Feedback on Utilizing Deep Learning AI to Analyze Scientific Models

Tingting Li[a,b], Leonora Kaldaras[c] , Kevin Haudek[a], Joe Krajcik[a]
a. CREATE for STEM Institute, Michigan State University
b. Washington State University
c. Texas Tech University

Abstract

Artificial intelligence (AI) is increasingly used to assist in educational assessment, yet discrepancies between AI scores and human scores can arise. This study investigates the alignment between a convolutional neural network (CNN) model and human raters in scoring student-constructed scientific models, with a focus on how human oversight can affect differences. We analyzed 1,151 student models from a high school curriculum, which was evaluated on a rubric with multiple categories and compared initial human scores to AI-predicted scores. Cases of AI-human score mismatches were flagged for expert human re-evaluation. Quantitative analyses assessed the level of AI-human agreement and the consistency of human scoring upon review, while qualitative analyses examined the persistently discrepant cases to uncover underlying issues. Results show that human scoring was not perfectly stable—some scores changed upon review—and overall AI-human agreement was high for many categories but varied where rubric criteria were ambiguous. Notably, certain rubric categories consistently showed discrepancies due to unclear scoring criteria or AI misinterpretations of student work. These findings demonstrate that while a well-trained AI can achieve performance comparable to human raters, human oversight remains essential to address nuanced or uncertain cases. The study contributes insights into improving rubric design and integrating AI in a "human-in-the-loop" grading process to enhance reliability and fairness in automated assessment.

Objectives

Modeling is crucial for enhancing students' knowledge-in-use, particularly through multimodal assessments in the realm of three-dimensional science learning (NGSS, 2013). Despite their significance, the intricate and diverse nature of these models complicates their evaluation (Lee et al., 2023). Assessing models intended to measure students' knowledge-in-use is further complicated by the multifaceted nature of the cognitive processes involved (Li et al., 2024). AI-assisted assessment has emerged as a promising approach to manage the heavy workload of grading complex student work. In science education, researchers have explored automated scoring of student-constructed scientific models and diagrams. For example, convolutional neural network (CNN) models have achieved high accuracy in classifying student-drawn responses into rubric categories—one study reported a CNN correctly predicted up to 97.7% of image-based responses, matching or exceeding typical human rater accuracy (von Davier et al., 2022). Moreover, the CNN was even able to correctly score some responses that human graders had initially scored incorrectly (von Davier et al., 2022) These advances suggest AI can potentially take on grading tasks with efficiency and consistency. However, fully autonomous scoring is not without challenges. Prior research highlights concerns about the reliability and trustworthiness of AI in nuanced evaluation tasks (Kortemeyer & Nöhl, 2024). In high-stakes contexts, automated grading is considered "high-risk," and regulations mandate that human oversight is obligatory when AI is used for evaluating learning outcomes (European Union, 2024). This underscores that human experts must remain in the loop to ensure grading integrity (Kashy et al., 2001).

A key challenge in AI-assisted grading is ensuring clarity and alignment of the scoring rubric for both humans and AI. Ambiguously defined criteria can lead to inconsistent judgments. As prior studies note, the language and descriptors in a rubric are critical—*an ambiguous rubric cannot be accurately or consistently interpreted by instructors, students or scorers*u.osu.edu. Even trained human raters can diverge in their scoring if rubric guidelines are unclear or open to interpretation. Rater training and calibration improve agreement but never eliminate differences entirely (Jonsson & Svingby, 2007). These issues imply that any AI trained on human-generated scores may also reflect or even amplify inconsistencies stemming from rubric ambiguity.

Given these considerations, it is important to examine how human and AI scores compare, where they diverge, and how involving humans in the loop can resolve discrepancies. Human experts bring contextual understanding and can judge complex or borderline cases, whereas AI provides efficiency and consistency in straightforward cases. The interplay between the two raises several questions. This study addresses the following research questions:

1. **RQ1**: Human Score Stability – How consistent are human scores upon reevaluation? Specifically, when examining cases flagged due to discrepancies between AI predictions and original human scores, do independent raters uphold the initial judgment or diverge, suggesting instability or ambiguity in rubric criteria?
2. **RQ2**: AI-Human Agreement – To what extent do AI-predicted scores align with human raters' reevaluation? What is the magnitude of agreement, and does it vary across different rubric categories?
3. **RQ3**: Scoring Inconsistencies – What characterizes cases where AI and human scores remain consistently inconsistent, despite reassessment?
    a. **RQ3a**: Persistent AI Disagreement – In which rubric categories does the AI systematically fail to align with human scorers, and what visual or representational features contribute to these mismatches?
    b. **RQ3b**: Human Instability – In which rubric categories do human raters themselves exhibit persistent disagreement, and what rubric ambiguities or representational complexities drive this variability?

By investigating these questions, we aim to illuminate the reliability of AI-assisted scoring and the essential role of human oversight. We anticipate that our findings will highlight categories where rubric definitions may need refinement and demonstrate how a human-in-the-loop approach can improve the overall scoring process. Ultimately, this work seeks to contribute guidelines for integrating AI into educational assessment in a way that enhances efficiency without sacrificing validity and fairness.

Methods

This study situated in an automated formative assessment system leverages AI to evaluate students' multi-modal assessments and offer tailored feedback. The system bolsters formative assessment practices within a high school physical science curriculum named "*Interactions*" which promotes three-dimensional learning, with curriculum materials aligned with the Next Generation Science Standards (NGSS, 2013). This study uses the "Electroscope" modeling task (Figure. 1) and employs a 13-category rubric for assessing students' knowledge-in-use performance and aligned to specific model components and/or relationships (Kaldaras et al., 2022). Categories 1-10 evaluate the comparative presence or absence of charges on electroscope components in scenarios A and B, while Categories 11-13 identify inaccuracies in responses. For instance, Category 13 checks whether the rod or the entire electroscope in Scenario A is

uncharged in a model. The human-human interrater reliability by Krippendorf's alpha exceeded 0.8 for most categories, indicating substantial agreement. For this study, we only focus on the first 10 categories (C1-C10) given the categories 11-13 were designed to capture inaccuracy.

Machine Algorithm Development and Validation. We developed algorithms based on Convolutional Neural Networks (CNN) to score student models and compared these predicted scores with those from human (Krizhevsky & Hinton, 2012). The validity of the CNN model was confirmed using a 10-fold cross-validation method. The 1151 student models were randomly divided into ten groups, assigning 10% for testing and 90% for training. The training set was further divided into validation and actual training sets using a 1:4 ratio. For feature extraction, we employed the ResNet-18 architecture, implemented our model in Pytorch, and optimized using Adam with a learning rate of 1e-4. The networks underwent training for 500 epochs on an NVIDIA GeForce GTX 1080Ti graphics card. We determined the human-machine scoring agreement (HMA) accuracy during both the training and validation phases. After each epoch, the validation accuracy was computed, and the network weights were stored. Following training epochs, validation accuracy was determined using the weights from the epoch with the highest validation accuracy (Lu & Tran, 2017). We averaged the validation accuracies across folds throughout iterative training. Table 1 provides the results as agreement between human and machine assigned scores.

