Brain, Digital, & Learning

pISSN : 2384-2474
eISSN : 2586-7490

RESEARCH ARTICLE

Science Teachers' Perception of Automated Scoring Scientific Argumentation in a Classroom

Suna Ryu*
Korea National University of Education

*Corresponding Author: sunaryu@knue.ac.kr

ABSTRACT

This study examines science teachers' perceptions of automated scoring systems (AS) in teaching scientific reasoning in the classroom. Using a combination of diffusion theory and digital assessment literacy as an analytical framework, focus group interviews were conducted to explore teachers' attitudes, understandings, and challenges with AS. Findings suggest that teachers primarily used AS as a supplemental tool rather than a primary assessment method, relying on its insights as an additional reference throughout the course. Teachers found that understanding the concepts underlying machine learning algorithms not only increased their confidence in AS, but also inspired new approaches to integrating AI with science content. A key contribution of this study is its detailed examination of teachers' first-hand experiences with AI-enhanced learning systems, which provides insights into how AS can be effectively incorporated into science education. The findings contribute to the design of supportive AI-enhanced assessment environments that are aligned with teachers' instructional goals and pedagogical values.

Key words: Teacher perception, automated scoring, diffusion theory, digital data literacy, machine learning

Check for updates

OPEN ACCESS

Brain, Digital, & Learning
2024, Vol. 14, No. 4, 559-575

https://doi.org/10.31216/BDL.20240032

Received: November 12, 2024
Revised: November 23, 2024
Accepted: November 25, 2024

Introduction

Automated Scoring (AS) is gaining recognition in educational research as the rapid development of Artificial Intelligence (AI) enables AS to reduce the time and cost of scoring while providing higher accuracy (Lee et al., 2019; Lee & Ryu, 2020; Shin et al., 2021; Wilson et al., 2021). In addition, AS provides immediate results and feedback throughout the learning process, allowing teachers and students to monitor and revise their teaching and learning (Lee et al., 2019; Lee & Ryu, 2020; Shin et al., 2021; Wilson et al., 2021). Despite the increasingly recognized potential of using AS, little is known about how teachers use and think about AS in the classroom (Shin et al., 2021).

Scientific argumentation has been a major focus of AS because scientific argumentation relies heavily on logical reasoning, evidence-based claims, and structure discourse(Berland &

McNeill, 2010; McNeill & Krajcik, 2012; Osborne et al., 2004; Ryu et al., 2015; Ryu & Sandoval, 2012, 2015). These characteristics make it particularly suitable for AS, as the algorithm can objectively evaluate components such as claim quality, evidence, and reasoning patterns. The type of argument that incorporates personal experience and/or cultural norms requires much more nuanced and difficult standards to evaluate (Haudek & Zhai, 2023).

While some studies have explored teachers' perceptions of AI-based learning systems like AS, these studies have primarily focused on introductory and hypothetical questions, reflecting teachers' relatively limited experience with AI-based learning technology like AS. Therefore, more research is needed to deeply explore teachers' perceptions based on their experiences and understanding of AS.

Understanding teachers' perceptions can contribute to improving educational research in this area because teachers' perceptions influence the way these technologies are used in the classroom (Cope & Ward, 2002). In addition, teachers are known to be reluctant to use new AI-enhanced technologies, especially when there is conflict, confusion, and disagreement between the teacher and the AI about goals, rules, and instructional strategies (da Rosa, 2024; Lin, 2022; Nebel et al., 2016; Tan et al., 2022). The present study explored teachers' perceptions of automated assessment as they learned and experienced AS for scientific argumentation. Specifically, the study also examined how teachers' conceptual understanding of the automated scoring model's algorithm influenced the use of AS as a classroom assessment.

To explore teachers' perceptions of Automated Scoring, an AI-assisted learning system, we draw on Rogers' innovative diffusion theory model as an analytical framework (Rogers, 2003). He proposed five factors (relative advantage, complexity, compatibility, trialability (experience), and observability (perceived trust) that influence the degree of adoption of innovative technologies (Rogers, 2003). Relative advantage suggests that users prefer to adopt and use innovative technologies over existing technologies when they have advantages over existing technologies. Complexity refers to the degree of difficulty in the operation and usefulness of innovative technologies. In the context of AS in science classrooms, we have addressed the difficulties, confusion, and conflicts that teachers face when dealing with AS. He defines compatibility as the degree to which an innovative product itself conforms to existing social norms, values, culture, past experiences, knowledge, needs, and desires. In the context of AS in science classrooms, compatibility refers to the extent to which teachers believe that the use of AS is consistent with instructional goals and strategies. In terms of experience, the study asked teachers about specific features of AS that they either liked or disliked. Regarding perceived trust, this study asked about the degree of teachers' willingness and intention to adopt AS in their science classrooms.

Because Roger's model is not specific to instructional purposes, the study also referred to digital assessment literacy (Eyal, 2012) to explore teachers' conceptions of science instruction and assessment using AS. The notion of digital assessment literacy emphasizes the need for teachers to evaluate the possibilities and capabilities of digital learning environments, especially as assessors who develop assessment items and rubrics in digital environments and understand the validation of collected data. More specifically for AI, the concept of AI literacy points to the need for both teachers and students to understand the basic concepts of AI, including the connections between AI applications and underlying concepts (Ng et al., 2021). AI literacy also emphasizes that teachers should not only use appropriate AI-based learning adaptation systems, but also introduce the application of AI concepts to relevant content. With reference to these two frameworks, this study explored the conceptual understanding of machine learning mechanisms for AS, the coordination between AS outcomes and curriculum contents/standards.

