Check for updates

Original Research

SAGE Open
July-September 2024: 1–15
© The Author(s) 2024
DOI: 10.1177/21582440241262864
journals.sagepub.com/home/sgo
Sage

Exploring Pre-Service Teachers' Cognitive Processes and Calibration with an Unsupervised Learning-Based Automated Evaluation System

Jiseung Yoo[1], Jisun Park[2], Minsu Ha[3], and Chelcea Mae Lagmay Darang[4]

Abstract

In the context of formative assessment in classrooms, the incorporation of automated evaluation (AE) systems and teachers' interactions with them hold significant importance. This study aimed to investigate the cognitive processes of pre-service teachers as they engaged with an AE system. We developed an unsupervised learning-based AE system, the Scoring Assistant using Artificial Intelligence (SAAI). SAAI calculates scores without relying on predefined labels and generates scientific keywords from student responses to provide constructive feedback. We collected a substantial number of constructed responses from students, and four pre-service teachers evaluated these responses initially without any external assistance and then re-evaluated them using SAAI scores as a reference point. Employing a mixed-methods approach, this study demonstrated a strong level of consistency between human raters and SAAI scores. Pre-service teachers also reflectively recalibrated their assessments and adjusted their rubrics to identify students' learning more accurately. This study highlights the practical application of AE in real classroom settings and demonstrates how AE can enhance efficiency and accuracy in K-12 science assessments, thus supporting teachers.

Keywords

automated evaluation, science education, formative assessment, unsupervised learning, calibration

Introduction

The rapid advancement of technology has led to significant improvements in the performance of automated evaluation (AE) systems, expanding their role as valuable tools for supporting formative assessments conducted during ongoing teaching and learning activities (Burrows et al., 2015; Foltz & Rosenstein, 2015). By incorporating cutting-edge techniques from natural language processing, machine learning, and deep learning, AE now provides more accurate and reliable evaluations for constructed responses, even with limited data. This has resulted in considerable time and cost savings in training and monitoring human raters (Ramesh & Sanampudi, 2022). Many AE models have been designed to handle small volumes of constructed response data from individual classrooms while maintaining low error rates, effective result analysis, and high reliability. These advancements underscore the potential of AE to enhance formative assessment, provide immediate feedback, and contribute significantly to educational settings.

While ensuring valid implementation of AE systems in real classroom contexts remains a challenge, research has emphasized the importance of teacher involvement and decision-making when integrating AE results (Brew &

[1]University of Maryland, College Park, USA
[2]Ewha Womans University, Seoul, Republic of Korea
[3]Seoul National University, Seoul, Republic of Korea
[4]Kangwon National University, Chuncheon-si, Gangwon-do, Republic of Korea

Corresponding Authors:
Jisun Park, Elementary Education Department, Ewha Womans University, 52 Ewhayeodae-gil, Seodaemun-gu, Seoul 03760, Republic of Korea.
Email: jpark29@ewha.ac.kr
Minsu Ha, Department of Biology Education, Seoul National University, 1 Gwanak-ro, Gwanak-gu, Seoul 08826, Republic of Korea.
Email: msha101@snu.ac.kr

Creative Commons CC BY: This article is distributed under the terms of the Creative Commons Attribution 4.0 License (https://creativecommons.org/licenses/by/4.0/) which permits any use, reproduction and distribution of the work without further permission provided the original work is attributed as specified on the SAGE and Open Access pages (https://us.sagepub.com/en-us/nam/open-access-at-sage).

Leacock, 2014; Noh et al., 2015). In particular, the incorporation of AE in classroom-based formative assessment necessitates context-based decision-making regarding different levels of teacher involvement (Han & Lu, 2021). Researchers have explored various models of harmoniously combining AE with human assessment practices to complement each other (Attali, 2009; Bridgeman et al., 2009; Shin et al., 2019; Williamson et al., 2012). However, much of this research has focused on the integration of human raters and AE within large-scale, high-stakes testing environments, determining the weighting of scores from both sources to compute final outcomes.

There is a need for more comprehensive research on how teachers respond to and adjust their scoring practices when integrating AE results into classroom-based formative assessment. Wilson et al. (2021) and Yuan et al. (2020) found that teachers adjusted the score results from AE with diverse objectives and approaches, such as addressing potential biases in AE scores and considering curriculum relevance. This adjustment is particularly crucial in contexts where the human rater effect is substantial or when notable discrepancies arise between AE results and teacher evaluations. Many studies have highlighted considerable variations in scoring severity, consistency, and decision-making processes among raters with differing levels of scoring experience (Attali, 2016). This means that teachers with limited assessment experience and grading expertise are more likely to exhibit a stronger rater effect and to display a wider range of dynamics and strategies in integrating the AE results into their own evaluations.

This study took a comprehensive methodological approach to carefully examine the extent to which teachers' assessment processes are integrated with the implementation of AE systems to accurately reflect student achievement. We examined the cognitive processes of teachers when utilizing AE systems within the context of science education (Bridgeman, 2013; Walsh et al., 2021). In contrast to viewing AE systems solely as tools for reducing teacher workload or sharing assessment responsibilities, our approach integrated these systems with the human raters' scoring processes to improve assessment validity by using them as valuable references (Wolfe & Wendler, 2020). By delving into the cognitive processes behind teachers' use of AE, particularly those with limited assessment and grading expertise, we aimed to shed light on how these systems can be effectively harnessed to optimize teachers' rubrics and the calibration processes.

In this study, we developed an unsupervised AE system called the Scoring Assistant Using Artificial Intelligence (SAAI) system, taking into consideration the practicality and cognitive processes of teachers using the system in real classroom contexts. Particularly in subjects like science, where assessment tends to involve evaluating evidence of students' learning and content-based features, such as specific word and sentence expressions, the system identifies important linguistic expressions in students' responses, clusters them based on extracted patterns, and conducts scoring accordingly (Basu et al., 2013). By adopting the mutual reinforcement principle of negative matrix factorization (Zha, 2002), the SAAI system we implemented in this study successfully achieved efficient summarization, scoring of student responses, and production of immediate feedback. This makes it highly adaptable to constructed responses and enables quick data processing in various assessment scenarios (Basu et al., 2013; Roy et al., 2016).

Williamson et al.'s (2012) suggestion of expanding the use of AE systems for novel-method approaches underscores the fact that the fundamental question of automated scoring for such applications is no longer "Can it be done?" but "How should it be done?" Enhancing a model's accuracy from 88% to 90% holds significance; however, it is equally crucial to investigate how teachers incorporate AE into their grading to make informed decisions and effectively address student achievement. Even less accurate models that offer greater extensibility and interpretability can be valuable if they provide useful information and insights, ultimately facilitating valid assessment practices and assisting teachers in their decision-making processes (Bauer & Zapata-Rivera, 2020; Bridgeman, 2013). By examining how teachers respond to and incorporate AE results in their classroom assessments, particularly in cases of substantial discrepancies, researchers can gain valuable insights into the complex interplay between human judgment and automated scoring systems. This understanding can inform the development of more effective and contextually aware AE systems that better support and complement teachers' professional expertise in evaluating student learning. In order to gather insights aimed at supporting teachers' assessment processes and expertise, this study focused on pre-service teachers rather than professional in-service teachers. The following specific research questions will guide this study:

1. To what extent is the SAAI system consistent with pre-service teachers' assessment results?
    a. What is the result of the correlation analysis between the pre-service teacher's first round of scoring and the SAAI's scoring?
    b. What is the result of the correlation analysis between the second round of scoring by pre-service teachers and the SAAI scoring undertaken while referencing the SAAI scoring results?

