International Journal of Artificial Intelligence in Education (2024) 34:1482–1509
https://doi.org/10.1007/s40593-023-00385-8

ARTICLE

Examining the Effect of Assessment Construct Characteristics on Machine Learning Scoring of Scientific Argumentation

Kevin C. Haudek[1] · **Xiaoming Zhai**[2,3]

Accepted: 27 November 2023 / Published online: 18 December 2023

Abstract

Argumentation, a key scientific practice presented in the *Framework for K-12 Science Education*, requires students to construct and critique arguments, but timely evaluation of arguments in large-scale classrooms is challenging. Recent work has shown the potential of automated scoring systems for open response assessments, leveraging machine learning (ML) and artificial intelligence (AI) to aid the scoring of written arguments in complex assessments. Moreover, research has amplified that the features (i.e., complexity, diversity, and structure) of assessment construct are critical to ML scoring accuracy, yet how the assessment construct may be associated with machine scoring accuracy remains unknown. This study investigated how the features associated with the assessment construct of a scientific argumentation assessment item affected machine scoring performance. Specifically, we conceptualized the construct in three dimensions: complexity, diversity, and structure. We employed human experts to code characteristics of the assessment tasks and score middle school student responses to 17 argumentation tasks aligned to three levels of a validated learning progression of scientific argumentation. We randomly selected 361 responses to use as training sets to build machine-learning scoring models for each item. The scoring models yielded a range of agreements with human consensus scores, measured by Cohen's kappa (mean = 0.60; range 0.38 − 0.89), indicating good to almost perfect performance. We found that higher levels of *Complexity* and *Diversity* of the assessment task were associated with decreased model performance, similarly the relationship between levels of *Structure* and model performance showed a somewhat negative linear trend. These findings highlight the importance of considering these construct characteristics when developing ML models for scoring assessments, particularly for higher complexity items and multidimensional assessments.

Keywords Argumentation · Assessment · Machine learning · Construct · Automated scoring · Learning progression

Extended author information available on the last page of the article

Introduction

Assessing students' argumentation, which is one of the critical scientific practices proposed in the *Framework for K-12 Science Education* (National Research Council, 2012) and the *Next Generation Science Standards* (NGSS; NGSS Lead States, 2013), is challenging due to the complexity of the construct (Gane et al., 2018). Aligned with these national calls for the future of science education, it has become necessary to evaluate students' knowledge-in-use and move away from simple memorization of knowledge (Pellegrino & Hilton, 2012). For students to demonstrate knowledge-in-use, the assessment tasks must engage students to analyze, evaluate and critique. Assessing such complex abilities and constructs has gone beyond the capacity of the traditional multiple-choice items, and thus many suggest using performance-based constructed response (CR) assessment items (Krajcik, 2021). CR assessments, requiring students to use their own language to demonstrate knowledge, are widely viewed as providing greater insight into cognition than closed form (e.g., multiple-choice) assessments. In the past, cost and time constraints have made CR assessments significantly more challenging to use in large-enrollment courses than multiple-choice assessments. However, the advances in artificial intelligence (AI) make it feasible to apply these NGSS-aligned assessments in instructional settings, with the potential to have a significant educational impact on students and teachers (Lee et al., 2021; Liu et al., 2016; Zhai et al., 2021a). These and other studies have shown that it is possible to create computerized scoring models that predict human scoring with accuracy measures approaching that of well-trained expert raters, and CR assessments reveal the heterogeneity of student thinking that is difficult to reveal using only traditional multiple-choice items (Zhai et al., 2020).

Scientific argumentation is the practice of reasoning within a domain by constructing and critiquing links between scientific claims and evidence (Osborne et al., 2016). Argumentation is also foundational to scientific inquiry, as it promotes a crucial justification for the evidence and claims used to explain phenomena (Walker & Sampson, 2013). As such, scientific argumentation is also identified as an essential scientific practice for students to learn as part of science education (NGSS Lead States, 2013). Critically, Osborne et al. (2004) found that the development of argumentation in scientific contexts is more challenging than that in socio-cultural contexts. Over the last decades, many studies have examined how to implement scientific argumentation in the classroom to improve scientific practice (Cavagnetto, 2010; Driver et al., 2000). Concurrent with these efforts are a number of studies that investigate assessing students' written scientific arguments, since writing is one of the prominent forms of engaging in argumentation (Lee et al., 2014; McNeill, 2009; Zhai et al., 2023).