Data Preparation. During the model's testing phase (N=327), any discrepancy between the AI's score and the original human-assigned score was flagged for further review. In other words, every instance where the deep learning model's predicted score did not match the *human original* score (across any of the 10 rubric categories) was identified as a flagged case. All flagged responses were independently re-evaluated by two expert human raters who were well-versed in the scoring rubric. During this rescoring phase, each rater was blinded to both the original human scores and the machine-predicted scores to prevent any biases or influences. They reassessed each response strictly according to the rubric criteria, assigning binary scores (0 or 1) independently. The raters also provided notes indicating the rationale behind their scoring decisions, especially when facing uncertainty or ambiguity. Importantly, no consensus discussions were held between the raters after rescoring. Instead, both sets of rescored data were preserved independently, allowing us to analyze inter-rater reliability directly. This design enabled us to measure and report exact inter-rater reliability between the two human raters without the potential influence of discussions or consensus-building, providing an authentic assessment of rubric clarity and scoring challenges.

Analytic Strategies for Each Research Question

For Research Question 1 (Human Score Stability), we aimed to examine the stability and consistency of human scores across time by comparing initial consensus-based human scores and subsequent rescoring performed independently by two new raters (Rater 1 and Rater 2). We used Fleiss' Kappa to evaluate the level of agreement among these three human-generated scores—original human consensus, Rater 1, and Rater 2—because Fleiss' Kappa is specifically designed for assessing agreement among multiple raters simultaneously. Cohen's Kappa was calculated to evaluate the level of agreement between the two independent raters (Rater 1 and Rater 2). This metric is suitable for examining pairwise agreement on binary-coded rubric categories. This analysis allowed us to comprehensively gauge the stability and reliability of human scoring across different time points and raters, identifying any systematic variability that might indicate inconsistencies in rubric interpretation or scorer judgment.

For Research Question 2 (AI-Human Agreement), we explored the magnitude and patterns of agreement between the AI-generated scores and human-generated scores, specifically after human rescoring. Initially, by definition, AI scores had zero percent agreement with the original human consensus for the cases selected, as flagged discrepancies were intentionally chosen based on disagreement. However, it remained essential to quantitatively evaluate the degree to which the human rescores aligned with AI scores upon reassessment. Therefore, we applied Fleiss' Kappa to measure the agreement among Machine predictions, Rater 1, and Rater 2 scores simultaneously. This analysis provided insight into whether the newly reassigned human scores tended to align more closely with AI assessments, thus offering important insights regarding the potential accuracy or systematic biases within the AI model's scoring mechanism.

Finally, for Research Question 3 (Scoring Inconsistencies), our strategy involved identifying and qualitatively analyzing persistent discrepancies that could not be resolved through rescoring. Specifically, we isolated cases where consistent disagreement was observed between AI and human raters, as well as between raters themselves. These cases were extracted based on their continuous disagreement across all scoring attempts—original human consensus, independent human rescores (Rater 1 and Rater 2), and AI predictions. A qualitative thematic analysis of these cases was then conducted, using rater annotations and comments to identify underlying causes for discrepancies. This allowed us to determine whether persistent scoring disagreements arose from ambiguities or deficiencies in rubric clarity, limitations in the AI scoring model's feature interpretation, or a combination of these factors. This qualitative analysis aimed to provide rich, nuanced insights into the conditions under which scoring remains problematic, directly informing future improvements in rubric design, AI model training, and overall assessment strategies.

Results

RQ1. Human Analysis Stability

To address RQ1, we examined the degree of consistency among human scores over time by comparing three sources: the original consensus human score (used as ground truth during AI training), and two independent rescoring judgments (Rater 1 and Rater 2) on cases previously flagged for AI-human disagreement. Cohen’s Kappa was used to evaluate pairwise agreement between Rater 1 and Rater 2, Fleiss’ Kappa1 was used to assess multi-rater agreement across all three human scores (Original, Rater 1, Rater 2), and Fleiss’ Kappa2 was used to assess multi-rater agreement across human rescores and AI scores (AI, Rater 1, Rater 2). Table 1 reports agreement statistics for each rubric category.

Table 1. Inter-Rater Reliability Metrics for Scores across Rubric Categories

Category	N Flagged Cases	Cohen's Kappa (Rater1 vs Rater2)	Fleiss' Kappa1 (Original, Rater1, Rater2)	Fleiss' Kappa2 (AI, Rater1, Rater2)
C1	15	1.000	0.030	0.154
C2	10	0.615	0.593	–0.222
C3	9	0.769	0.365	0.110
C4	21	0.712	0.744	–0.273
C5	12	0.211	0.443	–0.337
C6	28	0.314	0.375	–0.235
C7	17	0.866	0.748	–0.234
C8	17	0.757	0.753	–0.255
C9	20	0.211	0.214	–0.154
C10	15	0.167	0.351	–0.334

Substantial to near-perfect inter-rater agreement between Rater 1 and Rater 2 was observed in Categories C1, C3, C4, C7, and C8, with Cohen's Kappa values exceeding .70, indicating consistent interpretation of the scoring rubric by independent raters. However, in Category C1, despite a perfect Cohen's Kappa (1.00) between the new raters, the Fleiss' Kappa1 was extremely low (0.03). This pattern suggests that both new raters consistently disagreed with the original consensus score, revealing a significant instability in the original human scoring for that category. By contrast, lower levels of agreement were observed in Categories C5, C6, C9, and C10. For example, Category C9 had both low Cohen's Kappa (0.21) and low Fleiss' Kappa1 (0.21), indicating general disagreement among all raters and suggesting potential issues in rubric clarity or difficulty in interpreting student responses for that model component. These results suggest that while overall human scoring is relatively stable across time in some categories, certain rubric elements may suffer from interpretive ambiguity or inconsistent application, particularly in flagged cases where student responses challenge straightforward classification. This has important implications for both rubric design and the role of human oversight in AI-assisted assessment systems.

RQ2. AI-Human Agreement

To address RQ2, we examined the degree of agreement between the AI-generated scores and those produced by human raters after rescoring. Since all selected cases represented flagged discrepancies (i.e., AI and original human scores disagreed), we focused on evaluating whether the AI's predictions aligned more closely with the human judgments upon reassessment. Fleiss' Kappa2 was computed for each rubric category across three raters: the AI-predicted score, Rater 1, and Rater 2. This inter-rater reliability metric provides insight into the extent to which the rescored human judgments validate or diverge from AI-generated assessments.