The specific research questions guiding this study are as follows:

1. How do science teachers perceive the use of AS in science instruction?
2. How does conceptual understanding of AS help science teachers develop ideas for incorporating AS into their science teaching and assessment?

Theoretical Framework

Applying Rogers' Innovation Diffusion Theory as an Analytical Framework for Science Teachers' Perceptions of Automated Scoring

Rogers' Innovation Diffusion Theory (IDT) (Rogers, 2003) provides a structured approach for examining how science teachers perceive and potentially adopt automated scoring (AS) technologies in their classrooms. By focusing on Rogers' five key factors-relative advantage, complexity, compatibility, trialability, and observability-this model provides a detailed framework for analyzing the conditions that either promote or inhibit teacher adoption of AS. Using this framework, researchers can identify the particular challenges and motivators that shape science teachers' attitudes toward AS, providing insights that can inform the design of targeted support and professional development.

Rogers' Innovation Diffusion Theory (IDT) is well suited for analyzing the adoption of new technologies, including automated scoring (AS), in science classrooms. For example, previous studies highlight its applicability in educational contexts, particularly in exploring teacher perceptions and technology integration. Shin et al. (2021) examined science teachers' acceptance of AI tools, emphasizing factors such as perceived usefulness and compatibility with existing practices. Studies by Lin (2022) and Wilson et al. (2021) further demonstrate how innovations such as AI-enhanced assessment systems align with IDT constructs, including complexity and relative advantage, to explain their integration into the classroom. These findings provide strong support for IDT as a framework for understanding teacher perceptions and adoption processes in the context of AS in science education.

Relative advantage refers to teachers' perceptions that AS offers superior benefits over traditional methods, particularly in terms of efficiency, accuracy, and immediate feedback. Teachers may value AS for saving grading time, providing accurate scoring of complex responses, and supporting formative assessment through real-time instructional adjustments. Understanding these perceived benefits helps determine whether teachers view AS as supplementing or surpassing traditional assessments. Complexity refers to the perceived difficulty of using AS, which may hinder adoption due to its technical nature. Teachers may struggle to understand algorithmic processes or navigate complex interfaces, reducing confidence in AS. Addressing these barriers through training and user-friendly design can support smoother adoption. Compatibility refers to how well AS aligns with teachers' educational goals, values, and practices. Teachers are more likely to adopt AS if it supports curriculum standards and instructional strategies, such as promoting scientific reasoning or providing personalized feedback. Understanding these perceptions helps determine whether AS is seen as enhancing or disrupting existing instructional practices. Trialability allows teachers to experiment with AS in low-stakes settings, helping them build confidence and familiarity. Controlled trials and iterative learning allow teachers to explore the features of AS and refine their practices, reducing anxiety. Providing such opportunities can support smoother integration and inform strategies for effective AS adoption. Observability refers to the ability of teachers to see concrete, positive results from AS, such as improved assessment accuracy and better instructional decisions. Visible benefits increase teachers' confidence in AS and influence their willingness to adopt it. Applying Rogers' theory helps identify key factors, such as training and trial environments, to support the effective integration of AI-enhanced tools in science education.

Integrating Digital Assessment Literacy with Teachers' Perceptions of AS

Digital assessment literacy is a multifaceted framework that encompasses the skills, knowledge, and competencies necessary for educators to effectively design, administer, and interpret assessments in digital learning environments(Eyal, 2012). It goes beyond basic digital literacy and requires a deep understanding of various digital tools, assessment methods, and data analysis to create reliable and valid online assessments. This literacy also includes competencies related to the ethical considerations, security protocols, and accessibility standards that ensure fair and equitable digital assessment practices.

Educators' perceptions of Automated Scoring Systems(ASS) are closely tied to their digital assessment literacy. An advanced level of digital assessment literacy equips educators with the ability to critically evaluate the features and limitations of automated scoring systems. Teachers with high digital assessment literacy recognize the benefits of ASS in efficiently handling large amounts of data and providing immediate feedback, but they also understand the challenges, such as the limitations of algorithmic scoring in assessing complex cognitive and affective student responses. In this way, digital assessment literacy enables educators to weigh the reliability and validity of ASS against pedagogical goals and use the technology judiciously in their instructional practices.

Conversely, educators with limited digital assessment literacy may lack the framework to fully understand or trust automated scoring systems and may view them with skepticism or apprehension. For example, educators unfamiliar with digital data analysis may question the accuracy of ASS or be concerned about its fairness and transparency. Therefore, fostering digital assessment literacy is critical to shaping informed perceptions of ASS and empowering educators to effectively integrate such technologies while ensuring alignment with pedagogical values and instructional goals. Through professional development focused on digital assessment literacy, educators can become more adept at navigating the complexities of automated assessment, ultimately supporting the effective integration of AI-based assessment technologies in educational contexts.

Integrating AI Literacy with Teachers' Perceptions of AS

Integrating AI literacy with teachers' perceptions of automated scoring (AS) can foster a balanced and informed use of AI-driven assessment tools in the classroom. Teachers with a basic understanding of AI concepts, such as algorithms and data processing, are more likely to view AS as a complementary tool that can provide objective insights while recognizing its limitations, such as its inability to capture nuanced classroom dynamics. With AI literacy, teachers can develop a critical trust in AS, appreciating its potential for efficiency and fairness while remaining mindful of the biases and data limitations inherent in machine learning models. This understanding enables teachers to effectively use AS data for formative assessment, instructional adjustments, and individualized feedback. Ultimately, AI-savvy teachers can be better prepared to thoughtfully integrate AS and use it as part of a dynamic, responsive approach to assessment that supports both student learning and their professional growth in a technology-enhanced educational environment.

