Received: 19 August 2021 | Revised: 4 March 2022 | Accepted: 19 March 2022

DOI: 10.1002/tea.21773

RESEARCH ARTICLE

JRST | WILEY

Applying machine learning to automatically assess scientific models

Xiaoming Zhai[1,2] | **Peng He**[3] | **Joseph Krajcik**[3]

[1]Department of Mathematics, Science and Social Studies Education, University of Georgia, Athens, Georgia, USA

[2]Institute for Artificial Intelligence, University of Georgia, Athens, Georgia, USA

[3]CREATE for STEM Institute, Michigan State University, East Lansing, Michigan, USA

Correspondence
Xiaoming Zhai, Department of Mathematics, Science, and Social Studies Education, & Institute for Artificial Intelligence, University of Georgia, 105J Aderhold Hall, 110 Carlton St., Athens, GA 30602, USA.
Email: xiaoming.zhai@uga.edu

Funding information
National Science Foundation, Grant/Award Numbers: 2101104, 2100964; Lappan-Phillips Chair in the College of Natural Science at Michigan State University

Abstract

Involving students in scientific modeling practice is one of the most effective approaches to achieving the next generation science education learning goals. Given the complexity and multirepresentational features of scientific models, scoring student-developed models is time- and cost-intensive, remaining one of the most challenging assessment practices for science education. More importantly, teachers who rely on timely feedback to plan and adjust instruction are reluctant to use modeling tasks because they could not provide timely feedback to learners. This study utilized machine learning (ML), the most advanced artificial intelligence (AI), to develop an approach to automatically score student-drawn models and their written descriptions of those models. We developed six modeling assessment tasks for middle school students that integrate disciplinary core ideas and crosscutting concepts with the modeling practice. For each task, we asked students to draw a model and write a description of that model, which gave students with diverse backgrounds an opportunity to represent their understanding in multiple ways. We then collected student responses to the six tasks and had human experts score a subset of those responses. We used the human-scored student responses to develop ML algorithmic models (AMs) and to train the

computer. Validation using new data suggests that the machine-assigned scores achieved robust agreements with human consent scores. Qualitative analysis of student-drawn models further revealed five characteristics that might impact machine scoring accuracy: Alternative expression, confusing label, inconsistent size, inconsistent position, and redundant information. We argue that these five characteristics should be considered when developing machine-scorable modeling tasks.

KEYWORDS

artificial intelligence, artificial neural networks, deep learning, inclusive assessment, machine learning, natural language processing, scientific model

1 | INTRODUCTION

Society in the 21st century requires competent graduates with sufficient scientific knowledge to engage in public discussions on science-related issues (e.g., climate change), be critical users of scientific information related to their everyday lives, and continue to learn about science throughout their lives (National Research Council [NRC], 2012). In view of this need, the Next Generation Science Standards ([NGSS]; NGSS Lead States, 2013) set forth progressive and ambitious standards for students by integrating disciplinary core ideas (DCIs) and crosscutting concepts (CCCs) with scientific and engineering practices (SEPs). Such integrated three-dimensional (3D) learning requires students to construct models to explain phenomena and solve real-world problems, thus providing opportunities for students to fully appreciate the value of science and improve their scientific thinking (Krajcik & Merritt, 2012). Achieving these reform-oriented learning goals, however, requires the transformation of traditional assessment practices to performance-based assessments (Harris et al., 2019). Teachers who engage in such transformed assessment practices should be able to track students' progress and receive timely feedback to adjust instruction. By engaging in such practices, students will develop knowledge-in-use—the ability to apply scientific knowledge to solve problems and figure out solutions (Harris et al., 2019; National Academies of Sciences & Medicine, 2019).

However, achieving such promising assessment goals is not without challenge. Though assessment practices allow students to use ideas and provide opportunities for feedback to promote knowledge-in-use learning, they rarely take place in science classrooms (National Academies of Sciences, Engineering, and Medicine, 2019). This is particularly true for assessments that involve complex performance and representations, such as those targeting scientific modeling (Namdar & Shen, 2015; Schwarz et al., 2009). Scientific modeling assessments usually require students to visualize and represent their mental schema using drawings and/or writing (Schwarz et al., 2009). Drawings and writing can both provide explanations, but they may require different cognitive skills (Gilbert & Treagust, 2009). This is especially the case with emergent English learners (EELs) and students at the lower grades who might experience

more challenges when writing explanations as compared with drawing models. However, we have limited knowledge regarding the consistency of student-drawn models and their written descriptions of the models.

Moreover, teachers who rely on feedback to plan and adjust their teaching may not be willing to use the performance-based assessments if timely scoring is unavailable. Therefore, there is a need to explore scoring approaches to automatically assessing students' modeling performance (Furtak, 2017; Zhai, Yin, Pellegrino, et al., 2020). In prior studies (Gerard et al., 2019; Haudek et al., 2012; Lee et al., 2019; Liu et al., 2016), efforts have been made to apply machine learning (ML)—such as natural language processing (NLP)—as avenues to automatically score written constructed responses of explanation or argumentation; however, scientific models still remain one of the most challenging assessment tasks to automatically score (Zhai, Yin, Pellegrino, et al., 2020). The challenge is partly because students are required to use concrete representations to express their thinking of how and why phenomena occur when they construct models. The outcomes are usually represented in multiple ways, such as in free drawings (Gilbert & Treagust, 2009). Such representations reflect complex thinking and are challenging to differentiate via common computational technologies. The field does not yet know how to automatically score multirepresentations with high accuracy. Moreover, the nature of scientific modeling denotes the complexity of relationships among components in the systems under study. The greater the complexity of constructs assessed, the more challenging it is to use computational technologies to accurately differentiate students' modeling performances (Haudek & Zhai, 2021; Zhai, Yin, Pellegrino, et al., 2020).

Given the existing gaps, this research applies ML technologies to automatically score student-developed models. We used assessment tasks designed to align with the NGSS performance expectations (Harris et al., 2019; Zhai, Krajcik, & Pellegrino, 2021). The assessment tasks require middle school students to draw models using computer tools with text-based descriptions of the models. To facilitate classroom implementation of these tasks, we developed ML algorithmic models (AMs) to (a) automatically score student-drawn models and corresponding written descriptions and (b) validate ML scores with human experts' scores. Another unique contribution of this research is that we investigated how the drawn-model characteristics may impact the machine scoring accuracy, which is understudied (Zhai, Yin, Pellegrino, et al., 2020). Model characteristics can be used to inform assessment task design, which a recent meta-analysis of machine scoring deems critical to machine performance (Zhai, Shi, & Nehm, 2021). Thus, findings from this study will support future development of NGSS-aligned assessments that may be feasibly scored by ML algorithms. This study answers three research questions:

a. To what degree is student performance on drawn models consistent with their performance on written descriptions of the models?
b. How accurate are the scores assigned by machine algorithms to student-drawn and written descriptions of the models?
c. What characteristics of the drawn models may account for machine performance?

2 | AN INCLUSIVE PERSPECTIVE OF MODELING USING MULTIREPRESENTATIONS

The NRC's *Framework for K-12 Science Education* (NRC, 2012, pp. 56–57) reports that "scientists use models ... to represent their current understanding of a system under study, to

aid in the development of questions and explanations, and to communicate ideas to others." Hestenes (1992, p. 732) argues that "the great game of science is modeling the real world." Engaging students in developing and using models is considered a critical practice to improve students' scientific competence (Ke & Schwarz, 2021; Zhai, 2022; zu Belzen et al., 2019).

2.1 | Scientific models

Models provide a powerful tool with which to make sense of the world. Scientists use a variety of representations—including models—to explain or predict phenomena. A scientific model includes both abstraction and representation of the critical features and mechanisms of phenomena (Zhai, 2022). It represents a system that explains or predict phenomena (Shemwell & Capps, 2019), and can take a variety of forms; these can be categorized based on features, such as having a representational approach (e.g., drawings, graphs, diagrams), an epistemic purpose (e.g., explanatory or predictive), or a computational approach (e.g., system models or agent-based models; Harrison & Treagust, 2000).

Another function of models is communication—that is, models are a means to communicate one's understanding of phenomena. Given that human thoughts are invisible, one's understanding of phenomena must be expressed. In this process, one has to select the "modeling language," a form of representation that is understandable in the community. Such language can take the form of drawings (e.g., Tytler et al., 2020; Wilkerson-Jerde et al., 2015), graphs (e.g., Matuk et al., 2019), writing (e.g., Jong et al., 2015), simulations (e.g., Heijnes et al., 2018), mathematical formulas (e.g., Marshall & Carrejo, 2008), and so on. Such diverse multi-representations increase one's opportunity to develop and improve explanations because of the enriched approaches to communication. Although models may be represented differently, they share commonalities, such as generativity. Scientific models should be generative, as the model constructed to explain one phenomenon must be able to explain other related phenomena or predict future phenomena (Schwarz et al., 2009). Scientific models support theory generation, as they help scientists conceptualize problems and mechanisms, and figure out solutions.

Given the potential of modeling to improve science learning, it is essential to involve students in developing models. By developing models, students have opportunities to analyze, reason, synthesize evidence, and use scientific knowledge to explain and predict phenomena (Lehrer & Schauble, 2006b; Stratford et al., 1998). As such, models serve as representations of students' understanding. Through scientific modeling, students experience model construction, evaluation, testing, and use, mirroring the work that scientists do in their everyday practices (Lehrer & Schauble, 2012; Schwarz et al., 2017). Improving students' scientific modeling competence is therefore included in the *Framework for K-12 Science Education* (NRC, 2012).

2.2 | Multirepresentation of phenomena: An inclusive means for assessment

Multirepresentations in assessment practices could elicit students' knowledge-in-use and provide enriched, inclusive means for students to present their understanding. Scientists frequently select multirepresentations as a communication "language" or a combination of "languages" to investigate phenomena (Lehrer & Schauble, 2015; Zhai, 2022). Such activities have inspired educators to involve students in constructing multirepresentations, to aid in assessing students

with different learning capabilities. Multirepresentations usually contain information in terms of the mechanisms of the phenomena in different forms, which is processed using different sensory channels (e.g., auditory or visual). Such advantages could provide students with weaknesses in one sensory channel with opportunities to learn and represent their ideas in other sensory channels, creating inclusive learning and assessment possibilities.

Assessments that require the use of multirepresentations have the potential to better examine students' thinking. More importantly, multirepresentations are essential for students to demonstrate problem-solving abilities. Spiro (1988) argues that single representations might miss important facets of complex concepts, so that students may fail in applying their knowledge to problem solving. Therefore, science education should provide students with the opportunity to express their conceptual understanding, and engage them in multirepresentation activities to enrich their knowledge-in-use (Heijnes et al., 2018; Matuk et al., 2019; Singha & Loheide II, 2011; Tytler, 2021). In this study, we focus on student-drawn models and their text-based descriptions of those models, since these types of representations are frequently used in K-12 science classrooms to account for the mechanics of phenomena.

