater comparisons with expert human scorers and repeated scoring trials, the system demonstrated substantial agreement and almost perfect score stability. From an educational measurement perspective, these findings suggest that image-native AI scoring may meet a meaningful reliability threshold for human-supervised use in mathematics assessment. However, these findings should be interpreted as reliability evidence rather than as full validation, and broader claims about operational use require further evidence regarding construct representation, fairness, and the consequences of score use.

ACKNOWLEDGMENTS

The authors gratefully acknowledge the College Board for publicly releasing AP Calculus AB/BC free-response question scoring guidelines and sample student responses, which made this research possible.

FUNDING INFORMATION

This research was supported by the College of Education, Korea University Grant in 2025.

AUTHOR CONTRIBUTIONS STATEMENT

This journal uses the Contributor Roles Taxonomy (CRediT) to recognize individual author contributions, reduce authorship disputes, and facilitate collaboration.

Name of Author	C	M	So	Va	Fo	I	R	D	O	E	Vi	Su	P	Fu
JiEun Janet Song	✓	✓	✓	✓	✓	✓			✓		✓			
Young-Seok Oh				✓						✓			✓	
Dong Joong Kim	✓									✓		✓	✓	✓

C	: **C**onceptualization	I	: **I**nvestigation	Vi	: **Vi**sualization
M	: **M**ethodology	R	: **R**esources	Su	: **Su**pervision
So	: **So**ftware	D	: **D**ata Curation	P	: **P**roject Administration
Va	: **Va**lidation	O	: Writing - **O**riginal Draft	Fu	: **Fu**nding Acquisition
Fo	: **Fo**rmal Analysis	E	: Writing - Review & **E**diting		

CONFLICT OF INTEREST STATEMENT

Authors state no conflict of interest.

DATA AVAILABILITY

To support the reproducibility of this study, the source code for the automated batch grading system, prompt templates, and statistical analysis scripts is publicly available in a GitHub repository at https://github.com/JiEunJanetSong/AutomatedScoring. All AP Calculus FRQ sample responses and scoring guidelines used in this research are publicly accessible through the College Board's AP Central exam pages for AP Calculus AB and BC.

REFERENCES

[1] American Educational Research Association, American Psychological Association, and National Council on Measurement in Education, *Standards for educational and psychological testing.* Washington, DC: American Educational Research Association, 2014.

[2] M. T. Kane, "Validating the interpretations and uses of test scores," *Journal of Educational Measurement,* vol. 50, no. 1, pp. 1–73, Spring 2013, doi: 10.1111/jedm.12000.

[3] College Board, "College Board transitions most AP exams to digital this May." Newsroom.collegeboard.org. Accessed Mar. 20, 2025. [Online]. Available: https://newsroom.collegeboard.org/college-board-transitions-most-ap-exams-digital-may

[4] College Board, "Accelerating our transition to digital AP exams." Allaccess.collegeboard.org. Accessed Mar. 20, 2025. [Online]. Available: https://allaccess.collegeboard.org/accelerating-our-transition-digital-ap-exams

[5] College Board, "AP Calculus AB exam." Apcentral.collegeboard.org/. Accessed Mar. 20, 2025. [Online]. Available: https://apcentral.collegeboard.org/courses/ap-calculus-ab/exam