Methods

Research Context and Participants

This study was conducted as part of larger study to facilitate scientific argumentation and develop an AS model for spoken scientific argumentation (Lee & Ryu, 2020, 2021). During three consecutive academic years 196 10th grade students (104 male, 92 female) and 103 11th grade students (76 male, 27 female) participated in a scientific argumentation design-experiment by dealing with 5 science-oriented, 5 socioscientific issues and 2 science experiment topics for each year. As machine learning techniques for automated scoring, we applied support vector machines, decision tree, random forest, and artificial neural network. In this study, the scoring model used scientific terms and discourse markers of scientific argumentation as classification criteria for scoring. Teachers were able to see both individual and group argumentation performance depending on the frequency of argumentative elements. Regarding the accuracy of scientific content knowledge, the AS model utilized a combination of relevant science terms and morphemes.

Of the teachers involved in the larger study, five volunteered to participate in focus group interviews, including three chemistry teachers, one biology teacher, and one general science teacher, each with more than five years of teaching experience. These teachers served as an expert panel in the development of the argumentation scoring criteria, comparing their own scoring to that of the AS system. Prior to the focus group interviews, they attended training sessions to gain basic knowledge of AS and machine learning concepts that would enhance their science teaching practices and support the development of AS. In the first session, the teachers were introduced to the basic machine learning concepts underlying AS. In the second session, they collaborated to build an expert system designed to improve the accuracy of AS, specifically in identifying refutations within scientific arguments. This expert system, a tool that replicates the decision-making processes of human experts, was built to capture teachers' nuanced understanding of argumentation. These teachers, selected based on their familiarity with the literature on criteria for evaluating scientific argumentation, contributed to the design of a step-by-step decision tree, incorporating specific discourse markers and scientific terms to refine the algorithm. Through this hands-on process, the teachers actively shaped a more accurate AS model and contributed their expertise directly to the development of the system.

After participating in the development of the AS model, they used the AS models in their science classrooms prior to the focus interviews. A sample of 2,605 utterances about molecular structure was used to revise the automated algorithm. The focus group teachers classified the utterances into five argumentative elements and performed text preprocessing, such as decision tree, on the classified utterances. After the participation, the teachers used the same automated scoring topic for teaching and feedback. The essay topic was that water and ammonia have similar molecular weights. However, the boiling point of water is 100℃, while that of ammonia is -33.4℃. Let us explain the reason in terms of molecular structure.

Data Collection and Analysis

Focus Group Interview

The primary data source was focus group interviews. A focus group interview is a qualitative data collection method in which a small, diverse group of people are interviewed in a group setting to discuss a specific topic. Guided by a moderator, the goal of a focus group interview is to elicit in-depth, interactive responses, as participants share ideas, experiences, and viewpoints(Cheng, 2007; Dilshad & Latif, 2013). This study chose the focus group interview method because it is particularly useful for exploring complex issues. The author played as a moderator and asked semi-structured questions. The focus group method was useful for investigating teachers' ideas and considerations for the use of AS in the classroom as this method facilitated active discussion among teachers. The semi-structured interview questions were developed based on Rogers' innovative diffusion theory model, digital assessment literacy, and AI literacy. The interview data were fully transcribed. A descriptive approach was used to analyze the interview data. Supplementally, we also referred to the field notes as well as classroom video data to understand classroom context and confirm AS data.

Data triangulation using a variety of sources enhanced the understanding of the research participants' responses. In addition, feedback from a science education expert and colleagues with extensive teaching experience was incorporated into the data analysis process to revise and refine the findings. In-depth validation techniques, such as member checking by research participants and peer review, were also used.

Focus Group Interview Analysis with Diffusion Theory

To analyze focus group interview scripts using diffusion theory, we focused on identifying teachers' perceptions of AS through key constructs: relative advantage, compatibility, complexity, trialability, and observability. Responses discussing the benefits, challenges, and ease of use of AS reflect relative advantage and complexity, while comments about how AS fits with existing assessment practices touch on compatibility. In addition, teachers' openness to experimenting with AS reflects trialability, and their reactions to visible results indicate observability. By categorizing responses according to the stages of adoption-knowledge, persuasion, decision, implementation, and confirmation-we can gauge where each teacher is in the adoption process and identify patterns such as higher acceptance, passive acceptance with caution, and doubters. Examining these elements provides insight into both barriers and facilitators to AS adoption, highlights factors that influence teachers' willingness to integrate AS into their practice and supports recommendations for effective implementation.

AS in the Science Classroom from Digital Assessment Literacy

To analyze teachers' responses after using AS in science classrooms, we used Digital Assessment Literacy (DAL) as an analytical framework. The analysis was organized around specific DAL competencies and used qualitative coding methods to effectively categorize and interpret the data. A coding scheme was created based on DAL constructs that included technical understanding, data interpretation, instructional integration, and ethical/critical reflection. This scheme included subcodes within each category, such as "understanding AS algorithms" under technical understanding, "identifying misconceptions" under data interpretation, and "ethical concerns about objectivity" under critical reflection. Next, using thematic coding, we applied these DAL-based codes to segments of the interview data where teachers discussed their experiences and views of AS in the classroom. Specifically, we categorized responses according to how teachers articulated their technical understanding, any perceived advantages or limitations in interpreting AS data, or their reflections on using AS findings in the classroom. Through segment-level coding, we extracted specific themes that highlight where teachers demonstrate strengths in DAL and where they encounter challenges.