Many of the current analytic frameworks for scientific argumentation rely on Toulmin's foundational perspective on argumentation (Sampson & Clark, 2008). Toulmin (1958) suggested essential domain-specific elements within an argument, while recognizing that some elements of the argument are universal across disciplines. Following this framework, different elements within an argument

hold varying functions. Claims are statements that assert a perspective, while data are used to support a claim and warrant to justify the use of data for a claim. Further work has expounded that argumentation activities in science include both constructions of one's own arguments as well as considering or rebutting arguments made by others (Berland & Reiser, 2011; Osborne, 2010). Osborne and colleagues (2016) furthered this idea by proposing a learning progression for how middle school students develop scientific argumentation competence. This progression identifies three different levels based largely on the coordination of argument elements and includes both constructing and critiquing arguments. The progression contends that students with higher levels of scientific argumentation competence are able to craft arguments with more elements, which is more cognitively demanding.

Using technological approaches, like natural language processing and machine learning (ML), to evaluate written arguments has advanced over the last decade (Lippi & Torroni, 2015; Song et al., 2014). Much of this work required detailed schemes to annotate the elements, purpose, relationships, and/or structure of arguments. These schemes are used to evaluate the quality or persuasiveness of arguments in the text (e.g., Chernodub et al., 2019; Nguyen & Litman, 2018). Using these approaches also opens the possibility of evaluating argumentation competence in education, thus providing automatic feedback at the time of learning and as part of formative assessment (Jordan, 2012). Thus, recent reports have included these automated techniques with assessments to support students with immediate and personalized feedback, as they develop proficiency in argumentation. For example, Wambsganss et al. (2020) developed an adaptive *Argumentative Learning* system to provide students with instant feedback on the structure of an argument with text. The authors showed that instant feedback improved both the number of arguments and the perceived persuasiveness of the student text. In a following study, Wambsganss et al. (2022) compared types of automated feedback and demonstrated that including comparison nudging as part of automated feedback leads to higher quality arguments as compared to automated feedback about the arguments or syntactic feedback alone. In other applications, Mirzababaei and Pammer-Schindler (2021) explored the use of a conversational agent with students to detect structural issues in student arguments. They note that the combination of content with argument structural elements may be useful in detecting additional inaccuracies. Finally, Zhu et al. (2020) employed an automated scoring system to produce feedback about scientific arguments within an online earth science module for students at the secondary level. They found that students who received contextualized automated feedback had similar learning gains but spent less time revising than students who received generalized automated feedback. This finding supports the idea that including content and context as part of scientific argumentation is important to help students develop scientific argumentation proficiency.

As part of science assessment, recent work employing AI techniques to assess short, written responses to scientific argumentation and scientific explanation tasks has shown promising results. These short text responses usually contain disciplinary ideas and present distinct challenges to accurately score over longer texts, such as essays used in many argumentation evaluation training corpora (Brew & Leacock,

2013; Shermis, 2015). Numerous studies across various science disciplines and education levels have developed automatic text scoring ML models for such content targeting science assessment items (Zhai et al., 2020). Liu et al. (2014) examined the accuracy of automated scoring of CR concept-based items with multiple levels. Overall, computer model performance showed between moderate to good agreement with human scores. However, the linguistic diversity of middle school students and pronoun resolution (multiple ways to use the same pronoun in the same sentence) contributed to scoring errors (Liu et al., 2014). Mao et al. (2018) utilized an ML-based scoring engine to evaluate and provide feedback on students' short responses to scientific argumentation items. The automated system generates feedback based on whether the student's answer is correct or needs revision. In the following studies, the system feedback was shown to improve students' written scientific arguments (Lee et al., 2019; Zhu et al., 2020). In these cases, the science disciplinary content is integrated as part of the structure of the response (i.e., explanation, argumentation), thus posing challenges for automatic scoring, as well as challenges to develop and align the coding schemes used for student arguments (i.e., content vs. argumentation element).