[6] College Board, "Hybrid digital AP exams: free-response booklets overview." Apcentral.collegeboard.org/. Accessed Mar. 20, 2025. [Online]. Available: https://apcentral.collegeboard.org/media/pdf/ap-hybrid-digital-exams-free-response-booklets-overview.pdf
[7] College Board, "Score setting and scoring." Apcentral.collegeboard.org/. Accessed Mar. 20, 2025. [Online]. Available: https://apcentral.collegeboard.org/courses/how-ap-develops-courses-and-exams/score-setting-and-scoring
[8] E. B. Page, "The imminence of grading essays by computer," *Phi Delta Kappan,* vol. 47, no. 5, pp. 238–243, 1966.
[9] T.-N. Truong, C. T. Nguyen, R. Zanibbi, H. Mouchère, and M. Nakagawa, "A survey on handwritten mathematical expression recognition: the rise of encoder-decoder and GNN models," *Pattern Recognition,* vol. 153, p. 110531, Sep. 2024, doi: 10.1016/j.patcog.2024.110531.
[10] E. Z. Orji, A. Haydar, İ. Erşan, and O. O. Mwambe, "Advancing OCR accuracy in image-to-LaTeX conversion—a critical and creative exploration," *Applied Sciences,* vol. 13, no. 22, p. 12503, Nov. 2023, doi: 10.3390/app132212503.
[11] R. E. Bennett and I. I. Bejar, "Validity and automad scoring: it's not only the scoring," *Educational Measurement: Issues and Practice,* vol. 17, no. 4, pp. 9–17, Dec. 1998, doi: 10.1111/j.1745-3992.1998.tb00631.x.
[12] J. Wu, W. Gan, Z. Chen, S. Wan, and P. S. Yu, "Multimodal large language models: a survey," in *2023 IEEE International Conference on Big Data (BigData),* Dec. 2023, pp. 2247–2256, doi: 10.1109/BigData59044.2023.10386743.
[13] S. Küchemann *et al.*, "On opportunities and challenges of large multimodal foundation models in education," *npj Science of Learning,* vol. 10, no. 1, p. 11, Feb. 2025, doi: 10.1038/s41539-025-00301-w.
[14] Y. Wu *et al.*, "An early evaluation of GPT-4V(ision)," *arXiv:*2310.16534, Oct. 2023.
[15] OpenAI, "OpenAI o1 system card," *arXiv:*2412.16720, Dec. 2024.
[16] G. Kortemeyer, J. Nöhl, and D. Onishchuk, "Grading assistance for a handwritten thermodynamics exam using artificial intelligence: an exploratory study," *Physical Review Physics Education Research,* vol. 20, no. 2, p. 020144, Nov. 2024, doi: 10.1103/PhysRevPhysEducRes.20.020144.
[17] P. G. Poličar, M. Špendl, T. Curk, and B. Zupan, "Automated assignment grading with large language models: insights from a bioinformatics course," *Bioinformatics,* vol. 41, Suppl. 1, pp. i21–i29, Jul. 2025, doi: 10.1093/bioinformatics/btaf196.
[18] T. Liu, J. Chatain, L. Kobel-Keller, G. Kortemeyer, T. Willwacher, and M. Sachan, "AI-assisted automated short answer grading of handwritten university level mathematics exams," *arXiv:*2408.11728, Aug. 2024.
[19] R. Patil, A. A. Kulkarni, R. Ghatage, S. Endait, G. Kale, and R. Joshi, "Automated assessment of multimodal answer sheets in the STEM domain," *arXiv:*2409.15749, Sep. 2024.
[20] E. Kasneci *et al.*, "ChatGPT for good? On opportunities and challenges of large language models for education," *Learning and Individual Differences,* vol. 103, p. 102274, Apr. 2023, doi: 10.1016/j.lindif.2023.102274.
[21] H. Khosravi *et al.*, "Explainable artificial intelligence in education," *Computers and Education: Artificial Intelligence,* vol. 3, p. 100074, Dec. 2022, doi: 10.1016/j.caeai.2022.100074.
[22] Y. Attali and J. Burstein, "Automated essay scoring with e-rater® V.2.0," *ETS Research Report Series,* vol. 2004, no. 2, p. i–21, 2004, doi: 10.1002/j.2333-8504.2004.tb01972.x.
[23] J. Burstein, M. Chodorow, and C. Leacock, "Automated essay evaluation: the criterion online writing service," *AI Magazine,* vol. 25, no. 3, pp. 27–36, 2004, doi: 10.1609/aimag.v25i3.1774.
[24] Z. Ke and V. Ng, "Automated essay scoring: a survey of the state of the art," in *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence,* Aug. 2019, pp. 6300–6308, doi: 10.24963/ijcai.2019/879.
[25] D. Ramesh and S. Sanampudi, "An automated essay scoring systems: a systematic literature review," *Artificial Intelligence Review,* vol. 55, no. 3, pp. 2495–2527, Mar. 2022, doi: 10.1007/s10462-021-10068-2.
[26] A. S. Lan, D. Vats, A. E. Waters, and R. G. Baraniuk, "Mathematical language processing: automatic grading and feedback for open response mathematical questions," in *Proceedings of the Second (2015) ACM Conference on Learning @ Scale,* Mar. 2015, pp. 167–176, doi: 10.1145/2724660.2724664.
[27] J. R. Landis and G. G. Koch, "The measurement of observer agreement for categorical data," *Biometrics,* vol. 33, no. 1, pp. 159–174, Mar. 1977, doi: 10.2307/2529310.
[28] S. Baral, A. Botelho, A. Santhanam, A. Gurung, L. Cheng, and N. Heffernan, "Auto-scoring student responses with images in mathematics," in *The Proceedings of the 16th International Conference on Educational Data Mining,* Jul. 2023, pp. 362–369.
[29] H. Mouchère, C. Viard-Gaudin, R. Zanibbi, and U. Garain, "ICFHR2016 CROHME: competition on recognition of online handwritten mathematical expressions," in *2016 15th International Conference on Frontiers in Handwriting Recognition (ICFHR),* Oct. 2016, pp. 607–612, doi: 10.1109/ICFHR.2016.0116.
[30] OpenAI, "GPT-4 technical report," *arXiv:*2303.08774, Mar. 2023.
[31] J. Wei *et al.*, "Chain-of-thought prompting elicits reasoning in large language models," in *Advances in Neural Information Processing Systems,* vol. 35, pp. 24824–24837, Dec. 2022.
[32] P. Y. W. Myint, S. L. Lo, and Y. Zhang, "Harnessing the power of AI-instructor collaborative grading approach: topic-based effective grading for semi open-ended multipart questions," *Computers and Education: Artificial Intelligence,* vol. 7, p. 100339, 2024, doi: 10.1016/j.caeai.2024.100339.
[33] M. Saleh and A. Tabatabaei, "Building trustworthy multimodal AI: a review of fairness, transparency, and ethics in vision-language tasks," *International Journal of Web Research,* vol. 8, no. 2, pp. 11–24, 2025, doi: 10.22133/ijwr.2025.503147.1264.
[34] M. Yiğiter and E. Boduroğlu, "Examining the performance of artificial intelligence in scoring students' handwritten responses to open-ended items," *Education and Science,* vol. 50, pp. 1–18, Mar. 2025, doi: 10.15390/EB.2025.14119.
[35] N.-J. Schaller, Y. Ding, A. Horbach, J. Meyer, and T. Jansen, "Fairness in automated essay scoring: a comparative analysis of algorithms on German learner essays from secondary education," in *Proceedings of the 19th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2024),* Jun. 2024, pp. 210–221.
[36] College Board, "AP Calculus AB exam questions." *AP Central*. Accessed Mar. 20, 2025. [Online]. Available: https://apcentral.collegeboard.org/courses/ap-calculus-ab/exam/past-exam-questions
[37] College Board, "AP Calculus BC exam questions." *AP Central*. Accessed Mar. 20, 2025. [Online]. Available: https://apcentral.collegeboard.org/courses/ap-calculus-bc/exam/past-exam-questions