2. How do pre-service teachers use the results of the SAAI system in the second round of the scoring process?

Literature Review

AE Systems

AE systems in science education have made significant progress in effectively identifying specific content-based features in students' responses, allowing educators to examine both how and what students currently understand (Burrows et al., 2015; Liu et al., 2016; Zhai & Nehm, 2023). Proficiency in constructed response tasks goes beyond simply recognizing what students know, since students are required to construct what they understand and convey their knowledge in their own words. In this regard, AE systems have been designed to have greater capability to assess semantic-domain features that showcase a deeper grasp of the subject matter than writing styles or syntactic elements can provide (Haudek et al., 2019; Mohler et al., 2011; H.-C. Wang et al., 2008). For example, when a student explains a scientific concept, demonstrates scientific reasoning, or solves a problem, these systems analyze the response using key expressions to determine whether it contains clear and rich domain knowledge. They also can measure the relevance and coherence of ideas using numeric data mapping.

Several studies (e.g., Foltz et al., 2020; Ramesh & Sanampudi, 2022) have emphasized that AE supports student learning through immediate feedback and teacher-student interaction, which formative assessment research (e.g., Black & William, 1998; Irons, 2008) has highlighted as key strategies for promoting and supporting learning. To be specific, by leveraging natural language tools, AE enables offering students immediate formative feedback based on their own responses that allows them to gain a more concrete understanding. This promotes self-directed learning, as students continuously monitor their learning process (Duit & Treagust, 2010). Teachers can also accurately identify learners' comprehension levels, enabling meaningful interactions and providing targeted remedial activities (Lai & Chen, 2010). The integration of AE in the formative assessment approach enables teachers to tailor their instruction to cater to individual student needs, elevating the efficacy of day-to-day teaching.

AE systems use both supervised and unsupervised learning approaches to evaluate student responses (Burrows et al., 2015; Shermis & Burstein, 2013; Yan et al., 2020). However, in real-life educational contexts, particularly in K-12 environments, the use of unsupervised learning algorithms in AE holds appeal. Supervised learning relies on labeled training data, where human raters have previously scored responses, to develop a predictive model based on extracted features. In contrast, unsupervised learning algorithms operate on unlabeled data, aiming to identify hidden patterns and structures within the responses without the need for pre-existing scores (Basu et al., 2013; Colarusso, 2022). This is advantageous in classroom contexts where every task is unique, making it impractical to require ongoing teacher involvement to create labeled data, particularly in scenarios where labeled training data is limited or unavailable (Roy et al., 2016). For this reason, Rupp (2018) predicted that the methodological setup of supervised approaches may gradually disappear, and unsupervised approaches may become more common in K-12 education settings.

Relationship Between Teachers and AE Systems

It is essential to consider the teacher's role as a main actor in the evaluation process when implementing AE systems. Several studies have reported the importance of teacher involvement and system flexibility for the effective implementation of AE systems. For example, Noh et al. (2015) revealed that higher teacher involvement significantly enhanced system performance compared to relying solely on automated grading. Brew and Leacock (2014) found that the process of extracting features from students' answers and calculating scores also needs to be adapted to align with teaching and learning, such as by adjusting weights based on lesson goals, tasks, or relevant keywords, to elevate the validity and reliability of systems. In the context of formative assessment, Han and Lu (2021) found that the successful use of AE requires context-based decision-making regarding different levels of teacher involvement. They argue that, considering the goals of formative assessment, a combination of teacher and AE scoring could address the limitations of relying solely on either automated or manual assessment. This approach would enable teachers to provide timely and targeted feedback while engaging with students immediately after assessment, thus fine-tuning their feedback practices.

Some researchers explored various models of how AE can be harmoniously combined with human assessment practices to complement each other. Williamson et al. (2012) introduced implementation policies for automated scoring. One approach involves using AE as a quality control tool on human scoring by comparing scores from human evaluators and AE systems; discrepancies beyond a set threshold trigger a review by a second human evaluator, ensuring that the final score reflects a consensus or a human judgment, as seen in the GRE's scoring method (Bridgeman et al., 2009). Another approach extends to averaging scores from an AE system and a human evaluator for a more balanced assessment. In

cases where significant differences arise, a secondary human review is sought. This strategy of blending scores has been applied, for example, in the TOEFL (Attali, 2009). Recently, Shin et al. (2019) explored how AE could aid in the calibration and feedback process for human raters, suggesting the potential for AE to replace human judgment in certain monitoring aspects, particularly focusing on the consistency and severity of ratings.

However, much of the other research (e.g., Bennet, 2003; Wolfe & Wendler, 2020), including that mentioned above, has focused on the integration of human raters and AE within the context of large-scale, high-stakes testing environments, determining the weighting of scores from both sources to compute final outcomes. Studies of the deployment and analysis of AE in real classroom settings have also mainly focused on evaluating the system's performance (e.g., Liu et al., 2016; H.-C. Wang et al., 2008) and the effectiveness of its feedback (e.g., Gao & Ma, 2019; Maier et al, 2016). This has led to an understanding of AE primarily through its sole validation rather than its integrated impact with teacher-led evaluations and adjustment in daily classroom environments.

Research pertaining to the teacher-AE combination in formative assessment has been a small but growing area of inquiry. Wilson et al. (2021) found that teachers put effort into aligning the decontextualized results from AE systems to their specific instructional systems, as they encountered challenges such as overly stringent scoring or differences in scaling systems. Teachers tried to ensure the systems effectively represent the intended learning objectives for curriculum-specific and contextualized implementation plans. In addition, Yuan et al. (2020) found that the deep learning-based AE model showed that it was necessary to be cautious about bias when dealing with extremely high or low scores. They suggest that teacher score adjustment is vital for enhancing AI algorithms, particularly in interpreting nuanced aspects like the student answers' context and depth. However, these studies have demonstrated exemplary practices of teachers harnessing AE scoring. They have not comprehensively explored the patterns and features of teachers' adjustment processes or how they compare and incorporate AE results with their own scoring to ensure assessment reliability and validity. In particular, when significant discrepancies emerge in the incorporation, teachers' responses and adjustments of AE scoring results could give us insights into how AE can effectively assist teachers.

Cognitive Processes of Human Rating and the Role of AE

The use of AE not only depends on accurate overall performance but also involves a deeper understanding of human ratings, including cognitive processes and substantive grounding. Bauer and Zapata-Rivera (2020) suggested that there are two cases in which establishing the relationship between humans and AE is necessary: when the intended use depends not just on overall performance but on more detailed aspects of knowledge or skill, as in feedback for learning, and when a deeper understanding of human ratings, in terms of both cognitive processes and substantive grounding, helps researchers create more strongly grounded processes. The cognitive processes that have for years resided in raters' heads hold insight into better AE processing. Finn et al. (2020) and Walsh et al. (2021) suggested a framework of sequential activities during training and scoring, with three key elements of cognitive activities: checking scoring rubrics, creating benchmark exemplars, and calibrating scores. While Walsh et al.'s study was conducted in the context of the GRE test, where rubrics and predefined training sessions were provided, our study focused on the assessment processes and cognitive activities that arise when teachers employ the AE system to support formative assessment. For this reason, to interpret our findings, we drew upon the factors proposed by Walsh et al. but reframed them to provide valuable insights into the evaluation processes undertaken by teachers and the cognitive activities associated with them within the K-12 educational setting.