Agreement between AI and human scorers remained low across most rubric categories, with negative Fleiss' Kappa2 values in 8 out of 10 categories. Only Categories C1 and C3 exhibited slightly positive Kappa values, suggesting minimal alignment. The poorest agreement was found in C5 ($\kappa = -0.337$) and C10 ($\kappa = -0.334$), indicating that AI systematically failed to align with human judgments in these areas. Based on (Fleiss' Kappa < 0, but Cohen's Kappa >

0.6), we identified four categories that present consistent human-AI disagreement, including C2, C4, C7, and C8. These results suggest that AI models trained on human scores may not generalize well to ambiguous or edge-case responses—precisely those most often flagged for review. This reinforces the need for: Iterative retraining of AI models with rescore data to improve boundary-case prediction; Rubric refinement for complex or poorly operationalized criteria; Continued human oversight in high-stakes or formative assessment contexts, particularly in scoring representations of student reasoning.

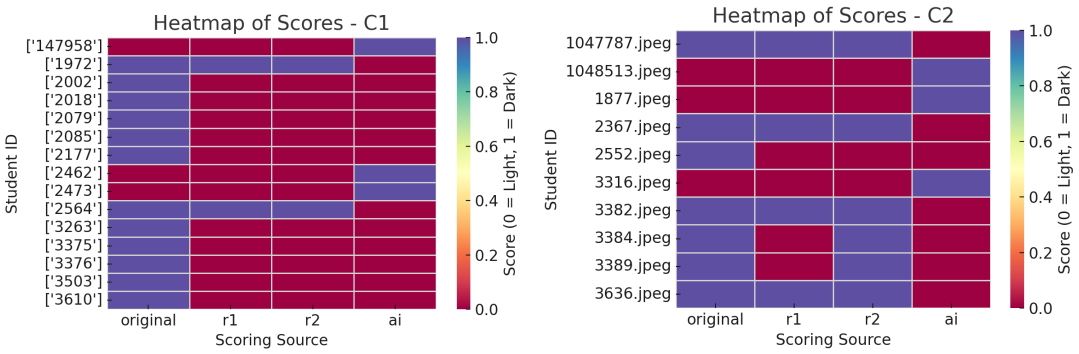
RQ3. Qualitative analysis.

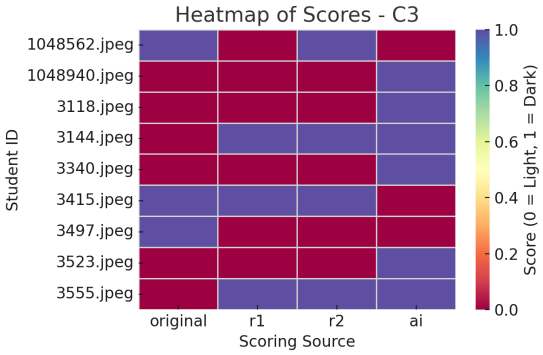
To address RQ3 (Scoring Inconsistencies)—which aims to uncover the underlying factors contributing to persistent disagreement between human and AI scoring—we implemented a structured, data-informed qualitative selection and analysis process grounded in our prior quantitative results. We began by identifying three distinct types of cases for qualitative follow-up based on well-defined disagreement patterns among AI and human scores from RQ2. Table 2 summarizes the logic behind each type and its interpretive value.

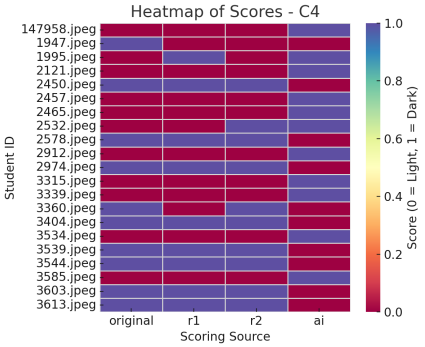
Table 2. *Qualitative Case Selection Strategy*

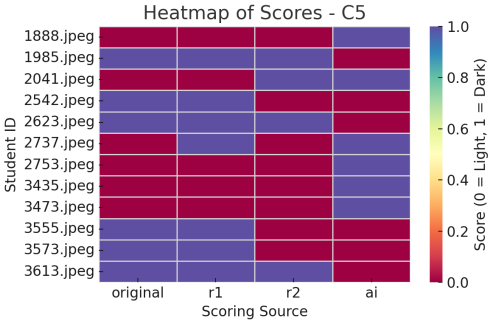
Case Type	Selection Logic	Interpretive Value
Persistent AI Disagreement	Original ≠ AI, Rater 1 ≠ AI, Rater 2 ≠ AI	AI systematically disagrees with all human raters, suggesting a failure to generalize or interpret certain features.
Human Instability	Original ≠ Rater 1, Original ≠ Rater 2, Rater 1 ≠ Rater 2	Indicates potential ambiguity in rubric criteria or interpretive differences among raters.

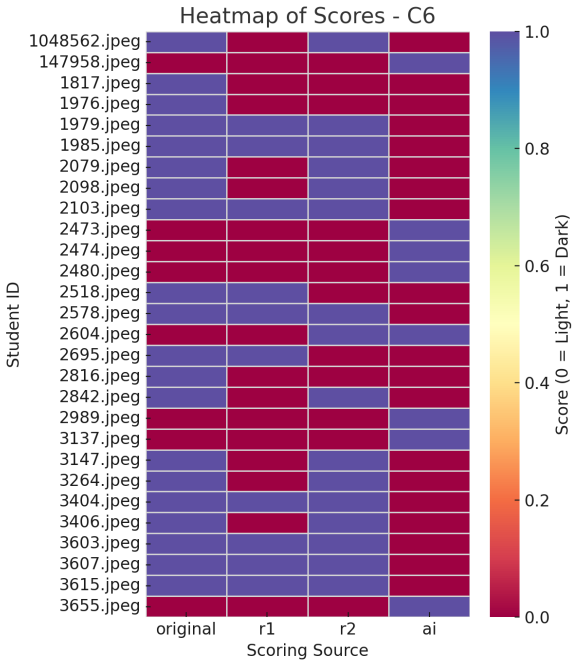
To identify the specific cases, we used heatmaps to map out each categories flagged cases.