Student-drawn models are a predominant type of mechanistic model to assess students' knowledge-in-use. A mechanistic model comprises typical components and their relationships to show why phenomena occur. Compared with other models, drawn models are easily adopted in assessment practices to demonstrate learning. To figure out phenomena, students use their existing knowledge to construct an initial model showing a mechanism of the phenomenon. They then evaluate and revise the model based on evidence they collect through investigations. Drawn models allow students to illustrate their knowledge, promote epistemic agency, and show creative thinking with diverse expressions (Stroupe, 2014). In contrast to written responses, drawn models require lower English proficiency and are less differentiated by students' cultural backgrounds. Thus, drawn models are more inclusive, equitable measures of students' science competence as compared with other representations.

Students' written descriptions of models are also used to assess their conceptual understanding. Similar to drawn models, written descriptions are explanations of phenomena. However, they can clarify aspects of the drawn model. Written responses are also cognitive tools for students to make sense of phenomena (Tversky, 2001). Jong et al. (2015) examined students' use of the modeling-based text regarding the ideal gas law and found that students improved their performance to explain gas-related phenomena. By engaging in writing refutational texts, students improved their ability to explain phenomena (Tippett, 2010). However, written descriptions of models are generally formed in a sequence, reflecting how people think logically; it can therefore be challenging to express spatial information (Gobert, 2005). Akaygun and Jones (2014) compared students' and instructors' drawn versus written modeling performances and found significantly different patterns. Written descriptions included more procedural information, such as the dynamic nature of equilibrium, while drawn models expressed more information on structural aspects. Stenning and Oberlander (1995) found that text permits more ambiguity than visualizations in representations of ideas. For example, "the particle is subjected to a force" does not entail the magnitude nor the direction of the force, while a model using an arrow could show the magnitude and direction of the force. Given the differences between drawn models and written descriptions, it is thus essential to assess students' modeling competence using both representation forms.

Due to individual difference in learning preferences and students' familiarity with the representational "language," multirepresentations allow them to select the best channel to express their knowledge. Schneider et al. (2022) employed a randomized controlled trial design to examine the efficacy of high school chemistry and physics project-based learning interventions.

They found that students' proficiency improved significantly when experiencing multi-representational modeling activities in project-based learning classrooms, as compared with students in traditional science classrooms. Ainsworth (2006) identified three functions of multirepresentations in expressing thinking: to complement, constrain, and construct. If multi-representations include different information about a system due to functional constraints, one can complement the other representations to engage students in constructing deeper understandings.

To meet inclusive learning goals, science teachers need to use performance-based assessment tasks that involve students in using knowledge to model the mechanism of phenomena and figure out solutions to problems (National Academies of Sciences, Engineering, and Medicine, 2019). National investments toward this effort have generated high-quality assessment tasks, such as the Stanford NGSS Assessment Project (Wertheim et al., 2016) and the Next Generation Science Assessment project (Harris et al., 2019). These projects share commonalities: assessments require students to apply scientific knowledge (i.e., DCIs and CCCs) and use scientific practices to make sense of phenomena or solve problems. Eliciting students' knowledge-in-use requires them to apply texts and other forms of representations to express their understanding. For example, the NGSA project—upon which the present study builds—developed over 100 performance-based assessment tasks, many of which require middle school students to draw representations and provide descriptions of their models.

Despite the great potential in engaging students in multirepresentational modeling, assessing and evaluating student models is challenging due to the complexity and diversity of the constructed models. Scoring student models is time-consuming and increases teachers' workload. Given the challenges, teachers may not be willing to engage students in modeling practices. To solve this problem, this study applied automatic scoring techniques to score middle school student-drawn and written descriptions of models.

3 | APPLYING MACHINE LEARNING TO AUTOMATICALLY ASSESS STUDENT-DEVELOPED MODELS

Though modeling competence is highly associated with students' ability to solve real-world problems and fully appreciate the value of science, it is challenging to assess students' modeling competence (as noted above). A review study (Namdar & Shen, 2015) that critically examined assessments of modeling identified only 30 empirical studies, and most of those assessments target students' conceptual understanding and affective aspects. Few studies directly focus on students' modeling performance. Existing research on assessing students' modeling competence relies primarily on human coding of the models, which is time- and cost-intensive. Employing ML to automatically score models expressed in drawn and written formats can reduce these constraints.

ML is a critical component of artificial intelligence, the aim of which is to enable technologies to cognitively work like human beings. However, computerized technologies were primarily featured as executing commands that humans pre-set in the system (Zhai, Yin, Pellegrino, et al., 2020). For example, in the traditional automatic scoring of multiple-choice items, computers could feasibly score student responses once the keys were set. However, if student responses were not pre-set in the system, such as with constructed responses or drawings, computer analysis could not evaluate the responses. This is because students' constructed responses or drawings tend to be so diverse that it was impossible to pre-set the keys in the scoring system. Without pre-set keys, computers could not assign scores to the responses. Knowing this

limitation, engineers began to explore the development of computer technologies that could solve complex problems or score new cases. After decades of effort, Mitchell (1997) proposed the contemporary concept of ML, which utilizes computers' ability to learn from "experience" and apply that learning to solve new problems. This is similar to how humans develop skills to solve new problems (Zhai, Yin, Pellegrino, et al., 2020).

Among multiple ML technologies, supervised ML is most appropriate for automatic scoring, given its accuracy. Supervised ML typically constitutes two phases: *learning/training* and *predicting/testing* (Nehm et al., 2012; Zhai, Shi, & Nehm, 2021). In the learning phase, computers are fed with existing data that have been labeled by humans to develop AMs. AMs denote the specific relationships between the data (e.g., student responses) and the labels (e.g., scores) so that computers can use AMs to assign labels to new data. Once AMs are constructed, a validation procedure may be employed to confirm their prediction accuracy. In automatic scoring, we usually validate AMs by comparing computer scores with human consent scores and calculating the agreement (e.g., Cohen's kappa). Once AMs are validated, they can be applied to automatically predict new data or score new responses.

ML has the potential to revolutionize science assessments by significantly improving the functionality and automaticity of scoring open-ended and drawn responses, as they are able to target complex constructs that traditional assessments cannot evaluate (Zhai, 2021). Prior studies have shown the substantial potential of ML to make evidentiary inference based on large-scale and complex data (e.g., Bertolini et al., 2021; Rosenberg & Krist, 2021), which are difficult to analyze using traditional statistical methods. These studies suggest that ML has the potential to improve the functionality of assessments to make accurate decisions based on evidentiary data and rigorous inference. Due to the improvement of the assessment functionality, ML could potentially assess complex constructs with developmental features. For example, in their study, Wilson et al. (under review) employed ML to examine students' learning progression regarding argumentation performance—traditionally difficult to assess. Using the AMs developed, they predicted students' levels of progression immediately after students submitted their responses, which is both time- and labor-saving. In this study, we also target a complex construct: scientific modeling. Using drawn models and written texts, we collected rich data to infer students' modeling competence. We employed two ML approaches to automating the procedure: convolutional neural network (CNN) and NLP.

3.1 | Convolutional neural network for drawn models

CNN, a breakthrough in image classification, has been broadly applied in many fields. It is a subcategory of artificial neural networks (ANNs) that uses artificial neurals to represent learning. ANNs mimic the information processing and communication functions in biological systems. An ANN consists of a large number of connected artificial neurons (called nodes), modeling the function of neurons in a biological brain. Neurons possess computational functions to process the input (i.e., learning) information and then transmit that information to adjunct neurons through neuron connections, which are called *edges*. As information is processed, each edge is assigned a weight which will increase or decrease the strength of the transmitted information. Typically, each neuron has a threshold for the strength of the information. When the strength of the information is above the threshold, the information will transmit to the next neuron. In a typical ANN structure, neurons are aggregated in layers, in which

information traverses among layers. If the ANNs include multiple layers in a neural network (i.e., input, hidden, and output layers), the method is called *deep learning*.

Though various CNNs have been created, their fundamental infrastructure has two basic layers: a *convolutional layer* and a *pooling layer*. The *convolutional layer* is the core building block of CNNs, used for feature extraction (e.g., edges and color). This layer utilizes a filter (also called a "kernel") to recode the information (i.e., assigning weights) from the input through mathematical operation (i.e., convolution) and then projects the information to a receptive field (see Figure 1, left). The filter is usually smaller than the input features and sweeps by stride to cover all incoming information. The size and stride of the filter are the primary factors determining the recoded information (e.g., edges, color, gradient orientation). With multiple convolutional layers, later layers can see aggregated information from the prior receptive field to formulate an understanding of the input image. Similar to the convolutional layer, the *pooling layer* is also a filter used to reduce the number of parameters. Instead of assigning weights to the input information, the pooling layer applies an aggregation function to the input features to populate the outputs. This activation will extract the dominant features while reducing the computational power required to process the data.

This study employed the ResNet-50 V2 CNN, developed by He et al. (2016). Compared with prior CNNs, which relied primarily on increasing layers, ResNet-50 V2 uses the function of skip connections (i.e., residuals) to avoid prediction accuracy becoming saturated as layers increase. To achieve this goal, He et al. (2016) introduced a residual block to operate the skip connection. Instead of learning from the outcome, the identity block learns from the residual (hence "ResNet"). The residual informs whether the information will skip certain layers, so that the added layers will continue decreasing errors. The ResNet approach may improve accuracy for CNNs with 1000+ layers (for those interested in the technical aspects, see He et al., 2016).

As a growing technology, CNN has achieved remarkable success in machine vision, the technology used to automatically inspect and analyze image data (Allen-Zhu & Li, 2019; Deng, 2012). Prior studies show that the rate of error can be lowered to 0.27% (Ciresan et al., 2011). Recent research compared the most popular residual ResNet performance on the CIFAR-10 data set with other machine deep learning approaches (Arora et al., 2019; Recht et al., 2018). Findings suggest that ResNet (96% test accuracy) significantly outperformed other algorithms, such as NTKs (77% test accuracy) and random feature kernels (85% test accuracy). Despite this advantage, ResNet has seldom been used to evaluate drawn models (Pei et al., 2019). In this study, we employed ResNet 50 V2 to automatically score student-drawn models.