Relatedly, in a study of machine-scorability for automated scoring of responses to CR items, assessment developers were able to predict with moderate accuracy whether science assessment items would lead to a good ML scoring model (Lottridge et al., 2018). This study posited a framework for examining item scorability by ML models but tested only a small set of science assessment items. The framework included criteria such as the number of concepts targeted by an item, the expected relationship(s) of the concepts in responses, and alignment between the assessment item and rubric. In a recent meta-analysis study, Zhai et al. (2021) expanded on factors that may influence the accuracy of automated scoring for science items by proposing a framework. The framework groups the factors into five key categories, including internal and external features of the assessment, examinee features, machine training and validation approaches, and technical features of the ML model. However, we have not yet realized research that examines how internal features of assessment tasks, which are based on scientific argumentation, relate to machine scoring accuracy. This study adopted a framework specifically constructed for ML-based assessments to fill this gap.

A Framework for Machine Learning-Based Assessment

Cronbach and Meehl (1955) suggest that a construct is a postulated attribute of humans, which can be reflected in test performance. For example, making evidence-based arguments in science is deemed as a type of attribute of competent students. That is, we should be able to infer students' attributes of evidence-based argumentation based on their test performance. In this case, making evidence-based arguments is an assessment construct. Science assessment practices fundamentally deal with evidence to infer the assessment constructs that delineate students' scientific competence (Zhai et al., 2021a). What makes assessment practice complicated is that the assessment construct is usually complicated, diverse, and

with developmental features. In their study examining the intersection of assessment in science and ML tools, Zhai et al. (2020) abstracted three fundamental features of the construct for ML-based science assessments: *complexity, diversity*, and *structure*. In this study, we adopt these three features as our analytical framework to examine how internal features of the assessment may impact ML model performance (Zhai et al., 2021b).

Assessment tasks in science can serve a range of purposes, from informing classroom instruction to programmatic evaluation (Pellegrino et al., 2001). Separately, the *Framework* (National Research Council, 2012) articulates that meaningful science learning in the classroom integrates three dimensions- scientific practices, crosscutting concepts, and disciplinary core ideas. Assessment items that only rely on declarative knowledge cannot fully assess the range of proficiency expected in three-dimensional science learning (Pellegrino et al., 2013). Thus, assessment tasks in science should move beyond factual recall and towards knowledge-in-use, in which students can demonstrate proficiency in using their knowledge to solve problems or think critically about problems (Gane et al., 2018; Harris et al., 2019). These new assessment tasks should engage students in higher-order thinking (Osborne, 2013) and focus on how students use knowledge, which requires them to analyze, evaluate and/or create as part of the assessment in order to demonstrate their proficiency. These kinds of assessments often rely on CR assessment formats so students can construct their answers and consequently, target different cognitive processes than forced-choice assessments (Bennett & Ward, 1993). These differences in how students engage in answering the item are reflected in *Complexity* of the assessment item. These item complexities are relevant within science classrooms, as task complexities are often noticed by teachers, along with dimensions of the NGSS, as part of classroom assessment items (Pellegrino et al., 2018).

Diversity reflects the combinations of different cognitive demands in performing a task. In dealing with three-dimensional science learning (National Research Council, 2012), the assessment is multifaceted. The cognitive demands while solving such a science problem may include multiple components or dimensions (e.g., practices, disciplinary core ideas). As noted above, assessments that engage students in multi-component tasks (e.g., using the law of conservation of energy to develop a scientific argument) are consistent with views of deeper science learning (Pellegrino et al., 2013). However, the number and combinations of these components targeted by the assessment tasks likely affect the sophistication of the construct. The more components engaged in the assessment task, the more diverse the assessment construct is. Students are required to demonstrate knowledge in use during problem-solving by showing their ability to conduct practices and understanding of scientific knowledge simultaneously. Compared to assessing such three-dimensional science learning, some assessments may only focus on one dimension of the construct, which is less diverse.

Zhai et al. (2020) further argue that the assessment construct should reflect cognitive developmental features and denote this feature as *Structure*. That is, students' competence and proficiency with science ideas and practices progress while they receive instruction. For example, research on students' learning progression explicitly lays out the cognitive structure of students' learning and progress (Alonzo &

Steedle, 2009; Osborne et al., 2016; Schwarz et al., 2009). Utilizing AI to assess cognitive models may face additional challenges over assessing descriptive and/or explanatory texts. For example, one may hypothesize that levels in cognitive models may differ in the order or arrangement of ideas, not just in the presence or absence of concepts (Wilson, 2009).