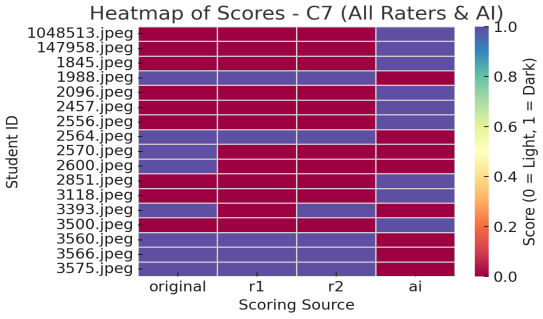


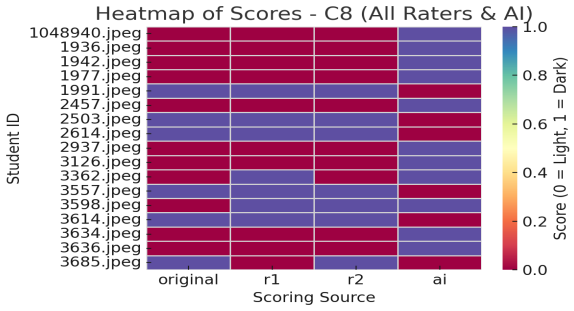

Heatmap of Scores - C3
1048562.jpeg
1048940.jpeg
3118.jpeg
3144.jpeg
3340.jpeg
3415.jpeg
3497.jpeg
3523.jpeg
3555.jpeg
Student ID
original
r1
r2
ai
Scoring Source
Score (0 = Light, 1 = Dark)
1.0
0.8
0.6
0.4
0.2
0.0

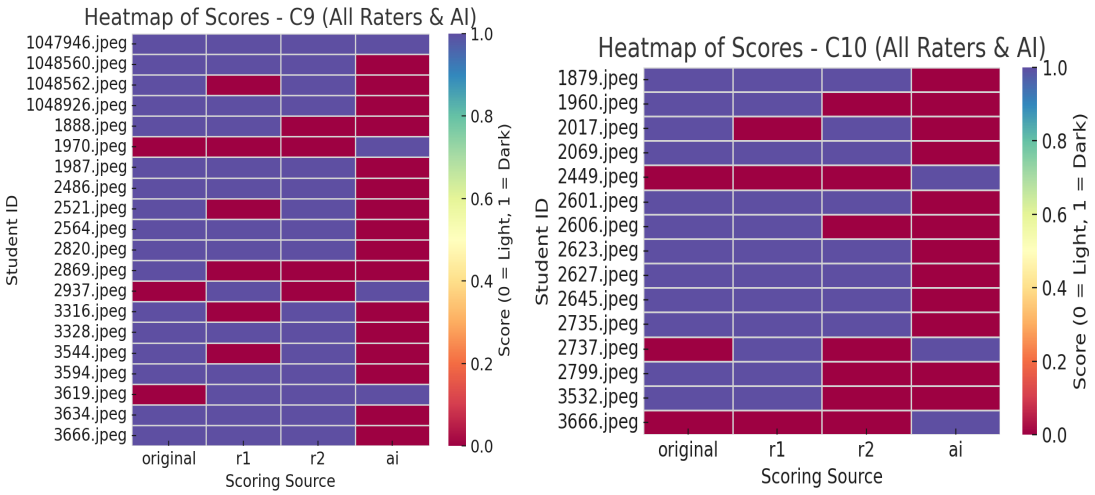

Heatmap of Scores - C4
147958.jpeg
1947.jpeg
1995.jpeg
2121.jpeg
2450.jpeg
2457.jpeg
2465.jpeg
2532.jpeg
2578.jpeg
2912.jpeg
2974.jpeg
3315.jpeg
3339.jpeg
3360.jpeg
3404.jpeg
3534.jpeg
3539.jpeg
3544.jpeg
3585.jpeg
3603.jpeg
3613.jpeg
Student ID
original
r1
r2
ai
Scoring Source
Score (0 = Light, 1 = Dark)
1.0
0.8
0.6
0.4
0.2
0.0


Heatmap of Scores - C5
1888.jpeg
1985.jpeg
2041.jpeg
2542.jpeg
2623.jpeg
2737.jpeg
2753.jpeg
3435.jpeg
3473.jpeg
3555.jpeg
3573.jpeg
3613.jpeg
Student ID
original
r1
r2
ai
Scoring Source
Score (0 = Light, 1 = Dark)
1.0
0.8
0.6
0.4
0.2
0.0


Heatmap of Scores - C6
1048562.jpeg
147958.jpeg
1817.jpeg
1976.jpeg
1979.jpeg
1985.jpeg
2079.jpeg
2098.jpeg
2103.jpeg
2473.jpeg
2474.jpeg
2480.jpeg
2518.jpeg
2578.jpeg
2604.jpeg
2695.jpeg
2816.jpeg
2842.jpeg
2989.jpeg
3137.jpeg
3147.jpeg
3264.jpeg
3404.jpeg
3406.jpeg
3603.jpeg
3607.jpeg
3615.jpeg
3655.jpeg
Student ID
original
r1
r2
ai
Scoring Source
Score (0 = Light, 1 = Dark)
1.0
0.8
0.6
0.4
0.2
0.0


Heatmap of Scores - C7 (All Raters & AI)
1048513.jpeg
147958.jpeg
1845.jpeg
1988.jpeg
2096.jpeg
2457.jpeg
2556.jpeg
2564.jpeg
2570.jpeg
2600.jpeg
2851.jpeg
3118.jpeg
3393.jpeg
3500.jpeg
3560.jpeg
3566.jpeg
3575.jpeg
Student ID
original
r1
r2
ai
Scoring Source
Score (0 = Light, 1 = Dark)
1.0
0.8
0.6
0.4
0.2
0.0


Heatmap of Scores - C8 (All Raters & AI)
1048940.jpeg
1936.jpeg
1942.jpeg
1977.jpeg
1991.jpeg
2457.jpeg
2503.jpeg
2614.jpeg
2937.jpeg
3126.jpeg
3362.jpeg
3557.jpeg
3598.jpeg
3614.jpeg
3634.jpeg
3636.jpeg
3685.jpeg
Student ID
original
r1
r2
ai
Scoring Source
Score (0 = Light, 1 = Dark)
1.0
0.8
0.6
0.4
0.2
0.0



RQ3a: When and Why AI and Humans Disagree

To investigate persistent discrepancies between AI and human scoring, we pull out cases where the AI-predicted score disagreed with all three human-derived scores: the original consensus score and the independent rescoring from Rater 1 and Rater 2. These cases are referred to as "Persistent AI Disagreement" and represent instances in which the AI model failed to align with any human judgment.