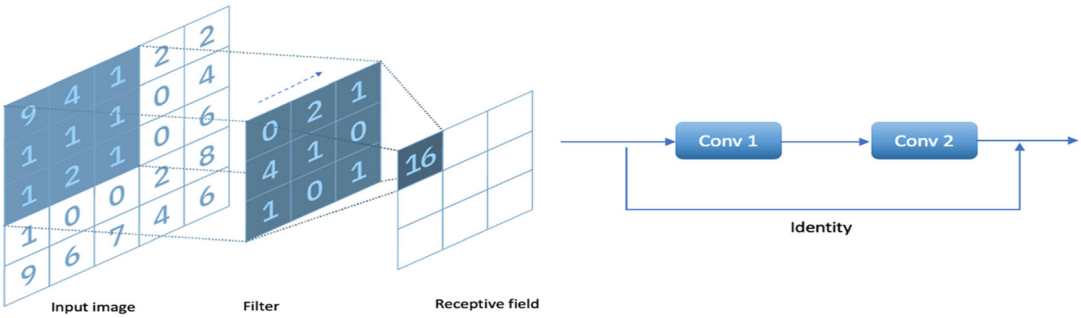


FIGURE 1 Convolutional neural network building blocks (left) and residual learning (right)

3.2 | Natural language processing for text analysis

NLP is a technique for characterizing and transforming texts by syntactic and/or semantic rules through statistical-based processing algorithms. Using NLP, computers can pause, segment, extract, and analyze text data. Unlike text mining, which uses words as units of analysis, NLP utilizes the underlying metadata, such as content or phrase patterns as units of analysis. Therefore, NLP is appropriate for analyzing short constructed responses in science. In their study, Nehm et al. (2012) describe how they applied the NLP procedure to their assessments of students' explanation of evolution. The first step was feature extraction, during which available features (e.g., words and word stems) were identified and extracted. Step two was algorithm construction; using statistical analyses, computers compared and differentiated features of responses containing a concept (labeled by human experts) and those not containing the concept (also labeled by human experts). The last step was algorithm validation, during which computer predictions were compared with human labels. This description represents the general procedure for applying NLP and is aligned with the ML application procedures articulated above. A review (Zhai, Yin, Pellegrino, et al., 2020) shows that more than 10 algorithms have typically been used for the evaluation of written responses to science tasks. For example, Lee et al. (2021) employed c-rater-ML, which built on a support vector regression model to score students' written argumentations and achieved robust human–machine agreements. In addition, an ensemble algorithm can integrate multiple algorithms into one package. The multiple algorithms work simultaneously and can eventually be assigned weights based on their performance to generate an ensemble AM, which can potentially overperform any individual algorithm (Maestrales et al., 2021; Zhai, Haudek, Stuhlsatz, & Wilson, 2020). The present study employed the ensemble algorithm.

4 | METHODS

4.1 | Participants

All our assessment items were deployed via a web portal which was accessible to teachers and students for free (NGSA, 2021). Given the wide visibility and high accessibility, by the time this study collected data, more than 40,000 middle school students in 3400 classrooms were registered as users. We downloaded student responses and employed Excel to randomly select a sample of 1050 student responses. We assigned random numbers to each student, regardless of their school and other information, and then reordered students and selected the first 1050. To protect students' privacy, researchers were blocked from students' demographic information. Because the assessment items are aligned with the NGSS performance expectations, this limited the users to middle school students in NGSS states. Given the large sample size and the sampling approach, the samples selected were highly representative.

4.2 | Assessment development

Though a variety of approaches can be adopted for ML-based science assessment practices, common procedures appeared in practice and literature. In prior research, Harris et al. (2019) developed an evidence-centered approach specifically for developing NGSS-aligned 3D

assessments. To forward automatic scoring, we integrated this approach with ML and developed an ML-based NGSA framework (Zhai, Krajcik, & Pellegrino, 2021; see Figure 2). This framework follows principles specified by Mislevy and Haertel (2006) and can be applied to most ML-based science assessments, including constructed responses, essays, simulations, game-based assessments, and interdisciplinary assessments. This framework reflects the procedures that we used to develop items and AMs, which comprised seven main steps: identifying target performance expectation, domain analysis, domain modeling, task construction, computer algorithm development, performance classification, and instructional decision making (see Figure 2).

We developed six modeling items and the respective rubrics and machine scoring AMs. Each item has two questions: the first question asks students to draw a model to make sense of the phenomena using online tools, and the second question asks students to write a description of the model. The questions focus on developing an explanation for the same phenomenon using alternative representations. The six items target one NGSS performance expectation at the middle school level: *MS-PS1-4. Develop a model that predicts and describes changes in particle motion, temperature, and state of a pure substance when thermal energy is added or removed*. To better elicit students' performance to infer their proficiency, we first conducted a *domain analysis* by unpacking the performance expectation, specifying the DCIs, CCCs, and SEPs (for details see Harris et al., 2019). We then conducted *domain modeling* by developing several fine-grain performance-based learning goals that also include aspects of the DCI, CCCS, and SEPs that we call learning performances (LPs). Finally, we developed items to assess the LPs. Table S1 in the supplementary material contains the six items.

To better illustrate the item, we present the "red dye diffusion (item R1)" example and the corresponding response (see Figure 3). This item was designed to assess one LP*: Students develop a model that explains how particle motion changes when thermal energy is transferred to or from a substance without changing state*. The item includes a video of how red dye diffuses in

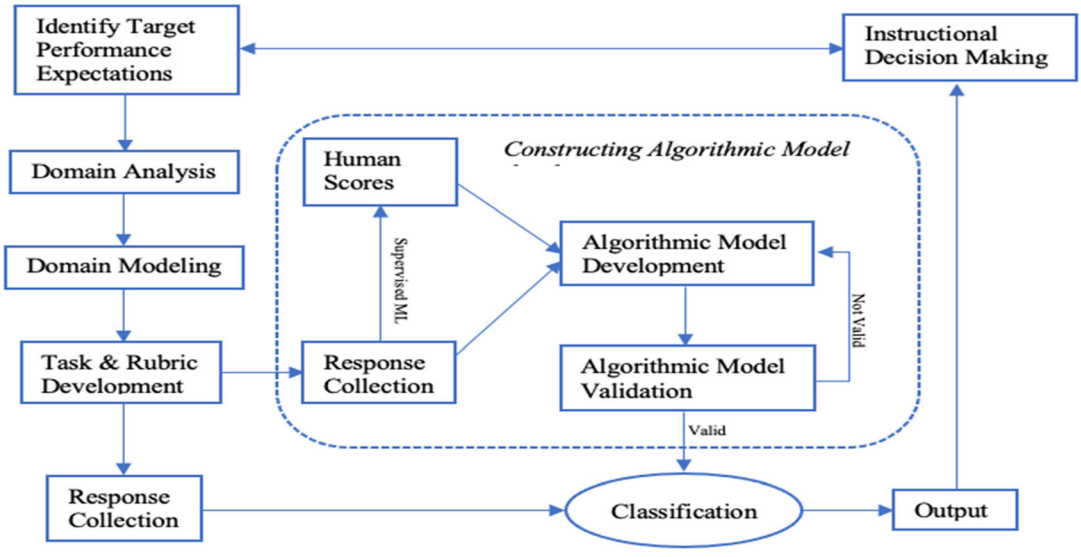


FIGURE 2 A framework for machine learning-based next generation science assessment (adopted from Zhai, Krajcik, & Pellegrino, 2021)

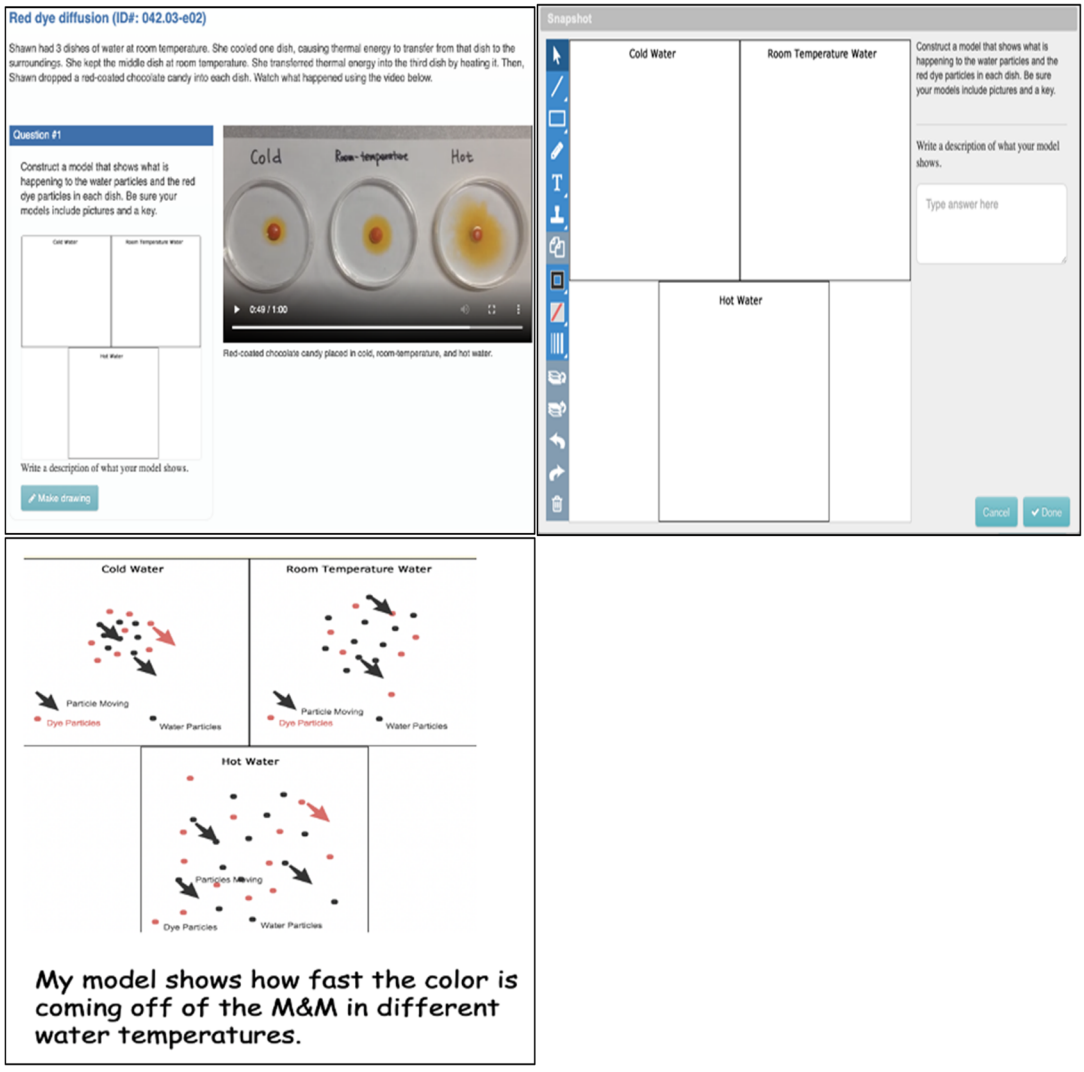


FIGURE 3 "Red dye diffusion" item screenshot (left), response interface (right), and a student response (bottom)

water at three different temperatures. Students watched the video and, on the same screen, they were asked to construct models to explain what they observed from the video and then write a description of their model. Once students clicked "Make Drawing," they were directed to a drawing pad with drawing tools, three drawing boxes, and a text box for writing (Figure 3, right).

To score student responses, we created two- or three-level analytic rubrics specifically for the items based on our design principle to rate student performance at the *beginning*, *developing*, and *proficient* levels. We determined the number of rubric levels based on the complexity of the relative aspects evaluated. For three items, we developed analytic rubrics that included three aspects, while the three remaining items contained two aspects. Table 1 shows the rubric for the example item, "red dye diffusion" (shown in Figure 3). We created two separate analytic aspects for scoring students on drawn models (three levels) and describing models (two levels).