Results

Teachers' Perceptions of AS

Teachers' perceptions of AS were categorized according to diffusion theory. *Relative advantage* refers to the preference for innovative technologies over existing ones. Teachers viewed AS as providing systematically documented information about students' reasoning. They noted that this wealth of information allowed them to track student progress throughout the course, rather than simply comparing starting and ending scores. Teachers also observed that fairness can be a unique benefit of automated grading. The following are excerpts from what teachers said:

"Chemistry 1T: AS allows for multiple assessments. Both parents and students want a credible assessment, and more resources with more frequent assessments can secure such credibility."

"General Science T:In our city, the performance test accounts for 40% of the total test score. Therefore, the school office encourages teachers to rate performance together and compare scores. However, it is extremely difficult to do it that way because the teachers' responsibility for their own scores is already a burden. Consequently, if AS is widely used, that would be useful."

"Biology T: The most important criterion is to secure objectivity. Computers always apply the same criteria, which improves fairness."

Teachers who recognized the relative advantages of AS expressed a higher expectation of the usefulness of AS and thus their acceptance to use AS because AS allow them to make multiple, credible assessments. Their acceptance, which is related to the recognition of advantages and usefulness, is similar to the previous result showing the relationship between perceived usefulness and teachers' acceptance of AI technologies in the classroom.

In contrast, a couple of chemistry teachers expressed concern that argumentation scores focused on quantitative features of structural elements (e.g., the number of pieces of evidence) and could be misused to measure the accuracy of scientific content. Another chemistry teacher expressed concern about the advantages of using AS, that unique and original ideas could be ignored.

"Chemistry 2T: Even though I could see the clear advantages of using AS in science teaching, I think that only the teachers could judge the quality correctly. For example, more number of evidence is not always better, but AS counts the number of evidence. Besides, I cannot trust that AS will evaluate the content accuracy well enough"

"Chemistry 3T: I agree with Chemistry 2T. One of the main reasons for writing a scientific argument is to have a chance to come up with greater, I mean unique and original ideas. I wonder how the AS could recognize his originality. After all, it is the teacher's job to evaluate the arguments. And the AS assessment could weaken the accountability of

teachers"

Teachers who were concerned about the potential risk of using AS mentioned that teachers should be responsible for scoring the assessment. This suggests that awareness of the potential risk may influence teachers' acceptance of the use of AS in the classroom.

Regarding "complexity," we explored the difficulties, confusion, and conflicts that teachers might face when dealing with AS in the classroom. Teachers were concerned that AS scores seemed to depend on the number of argument elements that appeared or the number of scientific terms, thus ignoring argument quality in terms of scientific literacy such as scientific novelty, critical thinking, and scientific knowledge. These concerns were related to the conflict over the discrepancy between AS scores and teacher-determined scores.

"Chemistry 3T: Scientific arguments should not be treated the same as general arguments in the literature. If you use a novel idea as an argument or evidence, you should get more credit, even if the quality of your writing isn't very high. Any type of information or query can be easily searched through the Internet. Therefore, the evaluation must address how one critically presents his genuine idea. I'm not quite sure if AS can assess such a point"

"Chemistry 2T:"It would not be possible to rate the accuracy of scientific knowledge based on the order of the keywords used. It is important to understand the content of the argument. The importance of understanding scientific knowledge through argumentation should not be underestimated"

A critical point is that these two chemistry teachers seemed to understand the scoring mechanism of AS. Chemistry 2T was aware of the mechanism that the order of keywords was used as a marker of the content accuracy of the AS algorithm, and Chemistry 3T expressed her concern that the AS algorithm in this study for scoring scientific arguments did not seem to be much different from the general essay scoring system. It should also be noted that these two teachers were familiar with the general benefits of using AS, such as saving time and money.

'Compatibility' was defined as the degree to which new technology, or a new product conforms to existing social norms, values, and cultures. Regarding compatibility, teachers pointed out that social norms for collaborative argumentation such as 'listen to each other' and 'giving everyone an equal opportunity to talk' should be considered in the AS score. Although teachers generally regarded AS as accurate and reliable, they noted that more points should be given to students who act as listeners, interrogators, and facilitators of group debates. As mentioned above, because AS scores depend on the number of argumentative elements present, dominant members are more likely to score higher than his/her peers.

"Biology T: The students' arguments have been completely transcribed. And machine learning algorithms evaluate the transcription. I'm not sure, but I'm a little more concerned now that the amount of conversation is the sole, main factor in determining the score. However, there might be a way to supplement the score by giving more credit to students who facilitated the conversation and helped resolve conflicts, who acted as negotiators and leaders. And this problem arose because this AS was trying to score individual students' participation in class discussion. It might not be a problem to grade a piece of written argumentation".

"General Science T: There are some students who facilitate an argumentation by interpreting/rephrasing other students' talk or by asking adequate questions. In doing so, these students make a scientific argumentation active, and thus contribute to make the quality of the argumentation better, which is like the role of a teacher. These students played a

teacher's role. I don't think that AS can consider this kind of contribution"

A teacher was concerned about teachers' authority in her science classroom. According to her, the reason that her students respect her authority is not because of her scientific knowledge superiority, instead, it mainly relies on her right to evaluate student performance.