Table 3. Persistent AI Disagreement Case Summary

Rubric Category	N Flagged Cases	N Persistent AI Disagreement Cases	Case IDs	Percent of Cases in Category (%)
C1	15	5	147958, 1972, 2462, 2473, 2564	33.3
C2	10	7	1047787, 1048513, 1877, 2367, 3316, 3382, 3636	70.0
C3	9	5	1048940, 3118, 3340, 3415, 3523	55.6
C4	21	17	147958, 2121, 2450, 2457, 2465, 2578, 2912, 2974, 3315, 3339, 3404, 3534, 3539, 3544, 3585,3603, 3613	81.0
C5	12	7	1888, 1985, 2623, 2753, 3435, 3473, 3613	58.3
C6	28	15	147958, 1979, 1985, 2103, 2515, 2623, 2644, 2683, 2753, 2922, 3018, 3118, 3132, 3435, 3473	53.6
C7	17	14	1048046, 1048722, 1049104, 1049638, 1979, 2103, 2578, 2644, 2683, 2922, 3018, 3118, 3132, 3287	82.4
C8	17	14	147958, 1972, 2121, 2515, 2644, 2683, 2753, 2974, 3132, 3404, 3435, 3473, 3539, 3613	82.4
C9	20	11	1048940, 1049638, 2103, 2457, 2465, 2753, 2922, 2974, 3132, 3534, 3613	55.0
C10	15	9	147958, 1888, 2103, 2450, 2623, 2683, 2974, 3315, 3539	60.0

RQ3a. To respond to RQ3a, we identified Categories 2, 4, 7, and 8 based on high human raters' agreement but persistent disagreement with AI-predicted scores (low Fleiss' Kappa2). Using heatmaps generated for each category, we systematically identified persistent AI disagreement cases. We then analyzed these cases individually within each category and synthesized common themes across the four categories to uncover potential reasons behind persistent AI-human scoring disagreements.

Textual Annotations Not Read by AI. Many students wrote words on their drawings (e.g., "no charge" or "electrons go here"). Humans naturally read these and factor them into scoring. The CNN, lacking text recognition, sometimes missed the student's intention. In some cases, the human gave credit due to an annotation, whereas the AI, seeing only unrecognized pixels, did not. *For example, model ID #2079 is illustrative.*

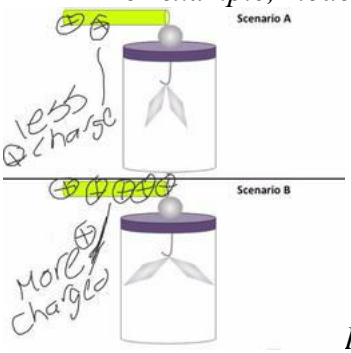


Figure 1. shows this student's model. The student wrote "less charge" near the rod in Scenario A and "more – charge" near Scenario B (note the hand-written notes in Figure 1). The human rater credited the student for indicating the difference in charge between the scenarios (fulfilling the "differences between scenarios" criterion and partially the explanation criterion). The CNN, however, only "saw" that the drawings of the electroscope in A and B were essentially identical visually (the leaves look the same in both, and no obvious graphical change apart from the small scribbles). It therefore did not mark the difference as present. This resulted in a discrepancy: the human score reflected the textual explanation of "less vs more charge," but the AI score treated it as if no difference was shown. This pattern occurred in several cases where student handwriting on the image carried key information. In those instances, the AI's limitation in reading text led to under-scoring relative to the human. One potential remedy for this in the future would be to incorporate an OCR (optical character recognition) component or to have students submit a written explanation separately that the AI can analyze in conjunction.

Unconventional or Abstract Representations. Some students used creative approaches that the model was not explicitly trained on. Figure 2 below presents an example, model ID #2842, where the student introduced an abstract representation.

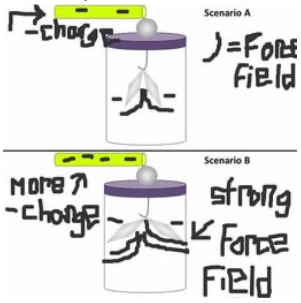


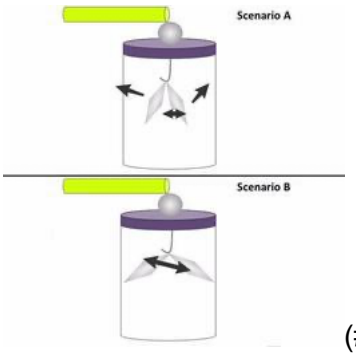
Figure 2. depicts Scenario A and B for that model. Instead of using standard "+" or "–" symbols for charge or drawing the leaves separated, this student drew bold black curves inside the jar and labeled them as a "force field." They also wrote text indicating

Scenario A is “neutral” and Scenario B has a “strong force field” with “more – charge.” The human rater recognized that the student was attempting to convey that Scenario B has a stronger effect (more charge, thus some kind of force field causing the leaves to move). It’s partially incorrect scientifically (an electroscope isn’t usually described in terms of a force field, and the student did not explicitly draw the leaves diverging), so the human gave only partial credit on the explanation and did not give credit for correctly showing the leaf separation (since the leaves in the drawing remain hanging, possibly an oversight by the student). The CNN, on the other hand, was thrown off by the unusual drawing. It saw a lot of extra black markings in Scenario B (the “field” lines) which were not present in Scenario A. Because in most training examples, “more lines or markings in B” correlated with showing a change or something happening, the model falsely assumed this student had depicted the leaves moving or a difference, and it gave credit for the scenario difference category where the human did not. This resulted in an AI over-score for that category. This case exemplifies an alternative expression issue: a student represented the concept in a non-standard way, leading to a mismatch in interpretation. The AI lacks the contextual understanding to know those lines were meant as a field (and that this was not what the rubric was looking for), whereas a human could make a nuanced judgment. These outlier representations often led to AI errors, since the model relied on pattern recognition and had difficulty with items outside its learned patterns.

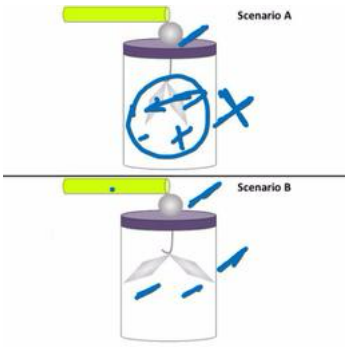
Ambiguous or Low-Quality Drawings. In some cases, the student’s drawing itself was unclear, and the AI and human simply interpreted it differently. For instance, a few students drew very light pencil marks or had cluttered sketches. In one case, a small arrow was drawn but almost invisible; the human missed it and gave no credit for showing charge movement, but the CNN’s image processing (which can detect subtle gradients) actually picked up something in that area and erroneously treated it as a valid arrow, thus giving a point. Conversely, there were cases where a student’s poorly drawn symbol might have been ignored by the AI (which didn’t recognize the shape) but a human deduced what the student meant. These discrepancies were essentially due to perceptual ambiguity. They underscore that even humans can disagree on what is in the image – and the AI might latch onto visual noise or miss faint details. We found that improving image quality (scans) and possibly instructing students to draw clearly could mitigate some of these issues for both human and AI scorers.

RQ3b. To respond to RQ3b, we identified Categories 5, 6, 9, and 10 based on both Cohen’s Kappa and Fleiss’ Kappa (Original, R1, and R2), which revealed persistent human-human disagreement across raters. We examined representative models in each category to investigate how differences in interpretation and rubric ambiguity may contribute to human scorer instability.