In Figure 3 (bottom), the student response presented water and dye particles and the spread of particle motions from cold to hot temperature. However, the model failed to show the random movement of particles that can be inferred from drawing, as all arrows were in the same direction. According to Aspect 1 in the rubric, this drawn model was scored as developing level (1 credit). However, the student's written response described the drawn model in the following way: "My model shows how fast the color is coming off of the M&M in different water temperatures." The text failed to describe how water temperature impacts the motions of water and dye particles. As such, based on Aspect 2 in the rubric, the written response was scored as beginning level (0).

4.3 | Human scoring

We recruited six content experts with expertise in K-12 science education, including two professors of science education and four graduate students who were involved in the project, to score the six modeling assessment tasks. The six experts formed three pairs, and each pair was assigned to assess three randomly selected assessment tasks. A whole-group training was organized to introduce the assessment tasks and explain the scoring procedure, which included three iterative phases: training, scoring, and confirming interrater reliability. In the training phase, the selected student responses were first partitioned into 10 portions (hereby called "portion(s)"), and raters scored one of the portions independently and made notes if questions arose. They then compared their scoring outcomes, discussed discrepancies and questions, and resolved issues. They also made minor revisions to the rubrics, if necessary. Raters independently scored the second portion of randomly selected responses. The interrater reliability was checked, and issues were resolved through discussion. The third and fourth portions were used to score and check interrater reliability until the agreement between two experts met a cutoff of Cohen's kappa $= 0.75$, which represents excellent reliability (Fleiss et al., 2013). If new scoring consistency rules were made between raters, they reviewed the scored responses in the prior portions to ensure consistency. The training process continued until the interrater reliability values met the cutoff.

Once the interrater reliability met the cutoff, two raters independently scored half of the remaining data, with one randomly selected portion of data scored by both raters to confirm interrater reliability. Raters did not know which portion was selected until they complete the scoring. The interrater reliability was calculated after scoring to ensure it met the Cohen's kappa > 0.75; otherwise, the raters had to check their scores and re-calculate the Cohen's kappa. The Cohen's kappa for both training and confirmation are presented in Table 2. We then calibrated the correlation coefficients of student scores between drawn and written descriptions of models to address Research Question 1.

4.4 | Computer algorithm development and validation

To answer Research Question 2, we developed AMs using AI-STEM, which employed CNN and NLP to score drawn models and written descriptions of the models, respectively. AI-STEM is a learning platform developed for automatic assessment practices, including (a) a web portal that connects to powerful computer cluster systems to develop AMs; and (b) mobile applications, AI-Scorer, which connects students, teachers, and the e-cloud service that houses AMs for

TABLE 1 The rubric for the "red dye diffusion" item

	Proficiency		
Aspects of proficiency	**Proficient (2)**	**Developing (1)**	**Beginning (0)**
Draw a model to explain how the transfer of thermal energy affects the change in particle motion and/or temperature.	Student develops a model that identifies both water and dye particles and their motion while describing that water molecules move faster at higher temperatures (and vice versa). Response includes ALL of the following: a. Water and dye molecules move slowly when the water is cold, faster when at room temperature, and fastest when the water is hot. b. The key identifies water and dye particles c. The key identifies particle's motion (faster/slower).	Student develops a model that partially identifies both water and dye particles and their motion while describing that water molecules move faster at higher temperatures (and vice versa). Response includes AT LEAST ONE BUT NOT ALL of the criteria listed in "Proficient."	Student do not develop a model that identifies both water and dye particles and their motion while describing that water molecules move faster at higher temperatures (and vice versa). Response includes NONE of the criteria listed in "Proficient."
Use the model to explain how the transfer of thermal energy affects the change in particle motion and/or temperature.	Student fully describes how the model explain when thermal energy is transferred to the water, water and dye particles move faster. Response EITHER includes: a. when thermal energy is transferred to the water (hotter), water and dye particles move faster. b. at the higher temperature, water and dye particles move faster	N/A N/A	Student does not describe how the model explain when thermal energy is transferred to the water, water and dye particles move faster. The response does NOT INCLUDE EITHER of the criteria listed in "Proficient."

TABLE 2 Human scoring interrater reliability measures (weighted Cohen's kappa)

	Weighted Cohen's kappa			Weighted Cohen's kappa	
Rubric category of task	**Training**	**Confirming**	**Task**	**Training**	**Confirming**
R1-1	0.73	0.85	H4-1	0.82	0.84
R1-2	0.93	0.77	H4-2	0.84	0.90
J2-1	0.70	0.84	H4-3	0.85	0.80
J2-2	0.94	0.94	H5-1	0.89	1.00
M3-1	0.91	0.88	H5-2	0.90	1.00
M3-2	0.95	0.95	J6-1	0.88	0.83
M3-3	1.00	0.94	J6-2	0.88	0.90
			J6-3	0.84	0.87

Note: For "R1-1," "R1" indicates the item ID; the "-1" indicates the rubric category.

scoring items and providing feedback to teachers and students. When this study was conducted, AI-STEM was under construction, but we were able to use the prototype of the platform to develop AMs. We applied CNN to develop a machine scoring AM for each of the drawn model questions. To protect student information and maintain consistency, the margin area of the drawn models was cropped. Student responses were randomly split into training and testing groups at a ratio of 4:1. We developed AMs using the training group and then applied the model to assign scores to the testing group data. We calculated the machine–human agreement. To further validate the AM, we applied it to score the testing group and calculated machine–human agreement. In the training and validation processes, machine–human scoring agreements were calculated and indicated by accuracy, 95% confidence interval, and Cohen's kappa. Given that the testing group of drawn models was new to the machine AMs, the accuracy calibrated in this process represents the scoring capacity when the algorithm was used to score new data. To identify the machine scoring patterns, we collected all the failed scoring cases of drawn models and analyzed the potential reason for the failure. We compared the failed scoring cases with the successful ones to identify patterns. We then compared and summarized the patterns identified across items.

For the written responses, we applied the ensemble approach to developing AMs. Similar to prior research (Jescovitch et al., 2021; Zhai, Haudek, Stuhlsatz, & Wilson, 2020), the data were randomly split into 10 equal groups. Nine of the 10 groups were used to train the computer and develop an AM, and the leftover group of responses was used to test the accuracy of the AM. Machine scores were compared with human scores to calculate accuracy, 95% confidence interval, and Cohen's kappa. The processes rotated 10 times so that each group had the chance to serve as training data and testing data. The average of the scoring accuracy represents the scoring capacity for new data. As described above, the AMs for drawn models and written descriptions were validated using different approaches, which is consistent with the convention in ML.

4.5 | Qualitative analysis of response characteristics

Although the classification process of machine scoring is usually not transparent, the classification is clearly based on all information provided. In this study, the input information came from

the drawn models. Theoretically, all visible differences between drawn models could be factors that account for machine scoring differences. Thus, to answer Research Question 3, we employed a qualitative *manual matching* approach to discern characteristics that might account for scoring differences. Specifically, we analyzed the machine-mislabeled drawn models by matching these models with identical, correctly labeled models (i.e., counter-labeled cases). This approach was comprised of four procedures (see Figure 4). The first of these was *categorization*, during which two researchers reviewed all mislabeled drawn models. They then reviewed the correctly labeled models and identified those that were identical to each of the mislabeled cases. They then categorize the mislabeled model and the identical models in groups. In the second procedure, *matching*, researchers reviewed each group and matched the most identical correctly labeled model to each mislabeled case. In the earliest stage, each mislabeled case was matched with multiple correctly labeled cases. Two researchers then identified the most similar pairs via discussion using a consensus procedure. The third procedure entailed *discerning characteristics*. By comparing correctly and incorrectly labeled models within pairs, researchers discerned the characteristics distinct between the paired models. In the fourth and final procedure—*commonization*—a cross-case comparison between pairs of models was conducted to identify the commonalities of the characteristics, which might account for the discrepancies labeled by computers. Researchers then labeled these characteristics according to their identified commonalities. This approach helped identify the characteristics of paired models. Given the diversity of students' drawn models and the large number of data pools, the matching was challenging and time-consuming. The findings are also deemed exploratory.

5 | FINDINGS

In this section, we first report the correlation between drawn models and written descriptions to justify the necessity of multirepresentations. We then report the machine scoring accuracy for both drawn models and written descriptions. Lastly, we report the qualitative analysis of the machine scores of drawn models.

5.1 | Associations of student performance on the drawn model and the written description

To examine the necessity of multirepresentations in eliciting students' conceptual understanding and explanation of phenomena, we calculated the associations of student performances on drawn and written descriptions in each modeling task. Table 3 presents the Pearson coefficients

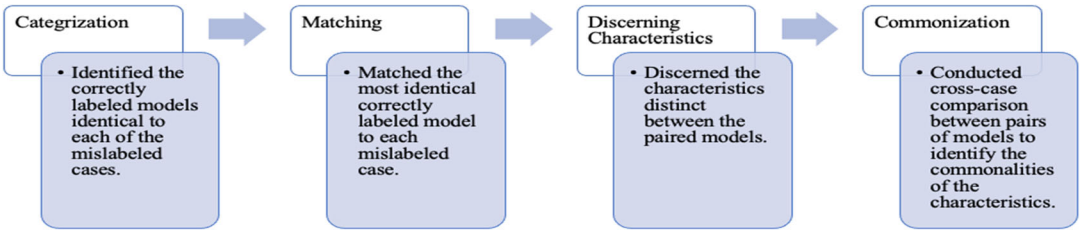


FIGURE 4 Procedures for the manual matching approach

of correlations between the drawn model and the written description in each task. The table shows significantly moderate to low correlations between the two types of representations, with coefficients ranging from 0.115 to 0.450. We found that the lowest correlation coefficient is the one between H4-1 (note that the H4 is the item ID and the -1 indicates the rubric category) and H4-3 ($r = 0.115$, $p < 0.05$), while that between H4-1 and H4-2 is moderate ($r = 0.442$, $p < 0.01$). A re-examination of the proficiency statements of the drawn model (H4-1) and the two written descriptions (H4-2, H4-3) revealed that the proficiency statements in the rubric between H4-1 ("develop a model to explain the change in the state of a substance resulting from the transfer of thermal energy") and H4-2 ("provide evidence that a model explains the particle motion changes because of the transfer of thermal energy") are tightly associated. Both questions require students to develop explanations accounting for the change in the state of the substance resulting from the transfer of thermal energy. In contrast, H4-3 rated students' ability to explain "a change in the state of a substance resulting from a change in particle motion." Though all three proficiency statements require students to develop an explanation, H4-3 ("change in particle motion") is a phenomenon that is more concrete and visible compared with H4-1 and H4-2 ("transfer of thermal energy"). To further examine the difference between students' performance on both representations, we transformed the raw scores into linear measures with equal intervals using Rasch measurement and found that the difficulty of the drawn model(s) was lower than that of the written descriptions for each item (Table 3; model fit refers to supplementary material Table S2).