"Chemistry 3T: Speaking of social norms and class culture, there is something that sincerely worries me. I believe that the results of AS are not necessarily accurate. So, I think there is a problem if students prioritize AS results over teacher grade results. In a high school, some degree of teaching authority comes from the right to grade students. It can be presumed that respect for teachers does not necessarily come from their knowledge, but because they have the authority to score and evaluate. Therefore, for the evaluation, I think it is necessary to confirm that the teacher's evaluation is more important than the AS result"

Teachers' feedback highlights concerns about automated scoring (AS) systems that may misalign with classroom social norms and undermine teacher authority. They note that AS may favor dominant speakers over students who enhance argumentation by listening, questioning, or facilitating - a misalignment with collaborative norms that value diverse roles. In addition, AS's focus on the quantity of participation may inadvertently encourage competitive behavior rather than inclusive, respectful discussion. Teachers also worry that reliance on AS could diminish their authority, as students may prioritize AS scores over teacher evaluations, which are integral to maintaining classroom structure and respect. To address these issues, one teacher suggested that AS systems should include indicators that value facilitative roles and maintain a model in which teacher input is essential, ensuring alignment with educational norms and preserving teachers' evaluative authority.

Regarding 'experience', the researcher asked teachers about specific characteristics of AS that they either preferred or disliked. Teachers recommended that AS can be used as a mapping tool to represent the overall quality of argumentation class and the participant structure in the classroom. What teachers hate about AS was that it forces students to adhere to idealistic argumentation structures rather than considering sound scientific knowledge or competencies.

"General Science T: We thought that automatic grading was only used to assess the performance quality of individual students. But I think it would be helpful to look at the class results as a whole. Although there are individual differences, I was able to see which element of the scientific argumentation was lacking when viewed as a whole, and in that regard, the connection of frequently used words was helpful"

"Biology T: I thought the students were actively participating in the scientific argument in their own way. However, I noticed that the percentage of non-Argumentation is very high. When you look around and observe, you can discover who is doing well and who is not. However, the AS results provide more detailed information, such as what percentage of claims there are and what percentage of evidence comes from group argumentation. This way, you can better understand which group is doing better or not."

"Chemistry 1T: Right. It would be nice to know the overall level. I knew the rate of justification or rebuttal was low, but the percentage of these elements was too low. If I could see the results every hour, I think I could have put more emphasis on justification or rebuttal."

Teachers' insights reflect a valid tension between the structural advantages of AS and its limitations in fostering authentic scientific discourse. When used as a mapping tool, AS provides teachers with valuable classroom-wide insights that are difficult to capture in real time, such as which argumentative elements are underrepresented. This aggregated data can be powerful in guiding instructional decisions and helping teachers promote balanced argumentation skills across groups.

However, teachers' concerns highlight a critical limitation. AS's tendency to favor rigid structures over content may inadvertently suppress meaningful contributions that don't conform to an "ideal" argumentative format. Authentic scientific inquiry often involves multiple approaches to argument and explanation, and students may demonstrate deep understanding in ways that deviate from prescribed structures. Thus, it seems essential that AS systems evolve to accommodate a broader range of scientific literacies. By allowing for teacher input and adapting to assess content and quality of knowledge alongside argumentative form, AS could better support learning environments that value both structure and genuine understanding. In this way, AS could complement rather than constrain pedagogical flexibility and student autonomy in scientific argumentation.

In the case of 'perceived trust', some teachers' willingness and intention to adopt AS in their science classrooms seemed to be positive despite personal concerns described above. They also pointed out that AS is an inevitable trend. At the same time, however, they pinpointed that AS cannot completely replace teacher evaluation of students. Teachers noted that AS can be used as supplemental and additional information.

"*General Science T: The need for classes using computer-based or meta-buses seems to have increased since the coronavirus. It is said that elementary science textbooks include the content of coding, but I think that coding and artificial intelligence will become more and more important in secondary schools, and the evaluation using artificial intelligence will also become more important.*"

"*Chemistry 1T: Nowadays, artificial intelligence is all the rage, and when a new technology comes in from school, everyone tells science teachers to use it first and provide them with training (laughs)*"

Teachers seemed to perceive that the use of AS and other artificial intelligence-enhanced learning environments will be broadly spread into their science classrooms. In addition, since science teachers are often taken as early adaptors for new instructional technologies in a school, they thought that they need to be ready for an AI learning environment including the AS.

Teachers' reflections on "perceived confidence" in AS reveal a cautiously optimistic outlook on its adoption, with an understanding that AS and AI-driven tools are likely to become integral to the future of science education. Despite personal concerns about the limitations of AS, most teachers in this group seemed to recognize its inevitability in educational trends, particularly in response to the increased emphasis on technology-enhanced learning following the COVID-19 pandemic. They saw AS as a valuable additional tool that could complement, rather than replace, traditional teacher assessment and provide additional insight into student performance.

Their balanced approach seems prudent, preserving the critical role of teacher evaluative judgment while integrating the efficiency and data-driven insights of AS. While AS can streamline certain aspects of assessment and help identify broad patterns, it cannot fully capture the nuances of student understanding or engagement that expert teachers observe directly. However, teachers' willingness to embrace AS as part of a broader AI learning environment demonstrates their adaptability and commitment to evolving pedagogical practices. The interview suggests that teachers who are generally receptive to AS seem to hold the view that AS remains a supportive rather than a dominant tool in assessment.

Teachers' Ideas of the Use of AS for Teaching and Assessment with Understanding of Algorithm

Teachers' ideas of using AS for their teaching and assessment after their using AS in science classes were examined based on the Digital assessment literacy framework(Eyal, 2012). As human experts, teachers developed a step-by-step decision tree that identified the most used morphemes and scientific terms from a set of 50 active scientific argumentation episodes. In these episodes, students discussed a chemistry topic: water and ammonia have similar molecular weights. However, the boiling point of water is 100℃, whereas ammonia is -33.4℃. Let us explain the reason in relation to the molecular structure.