Imprecise or Ambiguous Force Representation. Across Categories 5 and 10, models often used vague or inconsistent force arrows. For instance, Case 2017 (C10) showed arrows in both scenarios but failed to make Scenario B arrows clearly larger or bolder than those in Scenario A, as required by the rubric. In Category 5, Case 2041 featured bidirectional arrows between leaves, which some raters may interpret as electric field representation rather than explicit repulsive force. The lack of standardization in arrow size, direction, and meaning led to varying interpretations across human raters.

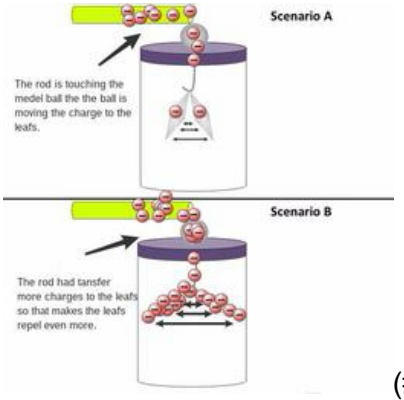


(#2017)

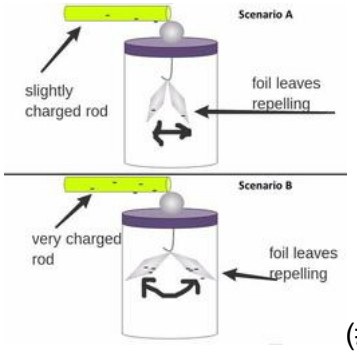


(#2041)

Confusion Between Charge Quantity and Effect. Category 6 and 9 models frequently blurred the distinction between the quantity of charge and its effect on the system. In Case 2518 (C6), students illustrated more charges in Scenario B, but without clearly showing charges on rob in B are more than rod in A. One rater took notes in the rescoring stage, "Pretty much the same amount of charge on each rod." Thus, this rater gave 0 while the other rater gave 1. Similarly, in Category 9, Case 3544 depicted more charges on the leaves in Scenario B, but did not show a noticeable difference from Scenario A in spread angle or repulsion, making it difficult for raters to judge rubric alignment. This disconnect between charge quantity and observable impact introduced subjectivity into scoring.

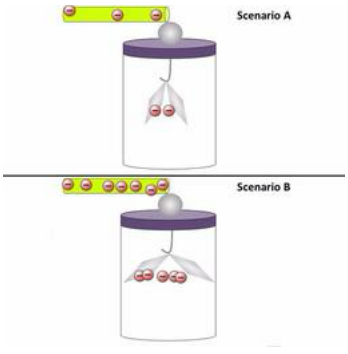


(#2518)

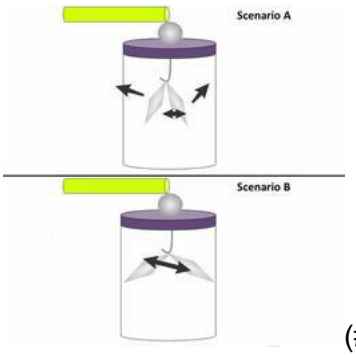


(#3544)

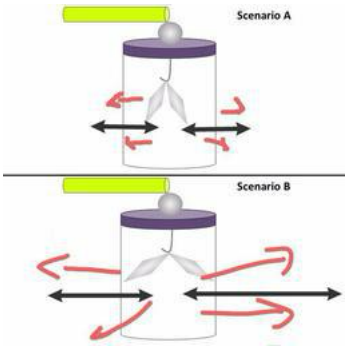
Implicit Visual Reasoning. In several Category 9 and 10 models, students embedded their reasoning in subtle visual cues rather than explicit annotations. For instance, Case 1888 (C9) used minimal symbols with slight variation in leaf angle between scenarios, requiring raters to infer whether the change was intentional and scientifically meaningful. In this case, one rater rescored it as "0" and gave rationale of "Charges not on the leaves." In Case 2017 (C10), arrows were drawn in Scenario B, but their placement and boldness were not markedly different from Scenario A. Similarly, for Case 2737 (C10), one rater gave "0" given "Arrows not between the leaves." These implicit cues led to inconsistent interpretation among raters who weighed subtle visual elements differently.



(#1888)



(#2017)



(#2737)

Discussion

Our study examined an AI-assisted scoring approach for complex student-generated science models, yielding findings with important implications for educational assessment practice and research. We discuss these implications in three areas: (1) rubric and assessment design, (2) human-AI collaboration in scoring**,** and (3**)** training and refining AI models with human feedback**.** We also consider the broader significance and limitations of this work.

Implications for Rubric Development and Assessment Design

The results highlight how the design and clarity of a scoring rubric directly impact both human and AI scoring performance. The variability we observed in human rater agreement across rubric categories suggests that some aspects of the construct were not as sharply defined or were inherently more complex to judge. This calls for rubric refinement: clarifying descriptors, providing anchor examples at each score level, or breaking a broad category into more discrete sub-components. Our analysis of discrepancies pointed out that certain student representations (like text annotations or alternative symbols) were handled inconsistently – sometimes credited, sometimes not. A well-tuned rubric could anticipate these possibilities. For example, if written labels on a diagram are acceptable evidence for a concept, the rubric (and rater training) should explicitly include that. Conversely, if the intention is to have students convey an idea visually, the rubric might specify that credit is only given for graphical representation, not just written notes. Being explicit about such criteria would likely improve human scoring consistency and also make it easier to train AI, which thrives on well-defined targets. Moreover, our findings resonate with observations by Zhai et al. (2022), who identified representational issues like *alternative expressions* and *confusing labels* as factors impacting automated scoring. The presence of these factors in our data (e.g., students inventing novel notation or using unclear labeling) suggests that test developers and educators should strive to design modeling tasks that minimize unnecessary ambiguity. This doesn't mean limiting student creativity, but rather guiding students on how to communicate their ideas clearly. For instance,

teaching students standard ways to denote charges or encouraging them to include a brief written legend for any unique symbols they use could reduce misinterpretation. In essence, improving the task design (instructions given to students) can lead to models that are easier to score reliably by both humans and machines. Finally, the rubric development process itself might benefit from iterative piloting with an AI in the loop. Traditionally, rubric refinement relies on human scoring trials and analyzing disagreements (Bresciani et al., 2009). Our study shows that an AI can highlight the same troublesome areas. If during rubric design one trains a provisional AI model on pilot data, the AI's confusion can point to rubric criteria that need tightening. For example, if the AI has trouble distinguishing scores for a certain category, it might be because the category is ill-defined or the features are too subtle – something that might also confuse human raters. In this way, AI can act as an "analytical lens" to examine rubric quality. Overall, our work reinforces that effective assessment of complex performances like modeling requires careful construct definition and may involve an ongoing cycle of rubric revision and training calibration.