BIOGRAPHIES OF AUTHORS

JiEun Janet Song is a master's candidate in Curriculum and Instruction at Korea University. Her research focuses on mathematics learning and discourse in emerging AI-mediated environments, informed by her experience teaching mathematics across IB, AP, A-level, and undergraduate calculus contexts. Using a commognitive framework and design-based research methods, she investigates student–AI mathematical discourse in higher education and explores pedagogically grounded design considerations for AI-supported mathematical modeling environments. She received the ICAT Best Presentation Award in 2025 and the ICEAS Best Presentation Award in 2026. She can be contacted at email: janetsong@korea.ac.kr.

Young-Seok Oh received the Ph.D. degree in Mathematics Education from Korea University. He is an Adjunct Professor in the Division of Artificial Intelligence and Data Science at Korea Cyber University, and a Lecturer in the Department of Mathematics Education at both Korea University and Seowon University. His primary research interests include mathematics for AI, AI-based mathematics education, and the convergence of mathematics education with agentic AI. He has published and presented on topics such as mathematical connections and AI-driven mathematical discourse, and has participated in national research projects related to AI mathematics textbooks and digital education innovation. He can be contacted at email: youngseokoh@koreacu.ac.kr.

Dong Joong Kim received his bachelor's degree in Mathematics Education from Korea University, his master's degree in Mathematics from the University of Georgia, and his Ph.D. in Mathematics Education from Michigan State University. He is a Professor in the Department of Mathematics Education at Korea University. His research interests include discourse analysis and the use of technology to promote student engagement in mathematics learning. His work explores connections between mathematics learning, teaching, and discourse. He can be contacted at email: dongjoongkim@korea.ac.kr.