5.2 | Accuracy of automatic scoring on student-drawn models

Table 4 shows the *accuracy* of the AMs for drawn models. The *accuracy* column indicates the percentage of human–machine agreements. For the training results, all six scoring models perform robustly, ranging from 0.95 to 0.98, with minor variation according to the 95% confidence interval. We also calculated Cohen's kappa, which indicates the human–machine agreement (excluding chance agreement) as ranging between 0.92 and 0.96, consistently indicating robust accuracy. The validation result column indicates the human–machine agreements from the testing samples, which were hidden from the computer when developing the AMs. The accuracy ranges from 0.82 to 0.89 and Cohen's kappa ranges from 0.64 to 0.82. The lower human–machine agreements compared with the training results indicate overfit of the models. According to the criteria for machine scoring (Nehm et al., 2012; Zhai, Shi, & Nehm, 2021), Cohen's kappa = 0.60–0.80 indicates *substantial*, and Cohen's kappa = 0.81–1.00 indicates *almost perfect*. Our results suggest that the machine scoring models are *substantial* with regard to scoring new data.

To provide insight into the findings, Table 4 also presents information on *sensitivity*, *precision*, *and prevalence*. *Sensitivity* indicates the percentage of machine-labeled cases among all the cases that should be labeled. High sensitivity indicates that, among the cases that should be labeled positive, the algorithm returns more positives than the cases that are not labeled positive. A score of 1 indicates that the cases labeled positive should be labeled positive, but says nothing about whether the labeled cases are those that should be labeled. In the training, results suggest that all the item categories achieved robust sensitivity (average = 0.96, SD = 0.02), while the validation results suggest a less robust sensitivity (average = 0.82, SD = 0.09).

Precision relates to the ratio of the correctly labeled cases against the total number of labeled cases. High precision indicates that, among the labeled cases (including cases that should be labeled and cases that should not be labeled), the algorithm returns substantially more cases

TABLE 3 The correlations between student performance on drawn models and on written responses

Task (N)	Drawn models (*d*)	Written description (*d*)	Pearson coefficient (*r*)	Task	Drawn models (*d*)	Written description (*d*)	Pearson coefficient
R1 (844)	R1-1 (0.18)	R1-2 (0.40)	0.306**	H4 (890)	H4-1 (−0.29)	H4-2 (0.28)	0.442**
J2 (883)	J2-1 (0.30)	J2-2 (0.46)	0.365**		H4-1 (−0.29)	H4-3 (−0.25)	0.115**
M3 (809)	M3-1 (−0.30)	M3-2 (1.13)	0.328**	H5 (834)	H5-1 (0.11)	H5-2 (1.59)	0.450**
	M3-1 (−0.30)	M3-3 (1.66)	0.339**	J6 (743)	J6-1 (−0.62)	J6-2 (−1.03)	0.438**
					J6-1 (−0.62)	J6-3 (1.51)	0.274**

Note: *d* = difficulty value calibrated from Rasch measurement. ** indicates statistically significant at level <0.01. Pearson coefficients were calculated based on raw scores.

TABLE 4 Machine–human scoring agreement for drawn responses to modeling assessment

	Training result						Validation result					
Task	**Accuracy**	**95% CI**	**Cohen's *k***	**S**	**Prec**	**Prev**	**Accuracy**	**95% CI**	**Cohen's *k***	**S**	**Prec**	**Prev**
R1-1	0.97	(0.96, 0.98)	0.95	0.97	0.99	0.52	0.86	(0.81, 0.91)	0.76	0.85	0.93	0.48
				0.98	0.95	0.39				0.90	0.78	0.43
				0.96	0.98	0.10				0.79	0.88	0.09
J2-1	0.95	(0.94, 0.96)	0.92	0.97	0.95	0.45	0.79	(0.73, 0.85)	0.64	0.85	0.80	0.47
				0.94	0.96	0.45				0.76	0.80	0.43
				0.93	0.94	0.11				0.64	0.68	0.10
M3-1	0.96	(0.95, 0.97)	0.95	0.98	0.94	0.37	0.83	(0.77, 0.89)	0.74	0.90	0.74	0.43
				0.96	0.98	0.39				0.82	0.89	0.37
				0.95	0.99	0.25				0.75	0.92	0.21
H4-1	0.97	(0.96, 0.98)	0.95	0.98	0.99	0.64	0.88	(0.83, 0.93)	0.76	0.92	0.94	0.64
				0.92	0.91	0.12				0.62	0.59	0.13
				0.98	0.97	0.23				0.89	0.86	0.24
H5-1	0.95	(0.94, 0.96)	0.94	0.95	0.83	0.14	0.82	(0.76, 0.88)	0.71	0.77	0.44	0.22
				0.97	0.98	0.52				0.84	0.92	0.48
				0.96	0.99	0.34				0.82	0.94	0.30
J6-1	0.97	(0.96, 0.98)	0.96	0.98	0.97	0.49	0.89	(0.84, 0.94)	0.82	0.95	0.88	0.52
				0.96	0.97	0.32				0.79	0.93	0.27
				0.97	0.97	0.20				0.91	0.85	0.21

Abbreviations: Prec, precision; Prev, prevalence; S, sensitivity.

that should be labeled than those that should not. A score of 1 indicates that all labeled cases are true but says nothing about whether all the truth cases are labeled. Our training algorithm returned an average precision = 0.96 (SD = 0.04), while the validation results suggest a less robust but sufficient average precision = 0.82 (SD = 0.13).

Prevalence indicates the proportion of cases among the rubric categories. Prevalence matters because the machine training effects are dependent on the sufficiency of training samples. Prevalence essentially provides insights beyond sensitivity and precision. For example, in some cases, we might have a low sensitivity for a specific rubric category while still achieving high accuracy. This might simply be due to the inspected rubric category having a low prevalence, and the sensitivity not significantly impacting the accuracy. For example, in our validation samples, the rubric category level 2 of item H4-1 had a lower sensitivity (0.62), which did not significantly weaken the accuracy (0.88) because the prevalence was only 0.13. Overall, though the sensitivity varied in the validation sample, the accuracy was relatively stable.

5.3 | Accuracy of automatic scoring on student-written responses

In Table 5, we present the *accuracy*, the 95% CI of the accuracy, the Cohen's kappa, *sensitivity*, *precision*, and *prevalence* from the cross-validation results for each written response question. The results indicate that the accuracy ranged from 0.86 to 0.94, and the 95% CI indicates that the accuracy for the items had minor deviations. Cohen's kappa for the nine items were all above 0.60, which is *substantial* according to Nehm et al.'s (2012) criteria.

All items with two rubric categories achieved robust sensitivity (above 0.95), while for items with three rubric categories there were individual categories with low sensitivity. A further

TABLE 5 Machine–human scoring agreement for written responses to modeling assessment

Task	Accuracy	95% CI	Cohen's k	S	Prec	Prev
R1-2	0.91	(0.89, 0.93)	0.74	0.97	0.91	0.75
J2-2	0.92	(0.90, 0.94)	0.81	0.96	0.93	0.70
M3-2	0.93	(0.92, 0.95)	0.81	0.97	0.94	0.77
M3-3	0.94	(0.92, 0.95)	0.76	0.98	0.94	0.82
H4-2	0.94	(0.92, 0.95)	0.81	0.98	0.95	0.78
				0.60	0.83	0.10
				0.88	0.92	0.12
H4-3	0.86	(0.83, 0.88)	0.78	0.94	0.89	0.39
				0.93	0.79	0.36
				0.39	0.93	0.25
H5-2	0.94	(0.92, 0.95)	0.87	0.96	0.93	0.56
J6-2	0.93	(0.91, 0.95)	0.85	0.95	0.93	0.58
J6-3	0.89	(0.86, 0.91)	0.62	0.98	0.89	0.78
				0.56	0.85	0.18
				0.00	NA	0.03

Abbreviations: Prec, precision; Prev, prevalence; S, sensitivity.

examination of prevalence indicates that these rubric categories correspond to a lower prevalence. Specifically, rubric levels 2 and 3 of H4-2 had a lower sensitivity of 0.60 and a slightly lower sensitivity of 0.88, respectively. However, due to the lower prevalence of 0.10 and 0.12, the overall accuracy was still robust (0.94). In general, the item rubric categories with lower sensitivity usually had lower prevalence. This may be the reason that our overall accuracy for all items was robust (except for item J6-3, which generated a lower Cohen's kappa = 0.62, which was still *substantial* according to Nehm et al.'s, 2012 criteria). Similar to sensitivity, our findings indicate that for all items with two rubric categories, the precision was above 0.91, which was robust. Item H4-3 with a three-level rubric had the lowest precision (0.79), which seemed to contribute to the lower accuracy (0.86), compared with the average accuracy = 0.92.

5.4 | Characteristics of drawn models

In this section, we present five characteristics of drawn models that may account for the machine mislabels. Although we identified more than five characteristics, we discuss the five that were most convincing. Our judgment regarding whether or not they were convincing was based on the comparison of the mislabeled and counter-labeled cases. A convincing pair of cases had to be identical except for one identified characteristic, which was the factor accounting for the computer mislabels. Table 6 summarizes the five characteristics—*alternative expression*, *confusing labels*, *inconsistent size*, *inconsistent position*, and *redundant information*—accompanied by examples.

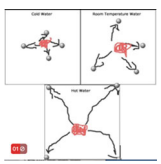
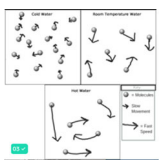
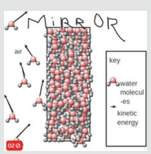
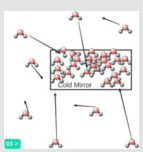
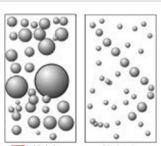
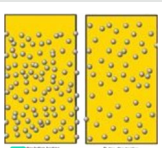
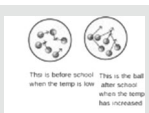
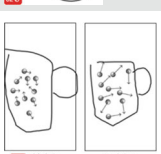
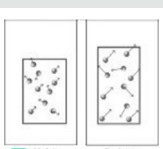
Alternative expression denotes students using different symbols or the use of the same symbols in different ways. Though students might have possessed a similar understanding, the diversity of symbolic language employed might create confusion for computers, yielding incorrect labels. In the example presented in Table 6, students used the length of arrows to represent the speed of dye particles and the distance between particles to represent temperature (i.e., kinetic energy). However, the mislabeled case inverted the directions of the arrows compared with the correctly labeled case, which led to incorrect labels.

Confusing label means that students made labels in the drawn models that are confusing to computers. In completing drawing tasks, students added labels to clearly indicate the components or relationships. If these labels were overly diverse, the computer tended to mislabel responses. For example, in the case presented in Table 6, the mislabeled case included handwritten words and some unsolicited labels. Although the ideas the student expressed in the mislabeled model were identical to the counter-labeled case, the drawn model was mislabeled.