Scoring rubrics serve as the established criteria for evaluating assignments and their content. To minimize the impact of the human rater effect (Bejar, 2017), external and internal mechanisms are employed to ensure the consistency and accuracy of their scoring, aligning it with the rubric across various scoring sessions. Raters are equipped with the rubrics at the beginning of their scoring sessions or when designing the evaluation tasks. This process aims to enhance raters' understanding of the goal of evaluation and improve the accuracy and consistency of their scoring. Benchmark exemplars, sometimes referred to as anchors, introduce the characteristics of responses that correspond to specific score levels to create certification tests to demonstrate the characteristics of an essay at a particular score level (Parke et al., 2006). Benchmarks also serve as validity or monitoring samples, along with actual responses from test takers, for monitoring rater performance during operational scoring. They help establish a shared understanding among raters and promote inter-rater consistency, enhancing the overall reliability and validity of the scoring process. Calibration sessions play a critical role in ensuring that raters comprehend and internalize the scoring rubrics, enabling accurate scoring across all score points. Calibration serves the dual purpose of reinforcing the scoring criteria and acting as a gatekeeping mechanism. This process also ensures consistency in ratings across and between scoring sessions by providing a reference point for raters to align their evaluations with the

Table 1. A Detailed Overview of How the SAAI System Operates Within the Website.

Stage	Feature
Teacher uploading the students' responses	Teachers upload students' responses, which can range from one sentence to multiple paragraphs, into Excel in a specified format for analysis.
System extracting keywords	The system computes the top 10 keywords or phrases from all responses using the principle of mutual reinforcement (Zha, 2002), presenting words, their ranks, and word scores.
System generating scores for students' responses	Scores for each response are computed based on the inclusion of the top 10 keywords within each response.
System generating missing keywords from students' responses	The system identifies and generates a list of keywords missing in each student's response.
Score range determination	Score ranges are set relative to the responses for each task, presenting students' answers, word counts, scores, and missing top 10 words.

established standards. Calibration further acts as a form of mutual validation, as it mirrors the cognitive process that automatic scoring algorithms aim to replicate, contributing to the mutual validity of the scoring process.

Methods

Development of the SAAI System

The SAAI system developed in this study adopted the mutual reinforcement framework (MRF) principle proposed by Zha (2002) to summarize and cluster student responses, resulting in a central response that is subsequently scored through saliency score calculation (Ha et al., 2023). Additionally, the system can automatically generate individualized student feedback based on 10 key expressions extracted from the central response. Zha introduced a method that simultaneously extracts topic words, summarizes documents, and calculates feature scores through the application of sentence clustering and the mutual reinforcement principle. The MRF principle employs non-negative matrix factorization to generate multiple eigenvectors and corresponding singular values from the collected text, specifically sentences. This overcomes the limitations of latent semantic analysis and ensures accurate summarization. In contrast to latent semantic analysis, which can face challenges in interpreting meaning intuitively and may extract only a limited number of sentences as representatives, non-negative matrix factorization excels at summarizing and extracting sentences that exhibit the highest similarity using positive eigenvalues. This approach enables a more effective and comprehensive representation of the underlying meaning in the text. Saliency scores are derived from sentences and terms using the MRF, incorporating both word usage and document structure. Also, the SAAI system further models documents as weighted graphs and applies spectral graph clustering to generate topic groups. By utilizing these techniques, the SAAI system effectively summarizes and scores student responses, providing valuable feedback based on the central response. The SAAI system generates a conclusive score for each student's response, which is influenced by the total word count within the dataset. For this reason, the scoring range varies across different classes and datasets and the observed scores span from 0 to 37.61 in this study. The SAAI system was designed as a web-based platform. When students' responses are uploaded, the system extracts keywords and generates scores of students' responses. Table 1 provides a detailed overview of how the SAAI system operates within the website layout and the information provided to users, including teachers.

Research Context and Data Collection

As the aim of this study included evaluating the effectiveness of the SAAI system, which scores students' constructed responses in unsupervised learning, we collected students' responses that had already been collected during assessments in real classroom settings. Students' responses were collected from high schools and a university in the Philippines during the 2022 to 2023 academic year. As shown in Table 2, students' responses in seven high school classes and one university course were collected in this study. Responses from 2,176 students were collected, resulting in a dataset of 32,144 responses. The responses were collected as a part of regular school assessments rather than specifically for the purpose of this study. All datasets collected were anonymized. Table 3 shows which subjects were assessed in each class and examples of questions and students' responses. These assessments included six disciplines: physics, biology, chemistry, general science, applying science to everyday life, and philosophy of science. The sets of assessments used in this study were questions that teachers who taught the classes used in their actual classes.

To examine the effectiveness of the SAAI system and its impact on teachers' rescoring processes, we involved four pre-service teachers in their third or fourth year of college as human raters. Pre-service teachers were chosen

Table 2. Summary of Student Responses and Pre-Service Teacher Raters.

Class	Student school level	Number of students	Number of responses	Number of questions	Pre-service teacher raters
a	University	253	3,795	15	A
b	High school	289	4,332	15	
c	High school	258	3,790	15	B
d	High school	251	3,673	15	
e	High school	259	3,882	15	C
f	High school	286	4,290	15	
g	High school	259	3,885	15	D
h	High school	321	4,497	15	

Table 3. Examples of Assessment Tasks.

Rater	Class	Discipline	Examples of questions and students' responses
A	a	Biology	Q. How are lysosomes formed? Q. Why is it important to maintain water balance in the cell? A. By the fusion of transport vesicles budded from the trans Golgi network with endosomes, which contain molecules taken up by endocytosis at the plasma membrane. A. Water balance is essential for maintaining homeostasis since our cells can die if we have too much or too little water. Water will rush into our cells if we have too much water in our extracellular fluids Because we lack a cell wall, our cells can explode if too much water enters.
	b	Practical science	Q. Thinking about the topic of diabetes, write a basic science question and an applied science question that a researcher interested in this topic might ask? Q. What do you think is the most salient feature of quantitative research in comparison to qualitative research? Why? A. Diabetes affect our body health How does diabetes work in our body? How much raise in diabetes if certain goods and drinks are taken? A. Quantitative research aims to the quality of research it deals with numbers and statistic while qualitative research aims to the quality of research it deals with words and meanings.
B	c	Biology	Q. Antibiotics are medicines that are used to fight bacterial infections. These medicines kill prokaryotic cells without harming human cells. What part or parts of bacterial cell do you think antibiotics target? Why? A. I think antibiotics target all the parts of the bacterial cell because a cell has parts that are working together and the antibiotics need to target all of it for the cell and all of its parts to be killed and not cause any more bacterial infections.
	d	General science	Q. In an uncertainty of measurement, how can systematic errors be removed? Q. Distinguish among scientific fact, hypothesis, law, and theory. A. Scientific fact is the statement that keep the relationship between variable A hypothesis is the tentative explanation that can be tested by research and the theories is a explanation of observation. A. Systematic error can be removed by using calibration process in which the result of measurement recorded by an instrument is compared with the standard result of a measurement.
C	e	Chemistry	Q. Nail polish remover (acetone) is liquid at room temperature. Is acetone in the gaseous state a vapor or a gas? Explain your answer. Q. Explain the procedure(s) that can be used to separate a mixture of salt and sand. A. Acetone in the gaseous state is a vapor. This is because the acetone in a room temperature is in liquid form/state, thus it is not considered as gas but instead a vapor. A. You can separate salt and sand through evaporation. This is because the salt are mixed with water, hence when we boil them together with sand, the water with salt can evaporate and leave the sand behind.
	f	Philosophy of science	Q. Have you experienced a situation where what you believed to be true turned out to be untrue? How did this experience affect you? A. Yes, because I felt disappointed due to expecting a lot from those things and I know that sometimes believing can deceive you.
D	g	Biology	Q. Why is cancer considered a malfunction of the cell cycle? A. Because cancer cells don't respond appropriately, they grow uncontrollably and can eventually damage the tissues around them.
	h	Earth science	Q. Why is Pacific Plate called "Ring of Fire"? A. The Ring of Fire, also referred to as the Circum-Pacific Belt, is a path along the Pacific Ocean characterized by active volcanoes and frequent earthquakes.