Human–AI Collaboration in Assessment

Our study provides evidence that human-AI collaboration can combine the strengths of each in the scoring process, resulting in more efficient yet trustworthy assessment. The CNN model achieved a level of agreement with human scores that was comparable to human rater (Rater 2) for many rubric categories. This suggests that AI models, once properly trained, can serve as effective assistants or second markers in routine scoring. For instance, in a classroom or large-scale exam setting, the AI could score all student models and flag only those cases where it is uncertain or where certain rubric criteria are met in unusual ways. A human instructor or assessor could then focus their limited time on reviewing these flagged cases. This approach could dramatically reduce grading time while maintaining quality. Our findings showed that about 22% of responses had any AI-human discrepancy; if an AI flagged roughly that proportion for human review, the human would not need to look at the other ~78% that the AI is confident on (assuming the AI's confidence aligns with correctness). In practice, an educator might still spot-check some of the AI-graded ones, but the workload would be much lower than grading 100% from scratch.

Crucially, we do not advocate for removing the human from the loop. The discrepancies we analyzed underscore why human oversight remains vital. AI is prone to particular failure modes (like missing meaning in text, or being fooled by odd drawings) that a human can catch. Meanwhile, humans have their own inconsistencies or lapses that AI can complement. The ideal system is a synergistic partnership: the AI rapidly analyzes and provides preliminary scores or identifies likely errors, and the human adjudicates the tricky cases and ensures fairness. This aligns with emerging regulations and ethical guidelines which mandate that algorithmic decisions in education be subject to human review (Nguyen, 2025). By designing a workflow where AI handles the heavy lifting and humans handle the edge cases, we satisfy the requirement for human oversight while leveraging AI's efficiency. Such a human-AI collaborative approach has been suggested in other domains (e.g., essay scoring or even medical diagnosis) as yielding better outcomes than either alone (**European Commission,** 2021). Our concrete results add credibility to this approach in the context of scoring student diagrams: we've shown the AI can match human judgment for the majority of straightforward cases and that the exceptions are manageable with targeted human attention.

From a practical standpoint, implementing human-AI collaboration could involve an interface for teachers where the AI scores are presented alongside indicators of confidence or flags for unusual features. Teachers could then quickly review those flagged by the system. This

not only saves time but can also function as professional development – by examining cases where the AI had trouble, teachers might become aware of common student misconceptions or creative strategies that they hadn't considered. In our study, reviewing AI-human discrepancies was illuminating; a teacher doing the same in real-time could similarly gain insights and adjust instruction. In sum, the role of AI here is to *augment* the human grader, not replace them, in line with the principle of maintaining fairness and transparency in AI-assisted education.

Improving AI Models with Human-in-the-Loop Feedback

A key contribution of this work is highlighting how human review data can be used to iteratively improve AI scoring models**.** During our analysis, we effectively performed an error analysis on the AI's outputs and used human judgment to categorize those errors. This kind of information is extremely valuable for refining the model. For example, once we identified that the CNN systematically missed text, one straightforward step for future model versions would be integrating an OCR module or a multi-modal model that processes text and images together. Similarly, knowing that the model misinterpreted certain abstract drawings, we could incorporate more diverse training examples or apply data augmentation techniques to expose the model to a wider range of representations. In general, every discrepancy case is an opportunity: if a human can determine the correct outcome, that example (originally mis-predicted by the model) can be added to the training set for the next iteration. This is a form of active learning or targeted retraining, where the focus is on model weaknesses.

One strategy in this vein, suggested by others, is to deliberately sample additional training data from cases the model finds challenging (von Davier et al., 2022). In large-scale assessments, one could use an approach where after an initial model is trained, a subset of responses (especially those where the model's confidence is low or it disagrees with a human rater) is sent to human experts for re-scoring. These newly confirmed labels then feed back into the model for retraining, hopefully improving its accuracy on those and similar cases. Our results support the efficacy of such a strategy: we saw that many of the model's errors were systematic, not one-off random noise. This means the model could likely learn from additional examples. For instance, providing the model with several examples of the "force field" type drawing (with correct labels indicating it does *not* count as showing the expected phenomenon) would help it adjust its internal representation and not give credit erroneously in the future. Over time, this human-in-the-loop training could greatly reduce the discrepancy rate. Essentially, the model becomes more robust by learning from the very cases that initially confounded it.

We also recognize that there is a limit to how much a purely image-based CNN can learn if some information is simply not present in the pixels (e.g., semantic meaning of words). Thus, another improvement direction is to broaden the AI's capabilities – for example, combining computer vision with natural language processing for a multi-modal scoring system. If students provide a separate written explanation along with the drawing (as was the case in some modeling tasks in other studies), an AI could analyze both and perhaps achieve deeper understanding. In our scenario, if no separate text is provided, an advanced model like a Vision Transformer or GPT-4V (vision-enabled language model) could be prompted with the image and rubric to generate scores (Chu et al., under review). Early research in this area (e.g., using GPT-4 Vision to evaluate drawings) indicates potential, though it also underscores the need for clear rubric instructions to guide the AI. Regardless of the specific technique, the common theme is that incorporating human insights (either by labeling data or encoding rules derived from human expertise) is essential for developing AI systems that are accurate, fair, and aligned with educational values.

Theoretical and Practical Significance

The outcomes of this study contribute to both the theory and practice of educational measurement in science. Theoretically, we add to the growing evidence that complex performance assessments can be reliably scored with the assistance of AI, which challenges the traditional notion that only selected-response or constrained tasks can be objectively graded. We show that with a detailed rubric and enough training data, even something as rich as a student's scientific drawing – which encapsulates a mix of conceptual understanding and creativity – can be assessed consistently. This opens up possibilities for broader use of performance tasks in science classrooms, alleviating the assessment burden that often limits their use. Our work also highlights an emerging conceptualization of reliability: when considering AI as part of the assessment process, one must consider *AI-human agreement* as analogous to inter-rater reliability. High AI-human agreement (especially approaching human-human agreement levels) can be interpreted as evidence of validity and consistency in the scoring process. Some scholars have even suggested that AI could serve as a "third rater" in validation studies of scoring, offering another lens on the reliability of scoring rubrics. We provide concrete data in this regard, and also nuance the interpretation by showing where AI and human differences revealed potential issues with the task or rubric.

Practically, for educators and assessment designers, our study offers a proof of concept and a set of guidelines for implementing AI-assisted scoring in a responsible manner. By identifying the types of rubric criteria that are amenable to AI scoring versus those that are tricky, one can design assessments that play to AI's strengths (e.g., visual pattern recognition of structures) and plan for human oversight on the subtle parts (e.g., interpreting reasoning). The guidelines emerging from our findings include: (1) Design rubrics with clear, observable indicators – this benefits both human and AI scoring. (2) Use AI as an initial scorer or second rater to significantly reduce grading time, but always include a mechanism for human review of flagged cases to catch the nuance AI might miss. (3) Continuously improve the AI model by analyzing its errors and feeding it additional training data or rules, especially focusing on systematic discrepancies (our analysis template could serve as a model for educators to periodically audit their AI's performance). (4) Maintain transparency with students – if AI is used in grading, inform students and, where possible, provide explanations for scores. Interestingly, AI can be used not just for scoring but as a feedback tool: for example, if an AI can detect that a student didn't show a difference between scenarios, it could automatically prompt the student to reconsider their model, thus acting as a formative feedback system. This blends assessment with learning in a powerful way.