Inconsistent size denotes that the sizes of the components in student-drawn models are inconsistent. In the drawing tool we provided, students could drag and draw components in any size they chose, which caused issues for the computer around identifying the critical information (such as the distance of the particles). In the example provided in Table 6, though both cases showed that butter particles are sparse in hotter water, indicating that the students had a similar level of understanding, the mislabeled case included particles with inconsistent sizes, which led to the mislabeling.

Inconsistent position is concerned with the position of the components drawn in the model. Table 6 presents examples in which both students possessed the same level of understanding. Both cases show that air particles inside the ball move slowly at a lower temperature and faster at a higher temperature. However, the mislabeled case positioned the two states vertically,

TABLE 6 Characteristics of drawn models, examples, and annotations

Category	Mislabeled case	Counter-labeled case	Annotation
Alternative expression			J1: Both cases show the dye particles are sparse and move faster (indicated by the arrows' length) when immersed at a higher temperature. Yet, the mislabeled case inverts the arrows.
Confusing label			H4: Both cases show that water particles are sparse and move faster (indicated by the arrows' length) before touching the mirror. Yet, the mislabeled case includes a confusing label box.
Inconsistent size			M3: Both cases show that butter particles are sparse in hotter water. Yet, the mislabeled case includes particles with inconsistent sizes.
Inconsistent position			R2: Both cases show that air particles inside the ball move slowly at a lower temperature and faster at a higher temperature. Yet, the mislabeled case positions the two states vertically, which is different from the correctly labeled case.
Redundant information			H5: Both cases show that water molecules move slowly (indicated using the arrow's length) before heating and faster after heating. Yet, the mislabeled case includes hand-drawn cups.

which was different from the correctly labeled case. As most students presented the two states horizontally, the computer was unable to correctly label the vertically presented case.

Due to the flexibility and the multifunctionality of the tools we provided, students sometimes input *redundant information* that confused the computer. In the example we provide in Table 6, students were expected to draw molecules and indicate both the speed and distances between molecules. The two cases presented show an identical understanding of the phenomena. However, the first case includes students' hand-drawn cups, which were confusing to computers.

6 | DISCUSSION

This study developed tasks to elicit students' modeling competence using drawn models and corresponding text-based descriptions. We found significantly moderate to low correlations between the two types of representations, indicating the necessity of multirepresentations for developing inclusive learning opportunities. To facilitate automatic scoring, we employed ML technologies to score student-drawn models and their written descriptions. Our research suggests that four of the six assessment items achieved excellent scoring accuracy through AMs, while the other two items achieved satisfactory accuracy. Our qualitative analyses identified five characteristics of drawn models that may significantly impact machine scoring accuracy. Findings in this study will contribute to the improvement and use of 3D science assessments—particularly drawn and written descriptions of models, as recommended by the *Framework for K-12 Science Education* (NRC, 2012) and the NGSS (NGSS Lead States, 2013)—and facilitate ML-based assessments in science education.

6.1 | Multirepresentations as a means of developing inclusive, equitable assessments

This study contributes to inclusive and equitable assessments by using visual and textual representations to assess students' modeling competence. Though modeling is critical to improve students' knowledge-in-use, assessing modeling competence in an inclusive and equitable manner is challenging. The findings contribute to the field by creating a multirepresentational form of assessment—asking students to draw and provide written descriptions of models. Our findings reveal significantly low to moderate correlations between drawn models and written responses. This finding suggests that although performance on drawn models is statistically associated with written descriptions, the performance consistency was low. This necessitates further exploration of multirepresentations in classroom assessments to promote inclusion and equity.

It is well acknowledged that the form of model (e.g., drawn or written) represents varying degrees of challenge for students. Our findings provide evidence that the overall mean scores on drawn models are higher than on written responses. The differences in scores can partially be explained by the additional skills needed to write explanations (e.g., students must write using technical words). For example, in the "red dye diffusion" item, students describe the particles and their motion, as well as how the motion is governed by the kinetic theory of gases. To correctly articulate the mechanism, students need to understand that a higher temperature results in a greater average kinetic energy of molecules that would result in speedier diffusion. This explanation demands a high level of writing proficiency and could potentially contaminate the inference of students' knowledge-in-use. Though students' descriptions were rated using

rubrics from varied perspectives that could thoroughly examine responses, the complex writing might make it challenging for some students to express their ideas. This is especially true with regard to invisible relations in a system that follow science principles (Ke et al., 2021; Lehrer & Schauble, 2006a; Namdar & Shen, 2015; Schwarz et al., 2017).

Consequently, written responses may decrease opportunities for students with reduced language proficiency to succeed in assessment practices. Even if an item requires little writing proficiency, it may remain challenging for emergent bilingual learners (EBLs) and students with writing disadvantages because it requires the use of technical words (Ryoo et al., 2018). Consequently, the written response itself might not sufficiently reflect the science proficiency of all students.

Drawn models appear more appropriate for EBLs and other students with lower levels of writing proficiency to communicate their understanding and resolve issues via text. Another accountability issue may be the sensory channels employed by students when expressing ideas. ChanLin (2001) found that novice students benefit from graphic information presented in problems compared with textual information. Individual differences in how learners use information sensory channels result from their background experiences and prior knowledge, and learning styles need to be considered in assessments. Using multirepresentations gives students equal opportunities to succeed in classroom assessment practices, thus contributing to inclusive, equitable science learning (González-Howard et al., 2017; Schwarz et al., 2017).

Our findings also reveal that students' performance differences between task forms might be associated with the specific performance expectations used in the rubric. Specifically, the evidence derived from Item H4 suggests that students' performance on drawn models (H4-1) was significantly associated with that on written descriptions in terms of rubric H4-2, yet not with H4-3. We suspect that the differences stem from the performance expectations. This finding suggests that the sophistication of the rubrics is equally important as the tasks developed for scoring accuracy and interpretations, if not more so.

Our findings suggest that multirepresentation is necessary for developing inclusive and equitable assessments. We embrace Lemke's (1990) idea that to express scientific ideas, text and speech need to work together with other representations.

6.2 | Insights into applying machine learning in responsive assessments

This study contributes to the field by reporting on the accuracy of ML to score students' multirepresentations of models. Though the science education community has adopted ML to automatically score open-ended assessments, a limited number of studies focus on the automatic scoring of models. According to a recent review (Zhai, Yin, Pellegrino, et al., 2020), most applications focus on students' written responses of scientific explanations or conceptual understanding (e.g., Ha & Nehm, 2016; Lee et al., 2021; Liu et al., 2016). As argued above, written responses are insufficient to elicit students' conceptual understanding and explanation of phenomena, while multirepresentations are essential (Gilbert & Treagust, 2009; LaDue et al., 2015). To our best knowledge, this is the first study employing ML to automatically score students' multirepresentations. The AI-STEM, which employs CNN and NLP, showed a powerful capability to score student-drawn models and written responses. The machine–human agreements reported in this study are as robust as human–human scoring, if not more so. The scoring accuracy for both drawn models and written descriptions provide robust evidence for the usability of the two ML approaches.

Though ML has been applied in automatic scoring, machine deep learning has rarely been applied in modeling assessments. In previous studies, machine deep learning was primarily applied to score written responses in science. For example, both Riordan et al. (2020) and Sung et al. (2021) applied machine deep learning—bidirectional encoder representations from transformers—to grade students' written responses, in order to examine their explanations and multimodal representational thinking. In both studies, the researchers found that machine deep learning outperformed traditional ML methods. Besides written responses, we found that limited studies had applied machine deep learning to score modeling, except one exploratory study (Zhai, Krajcik, & Pellegrino, 2021). The present study contributes to the field by demonstrates a comprehensive application of machine deep learning in scoring drawn models with high accuracy, showing the great potential of machine deep learning in complex classification (Allen-Zhu & Li, 2019).

This study also demonstrates the great potential of machine deep learning algorithms—specifically the ResNet-50 V2—in grading student-drawn models. In prior studies, automatic scoring technologies have been employed to assess student graphing competence. Vitale et al. (2015) developed technology to automatically score student-constructed position-time graphs to facilitate the use of assessments on graphing competence. Computer-executed commands encoded the spatial and numerical features of student graphs, such as the number or position of points and slope, to evaluate student graphs. However, students could not freely draw graphs because a limited number of graphing options were provided. While Vitale's technology is a constrained graphing tool, it represents an important step forward. In this study, we provided students with tools to draw models freely. We then employed ResNet-50 V2 to automatically score the responses. The success of this scoring can be attributed to the unique contribution of ResNet-50 V2: specifically, its use of skip connections to allow more hidden layers in the algorithm. Thus, ResNet-50 V2 could handle the diverse student-drawn models using its complex and powerful neural networks to process the training information. This approach may be used in other content areas, due to its flexibility.

This study also contributes to the field by exploring characteristics of drawn models that might impact machine scoring. Given that student models are developed and submitted in a digital setting, it is critical to ensure that assessment tools provide essential features to support students' model development and automatic scoring. Because few studies have explored automatic scoring of drawn models, we had limited knowledge regarding which characteristics might account for machine scoring and how to effectively design machine-scorable modeling tasks. This study employed a qualitative approach to identify characteristics that might account for machine scoring (in)accuracy. We found five major characteristics: (1) alternative expression, (2) confusing label, (3) inconsistent size, (4) inconsistent position, and (5) redundant information. Our findings indicate that unclear or redundant information might reduce the accuracy of automatic scoring of drawn models; these findings align with a prior meta-analysis of machine scoring accuracy. Zhai, Shi, and Nehm (2021) argue that "the assessment tasks and associated features are most accountable for machine–human agreement heterogeneity" (p. 12). This study contributes important knowledge to the field by specifying the features of student responses that inform task development so that student responses can be more feasibly scored using ML.

7 | IMPLICATIONS

Findings suggest that the development of machine-scorable NGSS assessments must consider not only what characteristics might account for machine scoring but also how to develop

assessments to utilize the many characteristics identified. In prior research (Zhai, Krajcik, & Pellegrino, 2021), we developed a validity inferential network for machine-scorable NGSAs by considering both the integrated nature of science learning and the characteristics of computer algorithms. The present study advanced this earlier study by providing empirical evidence for multirepresentational modeling assessments. According to the assessment characteristics we identified, one might consider adding more constraints to the assessment tasks to reduce the diversity of student responses or avoid redundant information. This proposal might work for ML scoring but could be contradictory to reform initiatives. Meeting the assessment goals of the NGSS requires that assessment tasks engage students in scientific modeling practices that provide sufficient flexibility to demonstrate knowledge-in-use (Pellegrino et al., 2014). This flexibility would significantly increase the diversity of student responses, and students might inadvertently use the flexibility to provide redundant and misleading information. Human experts could potentially differentiate this additional information from essential information, but it might confuse computer algorithms. Therefore, this issue cannot be resolved simply by decreasing or increasing constraints on assessment tasks. Researchers should consider the assessment goals and evidentiary inferences in assessment practices to determine the factors and how to manipulate these to improve ML of assessments.