Such different construct characteristics contribute to the challenges of developing accurate ML models for short text responses. Among other reasons which potentially impact the development of ML models, the underlying assessment construct as an internal feature of the assessment, as well as the training sources (Wang et al., 2021), may be particularly critical (Zhai et al., 2020). However, there is limited empirical data available about how differences in these construct features impact ML models, specifically for scientific argumentation assessments, and we have limited knowledge about the fluctuation of the human-machine scoring agreements due to a construct within a given performance assessment. We were interested in how the internal features of the assessment tasks may impact the performance of ML models for scoring student written arguments. Therefore, in this study, we investigated 17 scientific argumentation assessment tasks, which varied in item construct characteristics. We employed the same strategy to train the computer to develop 17 ML scoring models, one for each assessment item, holding the number of training and validation responses constant. We examined the resulting model performance to address the research question: *How are the item construct Complexity, Diversity and Structure associated with ML model performance?*

Materials and Methods

Study Context

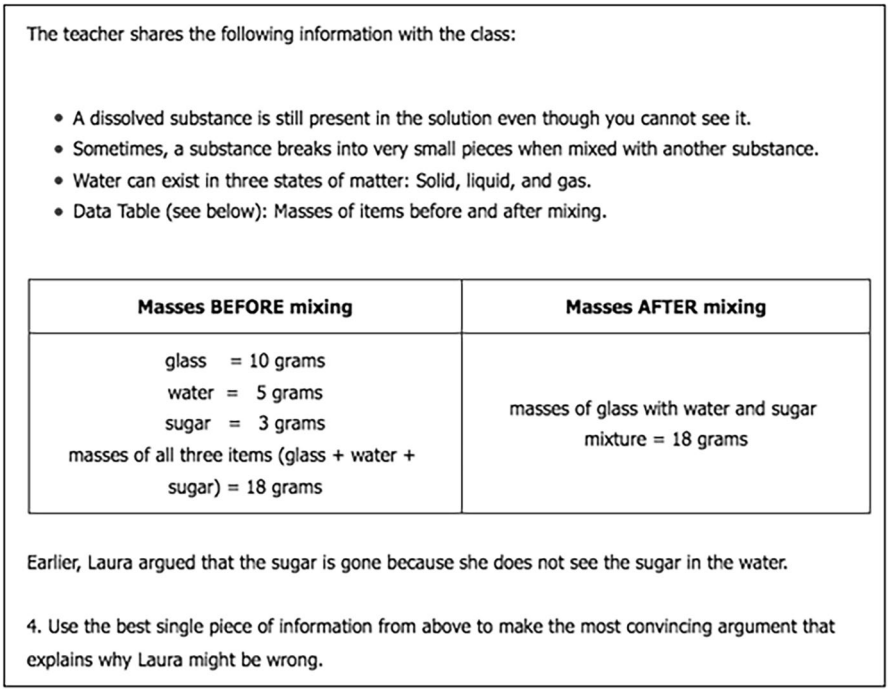
In a prior study (see Wilson et al., 2023), we developed ML scoring models for middle school argumentation assessment tasks, which are aligned to a learning progression of scientific argumentation competency validated by Osborne et al. (2016). A learning progression is based on a developmental approach to learning and describes "successfully more sophisticated ways of reasoning within a content domain" (Smith et al., 2006). For science education, learning progressions are useful for aligning curriculum, assessments and instruction that are consistent with a three-dimensional perspective of science learning (National Research Council, 2012). Therefore, assessing students using learning progression-based assessments can help provide guidance about where students are currently and what support they need to successfully advance in learning. The learning progression that we employed in this study includes three broad levels according to the intrinsic cognitive load. The higher level requires more connections (i.e., degree of coordination) to be made between claims and evidence. That is, level 0 requires no degree of coordination between claims and evidence, while level 1 requires one degree of coordination by means of a warrant. Level 2 requires providing a comparison and/or critique of multiple arguments. This requires students to make two or more degrees of coordination via warrants. Osborne et al. (2016) also considered the construction of students' own arguments

or rebuttal of others' arguments, with the latter more challenging than the prior. They employed some sub-levels to differentiate these cognitive activities for argumentation (details see Osborne et al., 2016).

We used deidentified responses collected previously from 932 students from science classrooms in grades 5–8 from two school districts in California. Students in participating classrooms were given a lesson in scientific argumentation before completing the assessment tasks. The original assessment items were delivered and student responses were collected electronically via Qualtrics. The number of responses varied for each question due to the changes in sequences of questions given to different school districts.