Limitations and Future Work

It is important to acknowledge the limitations of this study. First, the scope was a single type of modeling task in physics. The extent to which our results generalize to other science topics or different kinds of representations (e.g., ecosystem models, anatomical diagrams, etc.) needs investigation. Different tasks might introduce new challenges for rubrics and AI (for instance, coloring, 3D perspectives, etc., not present in our electroscope task). Second, our CNN model, while effective, was relatively domain-specific and did not incorporate text analysis. Future work could explore more advanced or generalizable models, such as those using both vision and language capabilities, to handle multi-modal student responses. Third, in our study the human was considered the gold standard. We did not undertake a separate expert resolution of discrepancies beyond our analysis; in an operational setting, one might convene experts to

adjudicate each AI-human disagreement to build an even higher-quality dataset for training. Doing so could further improve model performance but was beyond our research scope. Another limitation is that we did not deeply examine potential *biases* in AI scoring with respect to student subgroups. All students drew the same scenario, but it is possible that, for example, students with better drawing skills or English annotations were advantaged. While not directly observed, this is an area for cautious examination in any automated scoring system – ensuring that the model isn't inadvertently favoring neat drawings over messy ones in a way that doesn't correlate to actual understanding. Our recommendation of maintaining human oversight helps mitigate high-stakes risks here, but further bias analysis would strengthen confidence in such AI systems.

For future research, an exciting direction is to implement the human-AI collaborative scoring in real classroom settings. Studies could measure how teachers interact with such systems, the time saved, and any impact on student outcomes (e.g., does faster feedback from AI improve student learning or engagement in modeling?). It would also be valuable to explore students' perceptions: are students comfortable with AI involved in assessment and does it affect how they approach tasks? From a technical perspective, future work might compare different AI approaches (e.g., CNN vs. transformer models, or the use of synthetic training data to cover rare scenarios) to determine what methods yield the best reliability and validity in scoring. Finally, expanding the rubric and model to assess the *quality* of models (beyond checking for specific features) would move closer to more holistic assessment. For example, can AI judge the coherence or completeness of a model explanation? That remains a difficult challenge, but progress in AI may eventually support it, especially if combined with human insight.

Conclusion

This study demonstrates that AI-assisted assessment of student-drawn scientific models is not only feasible but can achieve a level of consistency comparable to human scoring, even on complex, open-ended tasks. By systematically comparing human and CNN-based scoring on a 13-category rubric, we found that the AI can mirror human judgments in many areas and that discrepancies, when they occur, carry meaning – highlighting either limitations of the AI or ambiguities in the task that can be addressed. The involvement of AI offers practical benefits in efficiency, yet our findings reinforce that the best outcomes arise from a thoughtful integration of human expertise and AI efficiency. In an era where educational AI is rapidly advancing, we provide an example of harnessing that technology to enhance assessment while upholding the principles of fairness, transparency, and pedagogical soundness. We hope this work serves as a foundation for further innovation in assessment practices, enabling educators to more readily use rich modeling and other performance tasks to deepen science learning, confident that they have reliable tools to assess student understanding. The partnership of teachers and AI, as evidenced here, holds great promise for the future of learning analytics and educational measurement in STEM education.

References

Bresciani, M. J., Oakleaf, M., Kolkhorst, F., Nebeker, C., Barlow, J., Duncan, K., & Hickmott, J. (2009). Examining design and inter-rater reliability of a rubric measuring research quality across multiple disciplines. Practical Assessment, Research, and Evaluation, 14(1).

European Commission. (2021). The EU AI Act: A groundbreaking framework for AI regulation. https://commission.europa.eu

He, P., Shin, N., Kaldaras, L., & Krajcik, J. (2024). Integrating artificial intelligence into learning progression to support student knowledge-in-use: Opportunities and challenges. In Jin, H., Yan, D., & Krajcik, J. Handbook for Science Learning Progression Research. Routledge.

Jonsson, A., & Svingby, G. (2007). The use of scoring rubrics: Reliability, validity and educational consequences. Educational research review, 2(2), 130-144.

Kaldaras, L., Yoshida, N. R., & Haudek, K. C. (2022, November). Rubric development for AI-enabled scoring of three-dimensional constructed-response assessment aligned to NGSS learning progression. In *Frontiers in Education* (Vol. 7, p. 983055). Frontiers.

Kashy, D. A., Albertelli, G., Ashkenazi, G., Kashy, E., Ng, H. K., & Thoennessen, M. (2001, October). Individualized interactive exercises: A promising role for network technology. In FRONTIERS IN EDUCATION CONFERENCE (Vol. 2, pp. F1C-8).

Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). Imagenet classification with deep convolutional neural networks. Advances in neural information processing systems, 25.

Kortemeyer, G., & Nöhl, J. (2024). Assessing Confidence in AI-Assisted Grading of Physics Exams through Psychometrics: An Exploratory Study. arXiv preprint arXiv:2410.19409.

Li, T., Liu, F., & Krajcik, J. (2023) Automatically Assess Elementary Students' Hand-Drawn Scientific Models Using Deep Learning of Artificial Intelligence. Proceedings of *the Annual Meeting of the International Society of the Learning Sciences (ISLS)*.

Lu, W., & Tran, E. (2017). Free-hand Sketch Recognition Classification. CS 231N Project Report.

National Research Council. (2013). Next generation science standards: For states, by states.

Nguyen, N. (2025, February 12). From regulation to innovation: What the EU AI Act means for EdTech. FeedbackFruits. https://feedbackfruits.com/blog/what-the-eu-ai-act-means-for-edtech

Sagherian, A., Lingaiah, S., Abouelenien, M., Leong, C. W., Liu, L., Zhao, M., ... & Qi, Y. (2022, June). Learning Progression-based Automated Scoring of Visual Models. In *Proceedings of the 15th International Conference on PErvasive Technologies Related to Assistive Environments* (pp. 213-222).

von Davier, M., Tyack, L., & Khorramdel, L. (2022). Automated scoring of graphical open-ended responses using artificial neural networks. arXiv preprint arXiv:2201.01783.

Sowjanya, A. M., & Mrudula, O. (2023). Effective treatment of imbalanced datasets in health care using modified SMOTE coupled with stacked deep learning algorithms. *Applied Nanoscience*, *13*(3), 1829-1840.