This study also suggests that a research agenda to systematically examine the assessment task features for automatic scoring of representations is needed. In a prior study, Zhai, Shi, and Nehm (2021) employed a meta-analysis approach to examine machine scoring accuracy and found that six factors significantly moderated human–machine accuracy, including assessment external features (e.g., length, rubrics) and internal features (e.g., number of concepts, depth of knowledge). This study aligns with the earlier study and provides evidence to advise researchers to provide adequate scaffolds to support student-drawn models. Such scaffolds may include sufficient alternative labels, fixed drawing positions, and clear prompts, while considering the potential constraints of students freely drawing models. However, the extent to which these strategies could improve machine scoring remains unknown and needs further study.

This study also highlights the potential to include scientific modeling in classroom assessment practices to provide inclusive, equitable, and customized science learning. Baumfalk et al. (2019) found that feedback and scaffolds support students to develop, evaluate, and revise their models to be more sophisticated and coherent compared with those without scaffolds. Teachers can also use the feedback to flexibly and efficiently adjust their everyday instruction (Lee et al., 2021). Modeling is difficult to include in classroom assessment practices due to the challenge of providing timely feedback to students. The high accuracy of automatic scoring of students' multirepresentations would boost teachers' confidence in using modeling for classroom assessment practices, allowing teachers to use multirepresentation modeling assessments to facilitate their students' modeling proficiency and provide timely feedback.

8 | CONCLUSIONS AND LIMITATIONS

Since the release of the *Framework for K-12 Science Education* (NRC, 2012), the field has been driven by its reform-oriented goal to cultivate students' knowledge-in-use. While prior studies have documented efforts to develop assessments (Harris et al., 2019), this study focuses on using ML technologies to promote classroom assessment practices. We specifically looked at a complex practice—scientific modeling—and developed ML algorithms to examine students'

multirepresentations: drawn models and text descriptions of the models. The scientific modeling assessments we used comprised both visualized and textual representations that provided students with diverse opportunities to express their scientific understanding. Our findings suggest that, though the drawn models were consistent with written description of the models, the coefficients were generally low. The study provides evidence that ML can score student models and the text-based descriptions of the models with a high degree of accuracy. The study also contributes to assessment task development by identifying five characteristics that might be critical for computer accuracy, to help future research develop scored modeling tasks for ML.

This study provides evidence for the potential of ML to evaluate drawn models and written descriptions that reflect students' science knowledge-in-use in a timely manner, which fills an important gap in implementing ML in science education. Two special issues, "Science Teaching, Learning, and Assessment with 21st Century, Cutting-Edge Digital Ecologies" (Neumann & Waight, 2020) and "Applying Machine Learning in Science Assessments" (Zhai, 2021), conclude that ML is a rapidly growing area of research, with extensive potential. The potential of ML-based assessments—including improving the possibility to assess complex constructs such as modeling, as well as increasing assessment functionality and automaticity (Zhai, Haudek, Shi, et al., 2020)—will likely occur through the integration of assessment practices and technological innovations.

While this study makes contributions to scoring ML-based multirepresentations, we acknowledge its limitations. First, although we found a statistically significant but low correlation between drawn models and written descriptions, we were unable to uncover the reason. Future studies should therefore employ qualitative methods to further uncover the reasons for this low correlation. Second, given the nature of ML—a "black box" that is not transparent when assigning scores—it is challenging to uncover the computer prediction process. We sought to partially uncover the accountability of predictions using qualitative methods; however, the findings could be further verified using quantitative methods. Third, due to the constraints of privacy protection, we were cautious in collecting students' demographic information. Though we believe that the sample information and sampling approach was sufficient to warrant the findings and conclusions, we recommend that future studies further examine other characteristics and how they might be associated with machine scoring performance. Lastly, this study only focused on assessing students' representational products instead of their process. As scientific modeling practice involves creating, revising, testing, and deploying the representations (Baumfalk et al., 2019; Chen, 2021; Schwarz et al., 2017), future studies should explore other activities to better infer students' modeling competence.

ACKNOWLEDGMENT

The authors are gratitude with Jie Yang, Tingting Li, Sisi Han, and Lehong Shi, and other colleagues of NGSA project. This study was partially funded by National Science Foundation (NSF) (Award # 2101104 and 2100964). Any opinions, findings, conclusions, or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the NSF. The project was also partially supported by the Lappan-Phillips Chair in the College of Natural Science at Michigan State University.

ORCID

Xiaoming Zhai https://orcid.org/0000-0003-4519-1931
Peng He https://orcid.org/0000-0002-2877-0117

REFERENCES

Allen-Zhu, Z., & Li, Y. (2019). What can resnet learn efficiently, going beyond kernels?. *Advances in Neural Information Processing Systems*, 32, 1–36.

Ainsworth, S. (2006). DeFT: A conceptual framework for considering learning with multiple representations. *Learning and Instruction*, *16*(3), 183–198.

Akaygun, S., & Jones, L. L. (2014). Words or pictures: A comparison of written and pictorial explanations of physical and chemical equilibria. *International Journal of Science Education*, *36*(5), 783–807.

Arora, S., Du, S. S., Hu, W., Li, Z., Salakhutdinov, R., & Wang, R. (2019). On exact computation with an infinitely wide neural net. *Advances in Neural Information Processing Systems, 32*, 1–36.

Baumfalk, B., Bhattacharya, D., Vo, T., Forbes, C., Zangori, L., & Schwarz, C. (2019). Impact of model-based science curriculum and instruction on elementary students' explanations for the hydrosphere. *Journal of Research in Science Teaching*, *56*(5), 570–597.

Bertolini, R., Finch, S. J., & Nehm, R. H. (2021). Testing the impact of novel assessment sources and machine learning methods on predictive outcome modeling in undergraduate biology. *Journal of Science Education and Technology*, *30*(2), 193–209.

ChanLin, L. (2001). Formats and prior knowledge on learning in a computer-based lesson. *Journal of Computer Assisted Learning*, *17*(4), 409–419. https://doi.org/10.1046/j.0266-4909.2001.00197.x

Chen, Y.-C. (2021). Epistemic uncertainty and the support of productive struggle during scientific modeling for knowledge co-development. *Journal of Research in Science Teaching*, *59*, 383–422.

Ciresan, D. C., Meier, U., Masci, J., Gambardella, L. M., & Schmidhuber, J. (2011). Flexible, high performance convolutional neural networks for image classification. In Twenty-second international joint conference on artificial intelligence.

Deng, L. (2012). The MNIST database of handwritten digit images for machine learning research [best of the web]. *IEEE Signal Processing Magazine*, *29*(6), 141–142. https://doi.org/10.1109/MSP.2012.2211477

Fleiss, J. L., Levin, B., & Paik, M. C. (2013). *Statistical methods for rates and proportions*. John Wiley & Sons.

Furtak, E. M. (2017). Confronting dilemmas posed by three-dimensional classroom assessment: Introduction to a virtual issue of science education. *Science Education*, *101*(5), 854–867. https://doi.org/10.1002/sce.21283

Gerard, L., Kidron, A., & Linn, M. (2019). Guiding collaborative revision of science explanations. *International Journal of Computer-Supported Collaborative Learning*, *14*(3), 291–324.

Gilbert, J. K., & Treagust, D. F. (2009). Towards a coherent model for macro, submicro and symbolic representations in chemical education. In J. K. Gilbert & D. F. Treagust, (Eds.), *Multiple representations in chemical education* (pp. 333–350). Dordrecht: Springer.

Gobert, J. D. (2005). Leveraging technology and cognitive theory on visualization to promote students' science. In J. K. Gilbert (Ed.), *Visualization in science education* (pp. 73–90). Springer.

González-Howard, M., McNeill, K. L., Marco-Bujosa, L. M., & Proctor, C. P. (2017). 'Does it answer the question or is it French fries?': An exploration of language supports for scientific argumentation. *International Journal of Science Education*, *39*(5), 528–547.

Ha, M., & Nehm, R. (2016). The impact of misspelled words on automated computer scoring: a case study of scientific explanations. *Journal of Science Education and Technology*, *25*(3), 358–374. https://doi.org/10.1007/s10956-015-9598-9

Harris, C. J., Krajcik, J. S., Pellegrino, J. W., & DeBarger, A. H. (2019). Designing knowledge-in-use assessments to promote deeper learning. *Educational Measurement: Issues and Practice*, *38*(2), 53–67.

Harrison, A. G., & Treagust, D. F. (2000). A typology of school science models. *International Journal of Science Education*, *22*(9), 1011–1026.

Haudek, K. C., Prevost, L. B., Moscarella, R. A., Merrill, J., & Urban-Lurain, M. (2012). What are they thinking? Automated analysis of student writing about acid–base chemistry in introductory biology. *CBE—Life Sciences Education*, *11*(3), 283–293.

Haudek, K., & Zhai, X. (2021). Exploring the effect of construct complexity on machine learning assessments of argumentation. *Paper presented at the 2021 annual conference of the National Association of Research in Science Teaching*, Orlando, Florida, pp. 1–22.

He, K., Zhang, X., Ren, S., & Sun, J. (2016). Identity mappings in deep residual networks. In European conference on computer vision (pp. 630–645)

Heijnes, D., van Joolingen, W., & Leenaars, F. (2018). Stimulating scientific reasoning with drawing-based modeling. *Journal of Science Education and Technology*, *27*(1), 45–56. https://doi.org/10.1007/s10956-017-9707-z

Hestenes, D. (1992). Modeling games in the Newtonian world. *American Journal of Physics*, *60*(8), 732–748.

Jescovitch, L. N., Scott, E. E., Cerchiara, J. A., Merrill, J., Urban-Lurain, M., Doherty, J. H., & Haudek, K. C. (2021). Comparison of machine learning performance using analytic and holistic coding approaches across constructed response assessments aligned to a science learning progression. *Journal of Science Education and Technology*, *30*(2), 150–167.

Jong, J.-P., Chiu, M.-H., & Chung, S.-L. (2015). The use of modeling-based text to improve students' modeling competencies. *Science Education*, *99*(5), 986–1018. https://doi.org/10.1002/sce.21164

Ke, L., Sadler, T. D., Zangori, L., & Friedrichsen, P. J. (2021). Developing and using multiple models to promote scientific literacy in the context of socio-scientific issues. *Science & Education*, *30*(3), 589–607.

Ke, L., & Schwarz, C. V. (2021). Supporting students' meaningful engagement in scientific modeling through epistemological messages: A case study of contrasting teaching approaches. *Journal of Research in Science Teaching*, *58*(3), 335–365. https://doi.org/10.1002/tea.21662

Krajcik, J., & Merritt, J. (2012). Engaging students in scientific practices: What does constructing and revising models look like in the science classroom? *The Science Teacher*, *79*(3), 38.

LaDue, N. D., Libarkin, J. C., & Thomas, S. R. (2015). Visual representations on high school biology, chemistry, earth science, and physics assessments. *Journal of Science Education and Technology*, *24*(6), 818–834. https://doi.org/10.1007/s10956-015-9566-4

Lee, H.-S., Gweon, G.-H., Lord, T., Paessel, N., Pallant, A., & Pryputniewicz, S. (2021). Machine learning-enabled automated feedback: Supporting students' revision of scientific arguments based on data drawn from simulation. *Journal of Science Education and Technology*, *30*(2), 168–192.