Assessment Tasks

We developed a total of 19 CR assessment items and associated response coding rubrics aligned to a learning progression (details see Wilson et al., 2023). These assessment items targeted different learning progression levels and engaged students in constructing and/or critiquing arguments, while using key disciplinary ideas related to the item context and appropriate for the grade level (see example items; Figs. 1 and 2). For each item, students had to write their own answer; a few items required students to choose between a fictitious character's argument followed by an open response portion


The teacher shares the following information with the class:

- A dissolved substance is still present in the solution even though you cannot see it.
- Sometimes, a substance breaks into very small pieces when mixed with another substance.
- Water can exist in three states of matter: Solid, liquid, and gas.
- Data Table (see below): Masses of items before and after mixing.

Masses BEFORE mixing	Masses AFTER mixing
glass = 10 grams water = 5 grams sugar = 3 grams masses of all three items (glass + water + sugar) = 18 grams	masses of glass with water and sugar mixture = 18 grams

Earlier, Laura argued that the sugar is gone because she does not see the sugar in the water.

4. Use the best single piece of information from above to make the most convincing argument that explains why Laura might be wrong.

Fig. 1 Sugar 4 item

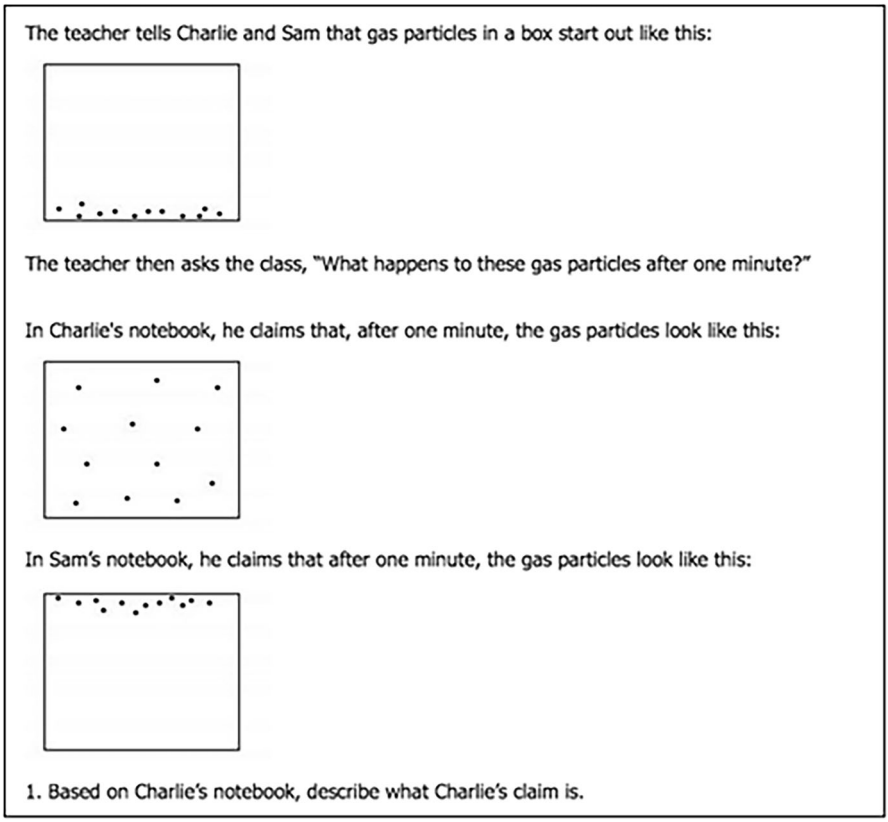


Fig. 2 Gases 1 item

which were later dropped from analysis (see Machine Learning Algorithmic Model Development section below). The items were divided into 3 item sets based on science contexts: five within sugar dissolving in water (S), eight within the kinetic motion of gases (G), and six within bacterial growth (B). Coding rubrics for student responses were developed for each item, and each rubric included two to four codes intended to capture the different quality of student performance in scientific argumentation, including integration of necessary disciplinary ideas, within a given item.