Teachers consented that the experience of interpreting and reviewing AS results helped them understand the process of developing students' thinking, thereby linking students' performance with instruction. They agreed that knowing students' thoughts and their misconceptions helped them to improve their science instruction.

"General Science T: ⋯. While I thought saving time with AS and more evidence for grading is a good thing, I didn't think about how to use the AS results for instruction because I felt that grading itself wouldn't say much about student performance. However, the AS results showed what caused the students to have misconceptions or confusion, which allowed me to understand the students' thinking through⋯"

"Biology T: ⋯What I realized from this result was that I have relied heavily on the group's final performance as a group performance, rather than their discussion process itself. Reviewing the final presentation tended to reflect the ideas and work of an academically advanced student rather than all group members. Although I walked around the class during the small group discussions, I couldn't hear each student's talk and therefore had difficulty picking up their thoughts. The AS results allow me to track a group's discussion, which allowed me to adjust my instruction accordingly"

It seemed that teachers did not think of AS results as an opportunity to obtain information on students' performance at first. Instead, they only thought that the results of automatic scoring could be reflected in the performance assessment. However, the enhanced understanding of AS through participation in revising the expert algorithm seemed to help them realize they could obtain in-process argumentation. Teachers noted that ideas about how to proceed with science class can be generated from this information.

For another thing, teachers came to not only consider the use of AS for their instruction but also reflect their habit of evaluating the classroom and its relation to the instruction. While teachers knew that the quality of the final product from group argumentation mostly relied on the ideas and efforts of academically advanced students in the group, it was hard for teachers to know the argumentation performance of every single student and thus provide feedback. Consequently, In-process information of AS seemed to give teachers ideas about how to provide individual feedback, which could be reflected in science instruction.

Interestingly, one teacher said that she was able to understand better the value of scientific argumentation for the development of scientific knowledge in the scientists' community, a characteristic of the Nature of Science by examining group argumentation's development process using AS. According to her, she thought argumentation could help students make their thinking clearer and visible, but not necessarily relate to the development of scientific knowledge through argumentation with students' limited content knowledge. However, the results from the expert system enabled her to trace how students' competing ideas—often silly ideas became shaped much better by securing and arguing evidence together. At this point, she said:

"Chemistry 1T: I know this might sound unrealistic, but I would like my students to build an expert system like I did, or at least see how this tree grows. Their dumb ideas got a lot better because of checking their evidence, which may help them understand what can be counted as solid evidence after all."

A teacher mentioned that the expected time saving by using AS allowed him to consider the integration of argumentation or other various activities into his science class.

"Chemistry 1T: The time savings and instant results give me much more room to spend more time on chemistry experiments or other activities. Honestly speaking, sometimes the fact that the results of student activity must be recorded and evaluated prevented me from planning such activities and discussions. Well, it's also true that students can engage in discussion more actively when they participate in experiments or other specific activities, compared to the times when they were only asked to discuss for discussion"

What stands out in designing assessments using AS is that teachers seem to have voluntarily come up with the notion of 'dynamic assessment' and customized assessment. The notion of dynamic assessment highlights the iterative process between assessment and instruction (Poehner & Lantolf, 2005). The result of the assessment is instantly reflected in the content of the instruction, and the changed instruction asks for the change in the assessment followed. A teacher mentioned, although she was aware of the reasoning process before she participated in composing an expert system, the understanding of scoring by knowing the algorithm contributes to the growing credibility of AS for its accuracy. Teachers also pointed out that decreasing the burden of scoring helped them utilize spoken argumentation as a performance assessment.

"General Science T: because basically, you can see what students talk all the time and also the system automatically scores it, teachers may assess students' performance more frequently, which enables teachers to know better the students and thus the classes"

Beyond the capability of assessing performance and in-process by using AS, a teacher pointed out the possibility of tracing growth using AS.

"Biology T: I always have an idea of assessing an individual student's growth, not just her achievement. To assess students' growth, a teacher must assess many aspects of students' performance and need to assess several times to see the growth."

"Chemistry 2T: What is important, though, AS does auto score and might be able to provide some kinds of instant feedback. Yet, what is important, here, it is still up to teachers, responsible for making decisions about what needs to be assessed, and how. Participation in the construction of an expert system helped me understand how AS score, and thus enabled me to think more specifically about what can be assessed best"

Teachers also saw the possibility of using AS as an effective way of conducting a peer assessment as well as a self-assessment.

"Chemistry 1T: Peer assessment has been stressed recently, but often students think that it is just an assessment for an assessment, which means that there is no such useful information from it. Because students are very much concerned about their reputation over other peers, they do not point out any specific points about learning. As teachers also know the concern, most peer assessments ask about peer affection or emotion, collaboration aspects of which results are almost always very positive. The AS may help students provide more productive feedback"

"Biology T: There is no right answer key. So, students never determine the quality of their scientific argument by themselves. AS results might help students assess themselves, if possible, regularly. In this way, students can be more motivated by obtaining independent control over learning argumentation"

With the growing importance of in-process assessment, teachers mentioned that they used peer assessment and self-assessment more than before. However, the limited resources for such assessments and less credibility seemed to prevent teachers from employing such assessment methods effectively in the classroom. Thus, the instant results and frequent assessing nature of the AS seem to be able to support a peer- or self-assessment more efficiently.