because their assessment expertise was likely more limited than that of in-service teachers, making them an ideal group to investigate how the SAAI system can support scoring quality control and how teachers incorporate the system when discrepancies between scores are large.

The scoring process in this study was carried out through the following steps. First, the human raters and SAAI systems independently scored students' responses. Pre-service teachers developed and used a five-tiered scoring rubric for assessments. The rubric's criteria included distinct components and detailed expectations for each discipline, along with sample answers to guide grading (Appendix A). In the first scoring round, teachers used these rubrics independently. Upon receiving the SAAI scores, they adjusted and revised their rubrics for the second round of scoring. Details of these adjustments are presented in the results section. Second, the results of the SAAI system were provided to human raters as a reference. Third, human raters were asked to re-score the students' previously scored responses. During both grading sessions, human raters assigned scores on a scale ranging from 0 to 4. A score of 0 indicated that the answer was completely incorrect and irrelevant to the given question, while a score of 4 represented a highly accurate and comprehensive response that demonstrated a deep understanding of the subject matter. This scoring scale provided a clear framework for assessing the quality and alignment of the students' answers with the given questions.

Data Analysis

To address RQ1, which compares human rating and the SAAI system's scoring, Pearson's correlation analysis was used to evaluate the correlation between the scores assigned by human raters and the scores generated by the SAAI system. A baseline model that employed term frequency-inverse document frequency (tf-idf) weights served as a benchmark for comparing the performance of the SAAI system so that researchers could assess its effectiveness and improvements in automating the grading process. In addition, two variations of the SAAI system were tested: one with stopwords included in the analysis and another with stopwords removed. The results of these variations demonstrate the nuanced capabilities of the SAAI system in different operational settings.

To address RQ2, we examined the extent to which the SAAI system influenced the scoring processes of the pre-service teachers. First, a paired *t*-test was used to compare the results of the correlation coefficient between the pre-service teachers' grading and the SAAI score in the first and second rounds. This statistical analysis identified significant differences between the two rounds, and we determined the potential influence of the SAAI system on changes in the pre-service teachers' scoring. In addition, we investigated changes in the patterns of teachers' scoring when using the SAAI system. The notable changes in each rater's scoring were examined using ANOVA and Tukey's test to identify statistical differences and features in the distribution of scores. The independent variable was the level of human scoring, which consisted of five categories: 0, 1, 2, 3, and 4. The dependent variable was the SAAI score assigned to each level of human scoring. These patterns of scoring changes were then categorized into the framework of sequential activities suggested by Walsh et al. (2021), and researchers interpreted the role of the SAAI system in influencing the scoring of the pre-service teachers.

Table 4. Correlation Between Human Raters and the Three Models.

Class	tf-idf (baseline)	SAAI stopwords removed	SAAI stopwords not removed
a	.524	.544	.549
b	.184	.178	.200
c	.192	.393	.458
d	−.019	.405	.282
e	.385	.384	.391
f	−.049	.338	.524
g	.294	.553	.593
h	.020	.483	.458

Results

Consistency Between Teachers' Scoring and SAAI Scores

Table 4 shows the correlation between human raters and the three models: the tf-idf as a baseline, the SAAI system with stopword removal, and the SAAI system with stopwords included. The results show that the average correlation of the baseline model, which used tf-idf to calculate keywords' weight, was approximately $r = .17$ ($p < .033$). The SAAI system achieved an agreement of approximately .40 ($p < .049$) with human raters on the dataset, which is similar to the reported result of .50 to .60 from other systems employing various unsupervised learning algorithms (Colarusso, 2022; Jadidinejad & Mahmoudi, 2014; Mohler et al., 2011; Sultan et al., 2016). This result indicated that the SAAI outperformed the baseline model, showing a relatively strong correlation between the evaluations generated by the system and human raters.

Table 5 shows the correlation coefficient between the SAAI model with stopwords removed and two sessions of human grading. There was a significant change in the correlation coefficient (r) from .40 ($p < .049$) in the first

Table 5. Correlation Coefficient Between the SAAI Model with Stopwords Removed and Two Sessions of Human Grading.

Class	First grading	Second grading
a	.544	.651
b	.178	.514
c	.393	.430
d	.405	.427
e	.384	.621
f	.338	.477
g	.553	.838
h	.483	.822

round of scoring to .68 ($p < .009$) in the second round of scoring. In the first round, 30% of the questions showed correlation coefficients above .5, whereas in the second round this percentage increased to 70%. Overall, the results demonstrated that the pre-service teachers' scoring was comparable to the SAAI system's scoring, and the correlation coefficients between the two improved after the second round of scoring. This finding demonstrated a higher average level of agreement between the scoring outcomes produced by the SAAI system and human scoring in the second round of grading sessions.

Changes Made in the Second Scoring When Using SAAI as a Reference

Examining the difference between the raters' first- and second-round scoring and how they utilized the SAAI system in their second-round scoring, we found significant differences between the first and second round scores of raters except for Rater B. The results showed that the SAAI system functioned as a peer reviewer by supporting the raters' grading processes, which does not necessarily mean that there were many grading errors. We used a *t*-test to examine whether there were statistically significant differences in the correlation coefficients between the two scoring rounds, aiming to demonstrate the impact of SAAI on the raters' rating. Table 6 show the average and standard deviation of *t*-values and the average and standard deviation of *p*-values across 15 questions for each class. The findings indicated that there were significant differences ($p < .000$) in the correlation coefficients between the SAAI system and the two rounds of scoring, except for Rater B. Rater B did not exhibit significant changes ($p > .05$) in the correlation with the SAAI system between the first and second rounds. These findings suggest that the integration of automated scoring systems led to noticeable variations in the scoring patterns of three out of four teachers during their evaluation processes.

To gain a deeper understanding of the aforementioned results in a more comprehensive and systematic manner, we used a combined analysis involving ANOVA and Tukey's test and categorized scoring change patterns of teachers from qualitative perspectives. The patterns of scoring change were categorized based on the framework of Walsh et al. (2021), which delineates three categories: checking scoring rubrics, creating benchmark exemplars, and calibrating scores.