Figure 1 presents a sample assessment task from the sugar dissolving in the sugar dissolving in water context. Previous items in the sugar set established a scenario using arguments from fictitious characters about why sugar cannot be seen when stirred in water. In this item, students are presented with one character's argument, a data table, and a set of statements from a fictitious teacher. Students are asked to critique the fictitious character's argument. This item is aligned with level 2a in the learning progression of Osborne et al. (2016), since it requires students to provide a counter-critique to another person's (i.e., the fictitious student's) argument. The coding rubric for responses to this item had four possible levels to rank students'

responses based on whether they were able to provide a valid critique using the provided evidence.

Coding of the Construct Characteristics of Assessment Tasks

Coding Scheme

We developed a coding scheme for the characteristics of the nineteen assessment items. We adapted dimensions from three existing frameworks (NRC, 2012; Osborne et al., 2016; Zhai et al., 2020) to capture different characteristics relevant to ML, science learning, and scientific argumentation. Using these frameworks, we examined three components (i.e., Complexity, Diversity, and Structure), which identified levels of complexity and sophistication relevant to our item sets (Table 1).

For *Complexity*, we identified four different tasks, ranging in difficulty, embedded in the assessment items. This component captures the cognitive processes the student must engage in to complete the question and is consistent with promoting knowledge in use in science learning (Harris et al., 2019). Items ranged from low-level tasks like identifying provided information in the item to high-level tasks like evaluating multiple pieces of information (Anderson, 2005). Additional examples and descriptions for coding of *Complexity* are provided in Supplemental Table 1. For *Diversity*, we examined if each item engaged students equally in different dimensions of science learning to produce the desired responses; our levels ranged from one to three based on identified dimensions in science learning. Here, we used the dimensions defined in the *Framework for K-12 Science Education* (i.e., disciplinary core ideas, cross-cutting concepts and science and engineering practices; NRC, 2012) to identify dimensions in the item and student responses. We assumed that items that engaged multiple dimensions of science learning (e.g., crosscutting concepts and argumentation) would be more difficult. Since items were designed to assess argumentation, all engaged at least one science practice (Level 1). We examined each assessment task and the associated scoring rubrics to identify if other dimensions were explicitly necessary for a student's response to be awarded the highest score for an item. Finally, for *Structure*, we examined which level of a learning progression for scientific argumentation the item was aligned with. Our item coding scheme ranged over three levels to reflect the critical activities within levels of an empirically validated learning progression.

Example Items and Codes

To illustrate high- and low-level item characteristics, we present two examples of item coding. First, we present item #4 in the sugar context (S4; Fig. 1) as an example of an item with an overall high complexity of underlying assessment constructs. We coded this item as a level 3 in *Structure*, since it requires providing a critique of a character's argument. We coded the item as a level 4 in *Complexity*, since it requires students to consider and evaluate multiple pieces of data and potential statements from the teacher. Then students must decide which piece of evidence to use and

Table 1 Developed coding scheme for three item characteristics relevant to machine-learning scored assessments

Characteristic	Level 1	Level 2	Level 3	Level 4
Complexity	Item provides students with the required knowledge to use in response	Item requires students to possess the knowledge to apply directly in response	Item requires students to use data or information to reach a conclusion	Item requires students to use, evaluate, or integrate multiple data or information to reach a conclusion
Diversity	Engages only one component of three-dimensional learning	Engages two components of three-dimensional learning	Engages three components of three-dimensional learning	N/A
Structure	Item requires identification of claim or evidence or provide a claim or evidence	Item requires the construction of a warrant or complete argument	Item requires a comparative argument, critique of an argument, or a counterclaim	N/A

incorporate the evidence into a critique that appropriately connects to the original argument. Finally, we assigned this item to *Diversity* level 3, since it engages students in all three dimensions of science learning to provide a high-quality response. The item is aligned with the practice of *Engaging in Argument*, specifically critique. This item engages students in disciplinary core ideas related to chemical and physical changes, to make sense of the provided data and provide a valid critique. Finally, it requires students to attend to the crosscutting concept of Energy and Matter, specifically the idea that matter (atoms) is conserved.