Teachers' reflections illustrate an evolving view of AS not just as a grading tool, but as an integral component for improving classroom practice and supporting dynamic assessment. Dynamic assessment is a formative approach that focuses on learning potential through interactive, adaptive feedback rather than on final scores. Unlike static assessments, it provides continuous guidance, allowing teachers to adjust instruction in real time based on student responses. This approach personalizes feedback to support individual learning needs, fostering deeper understanding and critical thinking. In science education, AS systems with dynamic assessment can provide immediate feedback during tasks such as argumentation, helping students correct misconceptions and allowing teachers to adjust instruction on the fly, directly integrating assessment with learning. Initially, teachers viewed AS primarily as a means to streamline grading, but their involvement in creating a decision tree to identify key elements of an argument allowed them to see AS as a tool for tracking students' thought processes and identifying misconceptions. This deeper understanding enabled teachers to use AS insights to tailor science instruction based on individual and group argumentation trends, thus helping students clarify and refine their thinking.

Teachers' reflections illustrate an evolving view of AS not just as a grading tool, but as an integral component for improving instructional practice and supporting dynamic assessment. Dynamic assessment is a formative approach that focuses on the learning potential of students.

Interestingly, the immediate feedback and comprehensive assessment capabilities of AS suggest opportunities for growth tracking, peer assessment, and self-assessment. Teachers see AS as a way to support peer feedback that goes beyond social dynamics and encourages students to objectively critique the quality of argumentation. Similarly, AS could facilitate self-assessment and motivate students by allowing them to monitor their progress independently. Finally, teachers' reflections indicate a shift toward seeing AS as a partner in teaching rather than a mere assessment tool, one that supports individualized learning and has an iterative, responsive quality.

Discussion

The integration of AS into educational settings represents a significant technological advance, but its practical application highlights nuanced challenges and opportunities for improvement (Ferrara & Qunbar, 2022; Lee et al., 2019; Lee & Ryu, 2020; Wilson et al., 2021). While AS has strengths such as efficiency, objectivity, and scalability, its limitations in assessing qualitative aspects-creativity, critical thinking, and the nuanced accuracy of scientific knowledge-must not be overlooked (Shin et al., 2021; Wilson et al., 2021). Incorporating teacher input to refine algorithms and adopting multidimensional assessment approaches can bridge the gap between quantitative measures and the deeper cognitive skills required for science learning.

Another critical consideration is the balance of authority between AS and teachers. Students may begin to prioritize AS results, potentially undermining the teacher's evaluative role and authority in the classroom. This dynamic underscores the need for a hybrid assessment model in which AS serves as a supplemental tool while teacher judgment remains central. By reinforcing the teacher's role in final assessments, education systems can maintain respect for pedagogical expertise while reaping the benefits of AS (Eyal, 2012; Ng et al., 2021).

In addition, fostering community awareness and collaboration-a key goal for 21st-century learning-requires a rethinking of how AS evaluates contributions in group settings (Molenaar, 2022). Current AS models risk favoring dominant voices and neglecting the importance of facilitative and inclusive roles. By developing algorithms that recognize diverse collaborative contributions, such as listening and conflict resolution, AS can better align with educational goals that emphasize inclusivity and teamwork.

In summary, the effective integration of AS into the classroom requires a dual focus: improving the technological framework to address pedagogical concerns and maintaining the human-centered elements of instruction. With careful calibration, AS has the potential not only to improve assessment practices, but also to foster richer, more equitable learning environments (Ferrara & Qunbar, 2022; Kaldaras & Haudek, 2022; Lee & Ryu, 2020; Shin et al., 2021).

Conclusion

This study examined teachers' perceptions of automated scoring (AS) in science classrooms, using diffusion theory and digital assessment literacy as an analytical framework. Teachers acknowledged the strengths of AS in terms of timeliness and accuracy and appreciated its potential to increase fairness and objectivity in assessment. However, teachers, even those with higher levels of acceptance of AS, generally viewed AS as a supplemental rather than a primary assessment tool, using its insights as an additional reference alongside traditional assessment methods. Their concerns reflect their intention to maintain the classroom culture of teacher authority and respect for students who contribute to the classroom. Compatibility issues arose, particularly regarding the limitations of AS in interpreting social cues, which teachers considered essential for comprehensive assessment.

Understanding machine learning concepts proved beneficial to teachers, enhancing their understanding of how AS works and sparking ideas for integrating it into teaching science. Teachers recognized AS's potential for diagnostic, dynamic assessment, with results informing individualized instruction. In particular, some teachers engaged with the AS algorithm and appreciated how reviewing and revising AS results could foster collaborative idea generation.

The findings of this study underscore the importance of considering teacher expertise and professional judgment when integrating AI tools such as AS. By recognizing AS as a complement to traditional assessments, teachers can leverage both the precision of automated analysis and the nuanced understanding of social and contextual dynamics that only human educators can provide. Moreover, teachers' engagement with the machine learning concepts underlying AS signals a promising shift toward greater AI literacy in education, which could empower teachers to shape AI tools to meet pedagogical needs. This collaborative approach, incorporating both technological insights and pedagogical expertise, has the potential to advance AI-enhanced learning environments that are not only efficient, but also aligned with the core values of holistic education.

An important contribution of this study is its exploration of teachers' firsthand experiences with AI-based assessment systems, which can inform the design of AI-enhanced learning environments. Although the small sample size limits the generalizability of the findings, the results suggest that AI used in conjunction with teacher evaluation and reflection could have a more balanced and effective impact on the classroom. Future studies with larger sample sizes are needed to build on these findings.

Acknowledgement

This work was supported by the Ministry of Education of the Republic of Korea and the National Research Foundation of Korea (NRF-2019S1A5A8036701)

Conflicts of Interest

The author(s) declared no conflicts of interest.

References

Berland, L. K., & McNeill, K. L. (2010). A learning progression for scientific argumentation: Understanding student work and designing supportive instructional contexts. Science Education, 94, 765–793. https://doi.org/10.1002/sce.20402

Cheng, K. W. (2007). A study on applying focus group interview on education. Reading Improvement, 44, 194–199.