Scoring Change Pattern 1: Revising Rating Scales in the Rubric. One scoring change pattern found in Rater D was revising rating scales in the rubric. During the first round of scoring, Rater D graded generously, while the SAAI system graded strictly. However, Rater D revised their rating scales strictly using the SAAI result as a reference. Table 7 shows the scoring change pattern of Rater D. Initially, Rater D gave scores of 3 to around 65% of students' responses. Assigning a score of 3 to most students initially and then revising their scores to 1 indicates that the teacher initially assessed the students' work as meeting a certain level of criteria (a score of 3), but upon further reflection, the teacher determined that the students' work did not actually meet that level of criteria and deserved a lower score. Indeed, this process aligns with the findings from the SAAI analysis, which revealed that students who received a score of 3 had primarily employed basic keywords related to the questions, implying a limited depth of comprehension. A considerable number of students' responses also lacked the

Table 6. *t*-Test Between the First Scoring's Correlation Coefficient and the Second Scoring's Correlation Coefficient.

	Rater A		Rater B		Rater C		Rater D	
Class	A	b	c	d	e	f	g	h
Mean *t* value	319.07	800.73	121.97	24.00	1,490.90	199.97	64.37	383.60
St.d *t* value	854.50	511.06	381.85	31.84	1,738.98	134.59	92.76	390.88
Mean sig.	.04**	.02**	.28	.34	.07*	.06*	.00***	.00***
St.d sig.	.10	.05	.33	.20	.25	.22	.00	.00

*$p < .05$. **$p < .01$. ***$p < .001$.

Table 7. Rater D's Grading Frequency Distribution.

	Initial scoring/first grading: Score frequency per question (%)					Revised scoring/second grading: Score frequency per question (%)				
	Score					Score				
Question no.	0	1	2	3	4	0	1	2	3	4
1–5, 7–15	1.78	5.59	25.17	65.97	1.70	5.26	70.19	13.62	10.44	3.40
6		5.47	29.19	64.59	0.73	3.65	25.91	62.04	7.66	0.73

Table 8. Example of Students' Answers, SAAI Scores, and Rater D's Scores on Question 5.

Student ID	5. Why is cancer considered a malfunction of the cell cycle?	SAAI score	First scoring	Second scoring
145	Cells divide uncontrollably and spread into surrounding tissues.	0.4282	2	1
28	It is a product of unchecked and excessive division.	0.6276	3	1

inclusion of the top 10 keywords. Despite Rater D initially assessing the students' responses as reasonably appropriate, the subsequent analysis from the SAAI system indicated that their answers did not adequately demonstrate a comprehensive grasp of the learning content.

Rater D seemed to revise the rating scale of their rubric using the SAAI results as a reference point. This adjustment increased discrimination in grading, allowing for more precise differentiation of student performance levels and enhancing rating severity. For instance, in Question 5 "Why is cancer considered a malfunction of the cell cycle?" during the first round of scoring, Rater D gave a score of 2 to a student's response of "Cells divide uncontrollably and spread into surrounding tissues" and gave a score of 3 to the response "It is a product of unchecked and excessive division." However, in the second round of scoring, both responses were re-scored as 1. As the SAAI system highlighted keywords such as gene, DNA, and control, the SAAI system gave a relatively higher score to the responses in including the concept of breakdown. Rater D revised the rubric to ensure that it captured the desired responses and accurately represented the expectations for each score level. This adjustment was based on assessing the extent to which individual responses contained the keywords highlighted in the feedback word list accompanying the SAAI results (Table 8).

Scoring Change Pattern 2: Changing Benchmark Responses and Criteria in the Rubric. In the scoring patterns of Rater A, scoring changes were observed in the benchmarks and example responses within the rubric. It is noteworthy that Rater A did not make overall changes to the entire scoring, but rather adjusted the benchmarks for specific responses, resulting in changes to the scores of corresponding answers. As mentioned above, Rater D made comprehensive adjustments to the scoring scale of the rubric, applying a more severe grading approach across all students' responses. The difference lay in the fact that Rater A only modified the scoring of specific responses based on the applicable benchmarks. These changes were directly influenced by Rater A's revision of the criteria and benchmarks in their rubric, guided by the insights provided by the SAAI system.

Notable differences in the frequency of scores for Question 1 are an explicit example of this explanation. In this specific question, students were tasked with providing a binary response, selecting either "yes" or "no" and providing a concrete explanation to support their choice. In the first round of grading, teachers assigned 0 points for incorrect "yes" responses, while a brief "no" answer received 1 point. However, in the second round of grading, teachers acknowledged the significance of not only providing the correct response but also offering a more elaborate explanation using relevant keywords. As a result, higher scores were awarded to students who answered "no" and provided a more comprehensive and keyword-aligned explanation, as indicated by the SAAI system. This adjustment was aimed at capturing the depth of understanding and analytical skills demonstrated by the students in their responses by changing benchmark responses (Table 9).

Scoring Change Pattern 3: Micro-Calibration in Inconsistent Scoring. Rater C made micro-calibration in their second scoring, which contributed to a more consistent and accurate assessment of student responses. During their

Table 9. Example of Students' Answers, SAAI Scores, and Rater A's Scores on Question 1.

Student ID	1. Do you consider viruses to be living things?	SAAI score	First scoring	Second scoring
197	No, because viruses are not living things. Viruses are complicated assemblies of molecules, including proteins, nucleic acids, lipids, and carbohydrates, but on their own they can do nothing until they enter a living cell. Without cells, viruses would not be able to multiply.	3.753	4	4
162	No. A virus is a parasitic intracellular organism with a subcellular life cycle. A cell's replication apparatus is necessary for replication. Nothing about its metabolism exists outside of the host cell.	2.423	4	3
89	No	0	1	0
94	I don't consider viruses as living things.	0	1	0
35	Yes, viruses contain bacteria and bacteria is classified as microorganisms.	0	0	0
51	Yes	0	0	0

Table 10. Example of Students' Answers, SAAI Scores, and Rater C's Scores on Question 8.

Student ID	8. Describe a situation in your life when you were able to engage in philosophy	SAAI score	First scoring	Second scoring
205	My example of dilemma that brought about my need to philosophize is when I have to choose between joining the revolution to promote alternative learning schools in …the insensitivity and apathy of people in society. Thus, this is an example of philosophizing.	8.065	2	3
280	My example of dilemma that brought about my need to philosophize is when I have to choose between joining the revolution to promote …The roof at chaos in our community these days is the insensitivity and apathy of people in society.	7.957	3	3
56	My example of dilemma that brought about my need to philosophize is when I have to choose between joining the revolution to promote …..is the empathy of communities. Thus, this is an example of philosophizing.	7.243	1	3

second grading session, they reviewed both students' answers and their first grading with the SAAI score. They also checked the feedback of the SAAI to determine which and how many of keywords each student had missed. This allowed them to revise their scoring approach and ensure greater consistency and fairness in evaluating individual responses.

This issue was particularly evident in Question 8, which required students to provide examples from the lesson or real-life scenarios to demonstrate their understanding. As presented in Table 10, Rater C's first-round scores revealed variations in the scores assigned to almost identical responses from Students 205, 208, and 56, with scores of 2, 3, and 1, respectively. The SAAI scores indicated that there were no significant differences in the missing keywords across a certain range of the students' responses, prompting Teacher C to recalibrate their scoring approach and make necessary adjustments. Although Rater C's initial evaluation of the responses exhibited more inconsistencies and discrepancies, they were able to elevate the reliability and validity of the overall scoring process through the second scoring session.