Lee, H.-S., McNamara, D., Bracey, Z. B., Liu, O. L., Gerard, L., Sherin, B., Wilson, C., Pallant, A., Linn, M., & Haudek, K. C. (2019). *Computerized text analysis: Assessment and research potentials for promoting learning*. The International Society of the Learning Sciences.

Lehrer, R., & Schauble, L. (2006a). *Cultivating model-based reasoning in science education*. Cambridge University Press.

Lehrer, R., & Schauble, L. (2006b). Scientific thinking and science literacy. In W. Damon, R. Lerner, K. A. Renninger, & I. E. Sigel (Eds.), *Handbook of child psychology* (pp. 153–196). John Wiley & Sons.

Lehrer, R., & Schauble, L. (2012). Seeding evolutionary thinking by engaging children in modeling its foundations. *Science Education*, *96*(4), 701–724.

Lehrer, R., & Schauble, L. (2015). The developing scientific thinking. In L. S. Liben & U. Müller (Eds.), *Handbook of child psychology and developmental science* (pp. 671–714). Wiley.

Lemke, J. (1990). *Talking science: Language, learning, and values*. Ablex Publishing Corporation.

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.

Maestrales, S., Zhai, X., Touitou, I., Baker, Q., Krajcik, J., & Schneider, B. (2021). Using machine learning to score multi-dimensional assessments of chemistry and physics. *Journal of Science Education and Technology*, *30*(2), 239–254.

Marshall, J. A., & Carrejo, D. J. (2008). Students' mathematical modeling of motion. *Journal of Research in Science Teaching*, *45*(2), 153–173. https://doi.org/10.1002/tea.20210

Matuk, C., Zhang, J., Uk, I., & Linn, M. C. (2019). Qualitative graphing in an authentic inquiry context: How construction and critique help middle school students to reason about cancer. *Journal of Research in Science Teaching*, *56*(7), 905–936.

Mislevy, R., & Haertel, G. (2006). Implications of evidence-centered design for educational testing. *Educational Measurement: Issues and Practice*, *25*(4), 6–20.

Mitchell, T. M. (1997). Does machine learning really work?. *AI Magazine*, *18*(3), 11–20.

Namdar, B., & Shen, J. (2015). Modeling-oriented assessment in K-12 science education: A synthesis of research from 1980 to 2013 and new directions. *International Journal of Science Education*, *37*(7), 993–1023. https://doi.org/10.1080/09500693.2015.1012185

National Academies of Sciences, Engineering, and Medicine. (2019). *Science and engineering for grades 6–12: Investigation and design at the center*. National Academies Press.

National Research Council. (2012). *A framework for K-12 science education: Practices, crosscutting concepts, and core ideas*. National Academies Press.

Nehm, R. H., Ha, M., & Mayfield, E. (2012). Transforming biology assessment with machine learning: Automated scoring of written evolutionary explanations. *Journal of Science Education and Technology*, *21*(1), 183–196.

Neumann, K., & Waight, N. (2020). The digitalization of science education: Déjà vu all over again? *Journal of Research in Science Teaching*, *57*(9), 1519–1528. https://doi.org/10.1002/tea.21668

NGSA. (2021). Next generation science assessment. https://ngss-assessment.portal.concord.org

NGSS Lead States. (2013). *Next generation science standards: For states, by states*. National Academies Press.

Pei, B., Xing, W., & Lee, H. S. (2019). Using automatic image processing to analyze visual artifacts created by students in scientific argumentation. *British Journal of Educational Technology*, *50*(6), 3391–3404.

Pellegrino, J. W., Wilson, M. R., Koenig, J. A., & Beatty, A. S. (2014). *Developing assessments for the next generation science standards*. Washington, DC: The National Academies Press. https://doi.org/10.17226/18409.

Recht, B., Roelofs, R., Schmidt, L., & Shankar, V. (2018). Do cifar-10 classifiers generalize to cifar-10? *Preprint*. arXiv: 1806.00451, 1–25.

Riordan, B., Bichler, S., Bradford, A., Chen, J. K., Wiley, K., Gerard, L., & Linn, M. C. (2020). An empirical investigation of neural methods for content scoring of science explanations. In Proceedings of the fifteenth workshop on innovative use of NLP for building educational applications.

Rosenberg, J. M., & Krist, C. (2021). Combining machine learning and qualitative methods to elaborate students' ideas about the generality of their model-based explanations. *Journal of Science Education and Technology*, *30*(2), 255–267.

Ryoo, K., Bedell, K., & Swearingen, A. (2018). Promoting linguistically diverse students' short-term and long-term understanding of chemical phenomena using visualizations. *Journal of Science Education and Technology*, *27*(6), 508–522.

Schneider, B., Krajcik, J., Lavonen, J., Salmela-Aro, K., Klager, C., Bradford, L., Chen, I.-C., Baker, Q., Touitou, I., Peek-Brown, D., Dezendorf, R., Maestrales, S., & Bartz, K. (2022). Improving science achievement—Is it possible? Evaluating the efficacy of a high school chemistry and physics project-based learning intervention. *Educational Researcher*, *51*(2), 109–121.

Schwarz, C. V., Passmore, C., & Reiser, B. J. (2017). *Helping students make sense of the world using next generation science and engineering practices*. NSTA Press.

Schwarz, C. V., Reiser, B. J., Davis, E. A., Kenyon, L., Achér, A., Fortus, D., Shwartz, Y., Hug, B., & Krajcik, J. (2009). Developing a learning progression for scientific modeling: Making scientific modeling accessible and meaningful for learners. *Journal of Research in Science Teaching*, *46*(6), 632–654.

Shemwell, J. T., & Capps, D. K. (2019). Learning abstraction as a modeling competence. In A. P. zu Belzen, D. Krüger, & J. van Driel (Eds.), *Towards a competence-based view on models and modeling in science education* (pp. 291–307). Springer.

Singha, K., & Loheide, S. P., II. (2011). Linking physical and numerical modelling in hydrogeology using sand tank experiments and COMSOL multiphysics. *International Journal of Science Education*, *33*(4), 547–571.

Spiro, R. J. (1988). *Cognitive flexibility theory: Advanced knowledge acquisition in ill-structured domains*. Center for the Study of Reading.

Stenning, K., & Oberlander, J. (1995). A cognitive theory of graphical and linguistic reasoning: Logic and implementation. *Cognitive Science*, *19*(1), 97–140.

Stratford, S. J., Krajcik, J., & Soloway, E. (1998). Secondary students' dynamic modeling processes: Analyzing, reasoning about, synthesizing, and testing models of stream ecosystems. *Journal of Science Education and Technology*, *7*(3), 215–234.

Stroupe, D. (2014). Examining classroom science practice communities: How teachers and students negotiate epistemic agency and learn science-as-practice. *Science Education*, *98*(3), 487–516.

Sung, S. H., Li, C., Chen, G., Huang, X., Xie, C., Massicotte, J., & Shen, J. (2021). How does augmented observation facilitate multimodal representational thinking? Applying deep learning to decode complex student construct. *Journal of Science Education and Technology*, *30*(2), 210–226.

Tippett, C. D. (2010). Refutation text in science education: A review of two decades of research. *International Journal of Science and Mathematics Education*, *8*(6), 951–970.

Tversky, B. (2001). Spatial schemas in depictions. In M. Gattis (Ed.), *Spatial schemas and abstract thought* (Vol. 79, p. 111). MIT Press.

Tytler, R. (2021). The role of visualisation in science: a response to "Science teachers' use of visual representations". *Studies in Science Education*, *57*(1), 129–139. https://doi.org/10.1080/03057267.2020.1766826

Tytler, R., Prain, V., Aranda, G., Ferguson, J., & Gorur, R. (2020). Drawing to reason and learn in science. *Journal of Research in Science Teaching*, *57*(2), 209–231.

Vitale, J. M., Lai, K., & Linn, M. C. (2015). Taking advantage of automated assessment of student-constructed graphs in science. *Journal of Research in Science Teaching*, *52*(10), 1426–1450.

Wertheim, J., Osborne, J., Quinn, H., Pecheone, R., Schultz, S., Holthuis, N., & Martin, P. (2016). *An analysis of existing science assessments and the implications for developing assessment tasks for the NGSS*. Stanford NGSS Assessment Project Team (SNAP).

Wilkerson-Jerde, M. H., Gravel, B. E., & Macrander, C. A. (2015). Exploring shifts in middle school learners' modeling activity while generating drawings, animations, and computational simulations of molecular diffusion. *Journal of Science Education and Technology*, *24*(2–3), 396–415.

Wilson, C., Haudek, K., Jonathan, O., Stuhlsatz, M., Cheuk, T., Donovan, B., Bracey, Z., Mercado Santiago, M., & Zhai, X. (Under review). Using automated analysis to assess middle school students' competence with scientific argumentation.

Zhai, X. (2021). Practices and theories: How can machine learning assist in innovative assessment practices in science education. *Journal of Science Education and Technology*, *30*(2), 1–11.

Zhai, X. (2022). Assessing high-school students' modeling performance on Newtonian mechanics. *Journal of Research in Science Teaching*, 1–41. https://doi.org/10.1002/tea.21758

Zhai, X., Haudek, K. C., Shi, L., Nehm, R., & Urban-Lurain, M. (2020). From substitution to redefinition: A framework of machine learning-based science assessment. *Journal of Research in Science Teaching*, *57*(9), 1430–1459.

Zhai, X., Haudek, K. C., Stuhlsatz, M. A., & Wilson, C. (2020). Evaluation of construct-irrelevant variance yielded by machine and human scoring of a science teacher PCK constructed response assessment. *Studies in Educational Evaluation*, *67*, 100916.

Zhai, X., Krajcik, J., & Pellegrino, J. (2021). On the validity of machine learning-based next generation science assessments: A validity inferential network. *Journal of Science Education and Technology*, *30*(2), 298–312.

Zhai, X., Shi, L., & Nehm, R. (2021). A meta-analysis of machine learning-based science assessments: Factors impacting machine-human score agreements. *Journal of Science Education and Technology*, *30*(3), 361–379.

Zhai, X., Yin, Y., Pellegrino, J. W., Haudek, K. C., & Shi, L. (2020). Applying machine learning in science assessment: A systematic review. *Studies in Science Education*, *56*(1), 111–151.

zu Belzen, A. U., Krüger, D., & van Driel, J. (2019). *Towards a competence-based view on models and modeling in science education*. Springer.

SUPPORTING INFORMATION

Additional supporting information may be found in the online version of the article at the publisher's website.

How to cite this article: Zhai, X., He, P., & Krajcik, J. (2022). Applying machine learning to automatically assess scientific models. *Journal of Research in Science Teaching*, 1–30. https://doi.org/10.1002/tea.21773