Cope, C., & Ward, P. (2002). Integrating learning technology into classrooms: The importance of teachers' perceptions. Journal of Educational Technology & Society, 5, 67–74.

da Rosa, J. A. (2024). Twenty-Years of Anti-Climate Change and Anti-Evolution Education Legislation in the United States. Science Education, 1-23. https://doi.org/https://doi.org/10.1002/sce.21907

Dilshad, R. M., & Latif, M. I. (2013). Focus group interview as a tool for qualitative research: An analysis. Pakistan Journal of Social Sciences, 33, 191–198.

Eyal, L. (2012). Digital assessment literacy - The core role of the teacher in a digital environment. Journal of Educational Technology & Society, 15, 37–49.

Ferrara, S., & Qunbar, S. (2022). Validity Arguments for AI‐Based Automated Scores: Essay Scoring as an Illustration. Journal of Educational Measurement. +Volume and page range

Haudek, K. C., & Zhai, X. (2023). Examining the Effect of Assessment Construct Characteristics on Machine Learning Scoring of Scientific Argumentation. International Journal of Artificial Intelligence. +Volume and page range

Kaldaras, L., & Haudek, K. C. (2022). Validation of automated scoring for learning progression-aligned next generation science standards performance assessments. Frontiers in Education. +Volume and page range

Lee, H. S., Pallant, A., Pryputniewicz, S., Lord, T., Mulholland, M., & Liu, O. L. (2019). Automated text scoring and real-time adjustable feedback: Supporting revision of scientific arguments involving uncertainty. Science Education, 103, 590–622. https://doi.org/10.1002/sce.21504

Lee, M., & Ryu, S. (2020). Automated scoring of scientific argumentation using expert morpheme classification approaches. Journal of The Korean Association For Science Education, 40, 321–336.

Lee, M., & Ryu, S. (2021). Automated Scoring of Argumentation Levels and Analysis of Argumentation Patterns Using Machine Learning. Journal of The Korean Association For Science Education, 41(3), 203-220.

Lin, H. (2022). Influences of artificial intelligence in education on teaching effectiveness: the mediating effect of teachers' perceptions of educational technology. International Journal of Emerging Technologies in Learning (Online), 17, 144–156. https://doi.org/https://doi.org/10.3991/ijet.v17i24.36037

McNeill, K. L., & Krajcik, J. (2012). Supporting Grade 5-8 Students in Constructing Explanations in Science: The Claim, Evidence, and Reasoning Framework for Talk and Writing. Pearson Allyn & Bacon.

Molenaar, I. (2022). The concept of hybrid human-AI regulation: Exemplifying how to support young learners' self-regulated learning. Computers and Education: Artificial Intelligence, 3, 100070. https://doi.org/https://doi.org/10.1016/j.caeai.2022.100070

Nebel, S., Schneider, S., & Rey, G. D. (2016). Mining learning and crafting scientific experiments: a literature review on the use of minecraft in education and research. Journal of Educational Technology & Society, 19, 355–366. http://www.jstor.org/stable/jeductechsoci.19.2.355

Ng, D. T. K., Leung, J. K. L., Chu, S. K. W., & Qiao, M. S. (2021). Conceptualizing AI literacy: An exploratory review. Computers and Education: Artificial Intelligence, 2, 100041. https://doi.org/https://doi.org/10.1016/j.caeai.2021.100041

Osborne, J., Erduran, S., & Simon, S. (2004). Enhancing the quality of argumentation in school science. Journal of Research in Science Teaching, 41, 994–1020. https://doi.org/10.1002/tea.20035

Poehner, M. E., & Lantolf, J. P. (2005). Dynamic assessment in the language classroom. Language Teaching Research, 9, 233–265. https://doi.org/10.1191/1362168805lr166oa

Rogers, E. M. (2003). Diffusion of Innovations (5th ed). The Free Press.

Ryu, S., Han, Y., & Paik, S.-H. (2015). Understanding co-development of conceptual and epistemic understanding through modeling practices with mobile internet. Journal of Science Education and Technology, 24. https://doi.org/10.1007/s10956-014-9545-1

Ryu, S., & Sandoval, W. A. (2012). Improvements to elementary children's epistemic understanding from sustained argumentation. Science Education, 96. https://doi.org/10.1002/sce.21006

Ryu, S., & Sandoval, W. A. (2015). The influence of group dynamics on collaborative scientific argumentation. Eurasia Journal of Mathematics, Science and Technology Education, 11. +page range https://doi.org/10.12973/eurasia.2015.1338a

Shin, S., Lee, J., Ha, M., & Park, J. (2021). A study on elementary and secondary science teachers' acceptance intention of automated computer scoring system using artificial intelligence. Journal of Educational Technology Research, 37, 101–120.

Tan, S. C., Lee, A. V. Y., & Lee, M. (2022). A systematic review of artificial intelligence techniques for collaborative learning over the past two decades. Computers and Education: Artificial Intelligence, 3, 100097. https://doi.org/https://doi.org/10.1016/j.caeai.2022.100097

Wilson, J., Ahrendt, C., Fudge, E. A., Raiche, A., Beard, G., & MacArthur, C. (2021). Elementary teachers' perceptions of automated feedback and automated scoring: Transforming the teaching and learning of writing using automated writing evaluation. Computers & Education, 168, 104208. https://doi.org/https://doi.org/10.1016/j.compedu.2021.104208

Authors' Information

Ryu, Suna: Korea National University of Education, Professor, First Author