Discussion

Our aim in this investigation was to provide empirical evidence supporting teachers in their assessment roles involving AE systems, while also advancing the feasible integration of these systems within K-12 educational settings. We sought to examine the impact of AE systems and how teachers utilize them in their assessment practices, with the understanding of teachers' cognitive processes of using AE, ultimately enhancing the overall quality of assessment. As part of our study, we analyzed the scoring of student data from the Philippines using both human raters and the SAAI system. It is important to note that the scoring process involved raters who had limited expertise in assessment. The scoring was conducted in two rounds, with the second round incorporating the SAAI system as a reference.

Based on the research questions, the findings of this study can be summarized as follows: First, we observed a significant level of consistency in the scoring outcomes between teacher ratings and the SAAI system. The SAAI model, developed using Zha's (2002) MRF algorithm for

unsupervised learning, demonstrated superior performance compared to the baseline tf-idf model. Out of the 120 items, approximately 60 items had correlation coefficients exceeding .6, which is a level similar to previous studies on power-grading data (Basu et al., 2013) and unsupervised models (Colarusso, 2022; Jadidinejad & Mahmoudi, 2014; Sultan et al., 2016), which reported correlation coefficients ranging from .5 to .6.

Second, the findings demonstrated the distinctive cognitive processes that emerge during the integration of AE and human scoring. In the second round of scoring, where SAAI scores were used as a reference, there was a noteworthy increase in the correlation between the teachers' assessments and the SAAI scores, except for Rater B. The average correlations coefficient in the second round stood at .7, showcasing the significance of this change. This alteration was statistically meaningful as per the *t*-test, implying that the incorporation of SAAI scores in the second round had an impact. Rater D revised the scoring scale of their rubric and Rater A adjusted the criteria of their rubric, aiming to align with intended learning outcomes and accurately reflect the expectations for student learning. These changes exemplified how teachers integrated their assessment practices with the AE system, guided by the SAAI-generated scores and the degree to which individual responses contained the keywords highlighted in the feedback word list accompanying the SAAI results. In addition, Rater C recalibrated their scoring and made overall adjustments to ensure greater consistency and fairness in assessing students' responses. They leveraged the SAAI score as quantitative data to identify and rectify errors and discrepancies between their assessments and the AE system. Although these three patterns of scoring adjustment were somewhat mixed among the three raters, each teacher exhibited a dominant pattern. For instance, in one of the tasks, Rater A recalibrated some students' scores, but less frequently than Rater C, who did not change the scales overall.

These findings suggest that the quantitative data generated by SAAI helped teachers in obtaining a clearer understanding of the level of achievement within their classrooms. The effectiveness of SAAI in promptly capturing the characteristics of student responses and accurately reflecting them through scores and feedback serves as valuable evidence to support the teacher's decision-making processes. Evidence-centered design, as introduced by Mislevy et al. (2006), supports this with a theoretical framework. This theory particularly expands on the intricate cognitive models and processes that teachers employ while utilizing AE, especially in constructed response tasks. This model can be used to explain that the response scoring process and the summary scoring process are particularly noteworthy, as they explain the strategies of teachers using AE. For instance, when it comes to evaluating constructed responses, relying solely on an observable binary of "right" or "wrong" is no longer sufficient. Using specific measurements to show how well students understand a topic and providing teachers with accurate data are vital parts of AE. On the other hand, the summary scoring stage involves a more complex process. In this process, human raters not only determine the extent to which their assumptions about a student's learning need to be shifted based on the evidence but also assign weights to the variables that contribute to the final score. It becomes clear that while human raters demonstrate active engagement in high-order cognitive processes during the summary scoring process, the "identification" activity—focused on quantifying evidence objectively—takes on a more substantial role, thus underscoring the increased importance of utilizing AE systems.

This study demonstrates the implication of unsupervised learning algorithms in a K-12 context. Particularly, when considering the evaluation of constructed responses within the context of daily formative assessment, the importance of unsupervised learning becomes evident, offering a means to circumvent the pitfalls of the "teach to test" approach (Williamson et al., 2012). Raters C and D revised their rubrics to reflect whether students exhibited a precise understanding of the question requirements in their constructed responses. Through the report from the SAAI with targeted keywords and distinctive features, teachers were able to adeptly evaluate both the comprehension of students and the areas where misconceptions may arise, all in one comprehensive assessment, as highlighted in *Knowing What Students Know* (NRC, 2001). In addition, these unsupervised techniques, known for their adeptness in capturing essential features and identifying evidence, exhibit remarkable versatility, as they demand little to no modification when confronted with a range of tasks and formats. Indeed, the scope of questions examined within this study spans from the philosophy of science to biology, showcasing the system's seamless and intuitive processing of students' constructed responses.

The results also highlight the potential of AE systems as a valuable tool for teachers, assisting them to enhance their assessment expertise and make well-informed scoring decisions. Rather than attempting to adjust scores from individual teachers, especially if these adjustments might vary from day to day, a potentially fairer procedure is to make early identification of raters who are too lenient or too severe and support them to grade more appropriately.

Promoting teachers' active reflection and ensuring consistent, accurate assessments. By providing teachers with opportunities to review their scoring practices in light of AE feedback and engaging them in meaningful

discussions about assessment standards and best practices, educational institutions can foster a culture of continuous improvement and enhance the overall quality of assessments. This requires real-time tracking and monitoring of rater behavior. Such tracking also addresses a related problem: the tendency of human raters to avoid the extremes of the score scale by overscoring low-quality responses and underscoring high-quality answers (Myford & Wolfe, 2009). Shin et al. (2019) shed light on these dynamics, revealing that individual raters underwent notable accuracy improvements and/or adjustments in their categorization over time, all facilitated by the instant incorporation of AE.

Our research highlights the immense potential of utilizing AE as a peer reviewer, a highly adaptable approach that can be applied to various local contexts worldwide. By fostering consistency and minimizing rater effects on assessment, AE serves as a powerful tool for the global educational community (Han & Lu, 2021; Z. Wang & von Davier, 2014). While specific implementation examples may differ based on subjects, standards, or test policies across nations, the core operation of AE remains a harmonized approach that can be adopted to various local contexts worldwide. Consequently, the methodologies presented in this study provide valuable insights into the prospective application of AE for rater monitoring. This research paves the way for future investigations and applications within the field of educational assessment, offering promising solutions to challenges faced by educators and policymakers worldwide in ensuring fair, reliable, and consistent evaluation practices.

In addition, the case of Rater B, who showed no significant adjustment changes in second-round scoring despite the incorporation of AE, lead us to consider practical implications and the broader issue of variability in technology adoption among educators. Zhai and Nehm (2023) emphasized the growing prevalence of AI in education and the necessity of tailoring AE systems to meet the diverse needs and expectations of individual teachers. The finding from the Rater B scenario underscores the potential for AE to reveal distinct rater behaviors but also stresses the importance of tailoring AE implementation to meet the unique needs of each teacher. While the reasons behind Rater B's resistance to adapting AE remain unspecified, the development and implementation of AE can focus on the support mechanisms for teachers who may be hesitant or unsure of how to integrate these systems into their teaching methodologies. The ultimate goal is making AE a useful, accessible, and beneficial tool for all teachers, thereby enhancing the overall quality of formative assessments.

This study has both limitations and recommendations for future research. First, the SAAI developed in this research generated task-level scores based on the word statistics embedded within the dataset, resulting in score variations influenced by responses. As such, it becomes imperative to explore methods of standardizing these scores. Moreover, to attain a more comprehensive understanding of the cognitive processes of teachers, there is potential for additional qualitative research, incorporating techniques such as interviews and observations. The case of Rater B, for example, who exhibited consistent grading during the second round, calls for a comprehensive analysis of the underlying mechanism that contributed to this consistency and an exploration of the specific ways in which AE is not integrated into certain contexts. Moving forward, the study could be further enhanced by delving into the intricacies between AE and teachers. This exploration would further uncover models that maximize the support for teachers using AE in their classrooms, thereby elevating the overall validity and reliability of the assessment process.

Appendix A. Test Scoring Rubric for General Science tasks

Points	Criteria
4	The response provides all aspects of a complete interpretation and/or a correct solution. The response thoroughly addresses the points relevant to the concept or task. It provides strong evidence that information, reasoning, and conclusions have a definite logical relationship. It is clearly focused and organized, showing relevance to the concept, task, or solution process. *Sample Response: "List the steps of the scientific method. Briefly explain each one."* 1. Ask a question: Ask a question about something that you observe: How, what, when, who, which, why, or where? 2. Do background research: Use library and internet research to help you find the best way to do things. 3. Construct a hypothesis: Make an educated guess about how things work. 4. Test your hypothesis: Do an experiment. 5. Analyze your data and draw a conclusion: Collect your measurements and analyze them to see if your hypothesis is true or false. 6. Communicate your results: Publish a final report in a scientific journal or by presenting the results on a poster.

(continued)

Appendix A. (continued)

Points	Criteria
3	The response provides the essential elements of an interpretation and/or a solution. It addresses the points relevant to the concept or task. It provides ample evidence that information, reasoning, and conclusions have a logical relationship. It is focused and organized, showing relevance to the concept, task, or solution process. *Sample Response: "List the steps of the scientific method. Briefly explain each one."* 1. Ask a question. 2. Do background research: Use library and internet research. 3. Construct a hypothesis: An educated guess about how things work. 4. Test your hypothesis: Do an experiment. 5. Analyze your data and draw a conclusion. 6. Communicate your results.
2	The response provides a partial interpretation and/or solution. It somewhat addresses the points relevant to the concept or task. It provides some evidence that information, reasoning, and conclusions have a relationship. It is relevant to the concept and/or task, but there are gaps in focus and organization. *Sample Response: "List the steps of the scientific method. Briefly explain each one."* 1. Ask a question. 2. Do background research. 3. Construct a hypothesis. 4. Test your hypothesis. 5. Analyze your data and draw a conclusion. 6. Communicate your results.
1	The response provides an unclear, inaccurate interpretation and/or solution. It fails to address or omits significant aspects of the concept or task. It provides unrelated or unclear evidence that information, reasoning, and conclusions have a relationship. There is little evidence of focus or organization relevant to the concept, task, and/or solution process. *Sample Response: "List the steps of the scientific method. Briefly explain each one."* Ask a question, hypothesis, do an experiment, analyze your data.
0	The response does not meet the criteria required to earn one point. The student may have written on a different topic or written "I don't know."

Declaration of Conflicting Interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: This work was supported by the Ministry of Education of the Republic of Korea and the National Research Foundation of Korea (NRF-2021S1A5A2A03061991)

ORCID iDs

Jiseung Yoo https://orcid.org/0000-0002-2613-6446
Jisun Park https://orcid.org/0000-0002-7740-2046
Minsu Ha https://orcid.org/0009-0006-6916-390X

Data Availability Statement

Data sharing not applicable to this article as no datasets were generated or analyzed during the current study.

References

Attali, Y. (2009, April). *Evaluating automated scoring for operational use in consequential language assessment—The ETS experience* [Paper presentation]. Annual Meeting of the National Council on Measurement in Education, San Diego, CABasu.

Attali, Y. (2016). A comparison of newly-trained and experienced raters on a standardized writing assessment. *Language Testing*, *33*(1), 99–115.

Basu, S., Jacobs, C., & Vanderwende, L. (2013). Powergrading: A clustering approach to amplify human effort for short answer grading. *Transactions of the Association for Computational Linguistics*, *1*, 391–402. https://doi.org/10.1162/tacl_a_00236

Bauer, M., & Zapata-Rivera, D. (2020). Cognitive foundations of automated scoring. In D. Yan, A. A. Rupp, & P. W. Foltz (Eds.), *Handbook of automated scoring: Theory into practice* (pp. 13–28). Taylor and Francis Group/CRC Press.

Bejar, I. I. (2017). Threats to score meaning in automated scoring. In K. Ercikan, & J. W. Pellegrino (Eds.), *Validation of score meaning for the next generation of assessments* (pp. 75–84). Routledge.

Bennett, R. E. (2003). Online assessment and the comparability of score meaning (ETS-RM-03-05). Princeton, NJ: Educational Testing Service.

Black, P., & William, D. (1998). Assessment and classroom learning. *Assessment in Education: Principles, Policy & Practice*, *5*(1), 7–75.

Brew, C., & Leacock, C. (2014). Automated short answer scoring. In M. D. Shermis, & J. Burstein (Eds.), *Handbook of*

automated essay evaluation: Current applications and new directions (pp. 136–152). Routledge.

Bridgeman, B. (2013). Human ratings and automated essay evaluation. In M. D. Shermis, & J. Burstein (Eds.), *Handbook of automated essay evaluation: Current applications and new directions* (pp. 221–232). Routledge.

Bridgeman, B., Trapani, C., & Attali, Y. (2009, April). *Considering fairness and validity in evaluating automated scoring* [Paper presentation]. Annual Meeting of the National Council on Measurement in Education, San Diego, CA.

Burrows, S., Gurevych, I., & Stein, B. (2015). The eras and trends of automatic short answer grading. *International Journal of Artificial Intelligence in Education*, *25*(1), 60–117. https://doi.org/10.1007/s40593-014-0026-8

Colarusso, D. (2022). *Unsupervised machine scoring of free response answers: Validated against law school final exams*. MIT Computational Law Report. https://law.mit.edu/pub/unsupervised-machine-scoring-of-free-response-answers/release/1

Duit, R., & Treagust, D. F. (2010). Conceptual change: A powerful framework for improving science teaching and learning. *International Journal of Science Education*, *25*(6), 671e688. http://doi.org/10.1080/09500690305016

Finn, B., Arslan, B., & Walsh, M. (2020). Applying cognitive theory to the human essay rating process. *Applied Measurement in Education*, *33*(3), 223–233. https://doi.org/10.1080/08957347.2020.1750405

Foltz, P. W., & Rosenstein, M. (2015, March). *Analysis of a large-scale formative writing assessment system with automated feedback* [Conference session]. *Proceedings of the Second (2015) ACM Conference on Learning@ Scale* (pp. 339–342).

Foltz, P. W., Yan, D., & Rupp, A. A. (2020). The past, present, and future of automated scoring. In D. Yan, A. A. Rupp, & P. W. Foltz (Eds.), *Handbook of automated scoring: Theory into practice* (pp. 1–10). Taylor and Francis Group; CRC Press.

Gao, J., & Ma, S. (2019). The effect of two forms of computer-automated metalinguistic corrective feedback. *Language Learning & Technology*, *23*(2), 65–83. https://doi.org/10125/44683

Ha, M., Choi, J., Lim, D., Choi, S., Choo, J., Shin, C. K., & Park, J. (2023). Development of an unsupervised learning-based automated evaluation system of descriptive assessment. *Brain, Digital, & Learning*, *13*(4), 339–351. https://doi.org/10.31216/BDL.20230021

Han, C., & Lu, X. (2021). Interpreting quality Haudek assessment re-imagined: The synergy between human and machine scoring. *Interpreting and Society*, *1*(1), 70–90. https://doi.org/10.1177/27523810211033670

Haudek, K. C., Santiago, M., Wilson, C. D., Stuhlsatz, M., Donovan, B., Buck-Bracey, Z., Gardner, A., Osborne, J., & Cheuk, T. (2019, April). *Using automated analysis to assess middle school students' competence with scientific argumentation* [Conference session]. National Conference on Measurement in Education. Annual Conference, Toronto, ON, United States.

Irons, A. (2008). *Enhancing learning through formative assessment and feedback* (Series: Key Guides for Effective Teaching in Higher Education). Routledge.

Jadidinejad, A. H., & Mahmoudi, F. (2014). Unsupervised short answer grading using spreading activation over an associative network of concepts. *Canadian Journal of Information and Library Science*, *38*(4), 287–303.

Lai, A., & Chen, D. (2010). Web-based two-tier diagnostic test and remedial learning experiment. *International Journal of Distance Education Technologies*, *8*(1), 31e53

Liu, O. L., Rios, J. A., Heilman, M., Gerard, L., & Linn, M. C. (2016). Validation of automated scoring of science assessments. *Journal of Research in Science Teaching*, *53*(2), 215–233.

Maier, U., Wolf, N., & Randler, C. (2016). Effects of a computer-assisted formative assessment intervention based on multiple-tier diagnostic items and different feedback types. *Computers & Education*, *95*, 85–98. https://doi.org/10.1016/j.compedu.2015.12.002

Mislevy, R. J., Steinberg, L., Almond, R. G., & Lucas, J. F. (2006). Concepts, terminology, and basic models of evidence-centered design. In D. M. Williamson, R. J. Mislevy, & I. I. Bejar (Eds.), *Automated scoring of complex tasks in computer-based testing* (pp. 49–82). Lawrence Erlbaum.

Mohler, M., Bunescu, R., & Mihalcea, R. (2011). *Learning to grade short answer questions using semantic similarity measures and dependency graph alignments* [Paper presentation]. *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies* (pp. 752–762). https://aclanthology.org/P11-1076

Myford, C. M., & Wolfe, E. W. (2009). Monitoring rater performance over time: A framework for detecting differential accuracy and differential scale category use. *Journal of Educational Measurement*, *46*(4), 371–389. https://doi.org/10.1111/j.1745-3984.2009.00088.x

Noh, E., Sung, K., & Lim, E. (2015). Analysis of scoring results in Korean constructed response automatic scoring program. *Journal of Educational Evaluation*, *28*(2), 523–551.

NRC. (2001). *Knowing what students know: The science and design of educational assessment*. National Academy Press.

Parke, C. S., Lane, S., & Stone, C. A. (2006). Impact of a state performance assessment program in reading and writing. *Educational Research and Evaluation*, *12*, 239–269. https://doi.org/10.1080/13803610600696957

Ramesh, D., & Sanampudi, S. K. (2022). An automated essay scoring system: A systematic literature review. *Artificial Intelligence Review*, *55*(3), 2495–2527. https://doi.org/10.1007/s10462-021-10068-2

Roy, S., Dandapat, S., Nagesh, A., & Narahari, Y. (2016). *Wisdom of students: A consistent automatic short answer grading technique* [Conference session]. *Proceedings of the 13th International Conference on Natural Language Processing* (pp. 178–187). https://aclanthology.org/W16-6324

Rupp, A. A. (2018). Designing, evaluating, and deploying automated scoring systems with validity in mind: Methodological design decisions. *Applied Measurement in Education*, *31*(3), 191–214. https://doi.org/10.1080/08957347.2018.1464448

Shermis, M. D., & Burstein, J. (Eds.). (2013). *Handbook on automated essay evaluation: Current applications and new directions*. Routledge.

Shin, H. J., Wolfe, E., & Wilson, M. (2019). Human rater monitoring with automated scoring engines. *Psychological Test and Assessment Modeling*, *61*(2), 127–148.

Sultan, M. A., Salazar, C., & Sumner, T. (2016). *Fast and easy short answer grading with high accuracy* [Conference session]. *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies* (pp. 1070–1075). https://doi.org/10.18653/v1/N16-1123

Walsh, M. M., Arslan, B., & Finn, B. (2021). Computational cognitive modeling of human calibration and validity response scoring for the Graduate Record Examinations (GRE). *Journal of Applied Research in Memory and Cognition*, *10*(1), 143–154. https://doi.org/10.1016/j.jarmac.2020.08.012

Wang, H.-C., Chang, C.-Y., & Li, T.-Y. (2008). Assessing creative problem-solving with automated text grading. *Computers & Education*, *51*(4), 1450–1466.

Wang, Z., & von Davier, A. A. (2014). Monitoring of scoring using the e-rater® automated scoring system and human raters on a writing test. *ETS Research Report Series*, *2014*(1), 1–21.

Williamson, D. M., Xi, X., & Breyer, F. J. (2012). A framework for evaluation and use of automated scoring. *Educational Measurement: Issues and Practice*, *31*(1), 2–13. https://doi.org/10.1111/j.1745-3992.2011.00223.x

Wilson, J., Huang, Y., Palermo, C., Beard, G., & MacArthur, C. A. (2021). Automated feedback and automated scoring in the elementary grades: Usage, attitudes, and associations with writing outcomes in a districtwide implementation of MI write. *International Journal of Artificial Intelligence in Education*, *31*(2), 234–276. https://doi.org/10.1007/s40593-020-00236-w

Wolfe, E. W., & Wendler, C. L. (2020). Why should we care about human raters? *Applied Measurement in Education*, *33*(3), 189–190.

Yan, D., Rupp, A. A., & Foltz, P. W. (Eds.). (2020). *Handbook of automated scoring: Theory into practice*. Taylor & Francis/CRC Press.

Yuan, S., He, T., Huang, H., Hou, R., & Wang, M. (2020). Automated Chinese essay scoring based on deep learning. *Computers, Materials & Continua*, *65*(1), 817–833. https://doi.org/10.32604/cmc.2020.010471

Zha, H. (2002, August). *Generic summarization and keyphrase extraction using mutual reinforcement principle and sentence clustering* [Conference session]. *Proceedings of the 25th annual International ACM SIGIR Conference on Research and Development in Information Retrieval* (pp. 113–120).

Zhai, X., & Nehm, R. H. (2023). AI and formative assessment: The train has left the station. *Journal of Research in Science Teaching*, *60*(6), 1390–1398.