We contrast this example with the first item in the context of gas diffusion (G1; Fig. 2). This is the first item in this gas diffusion context, and students are introduced to two fictitious characters and their differing ideas on how gases move within a box. These fictitious characters and their ideas are used repeatedly through the context subsets. The G1 item displays a relatively low complexity of constructs. We coded this item for *Structure* as a Level 1, since it asks students to identify another person's (i.e., character's) claim. We coded the item as a level 1 for *Complexity*, since it requires students to re-iterate information that is presented to them as part of the set-up of the item. Specifically, students must only describe Charlie's model. Finally, we assigned this item a *Diversity* code of level 1, since it engages students directly in only one dimension of science learning to provide a high-scoring response based on the accompanying coding rubric. The item is aligned with the practice of *Engaging in Argument from Evidence* but does not require explicit use of disciplinary ideas or crosscutting concepts to produce a high-quality response. Although subsequent items in this set will require additional dimensions, this item does not require students to directly use their knowledge of particle motion to respond to the prompt. That is, by describing the model provided, the responding student can score well on this item, regardless of if they can use the disciplinary idea of particle motion. Therefore, other science dimensions are not present for this specific item.

Coding of Assessment Tasks

After developing the assessment task coding schemes and defining generic examples for each characteristic at each level, two coders independently coded the nineteen items on all three components. One coder had a Ph.D. in science education and several years of classroom teaching experience at the secondary level; the other coder had a Ph.D. in a science discipline and several years of teaching experience at the post-secondary level. Both coders had several years of experience in assessment design and research. The interrater reliability of this initial coding showed a range of interpretations from none to perfect (Cohen's kappa of 0.11, 0.44 and 1; McHugh, 2012). Most disagreements differed by one level within a characteristic. Due to the small number of assessment items and most disagreements being near misses, the coders decided to resolve all differences by open discussion (Chinh et al., 2019). During these discussions, coders reviewed both the assessment items and coding rubrics to see which elements were necessary for a complete and full answer. They adopted the consensus codes of the items resulting from open discussions as the final codes for item characteristics.

Machine Learning Algorithmic Model Development

Human Coding of Student Responses

For this study, we used deidentified student text responses and associated human codes. Human coding of the argumentation responses was completed during assessment development (described in Wilson et al., 2023). Typical student responses for each item ranged from one short sentence to several sentences. Coding rubrics for each item had a variable number of levels (range 2–4) into which responses could be scored. For each set of items, two coders were trained on the associated coding rubrics for student responses to each item. The coders went through multiple rounds of training using a random subset of 150 student responses to each item. Training rounds were iterated until interrater reliability (Cohen's kappa) between the coders was 0.8 or higher on each rubric component or until three rounds of training were completed. Disagreements on scores during the training round were resolved by consensus discussion between coders. Interrater reliability between human coders (Cohen's kappa) for the last scoring round is provided in Table 2. (For further details on the response coding rubrics and human coding of student responses, please see

Table 2 The characteristics of the assessment tasks, levels in the response coding rubric and the interrater reliability of human coders on student responses

Item	Complexity	Diversity	Structure	# of levels in response rubric	# of responses	H-H reliability (Cohen's kappa)	Evenness
S1	2	2	2	4	775	0.81	0.797
S2	1	1	1	2	765	0.75	0.999
S3	1	1	1	2	763	0.83	0.998
S4	4	3	3	4	754	0.53	0.873
S5	4	3	3	4	744	0.72	0.912
B1	1	1	1	3	549	1.00	0.786
B2	1	1	1	3	527	0.97	0.777
B3	2	3	2	4	498	0.92	0.855
B4*	3	2	3	4	449	0.81	0.732
B5*	4	3	3	4	411	0.97	0.895
B6	2	1	2	3	361	1.00	0.487
G1	1	1	1	2	848	0.82	0.597
G2	1	1	1	2	840	0.85	0.758
G3	3	3	2	4	801	N/A	0.795
G4	3	3	2	3	770	N/A	0.974
G5	4	3	2	3	669	N/A	0.628
G6	4	3	2	3	642	N/A	0.941
G7	3	2	3	3	597	0.75	0.771
G8	4	2	3	4	548	0.77	0.635

* denotes item dropped from further analyses; *H-H* Human-human, *S* Sugar, *B* Bacteria, *G* Gas

Kaldaras & Haudek, 2022 and Wilson et al., 2023). For some items in the gases context, we only have after-discussion consensus scores for the training round and cannot provide Cohen's kappa (N/A in Table 2). The remaining student responses not used during coder training were split into two subsets, and each coder scored one subset of responses independently using the coding rubric.

We calculated a diversity index, evenness (Pielou, 1966), for the human assigned scores to responses to each item. This measure represents the distribution of response scores at the various levels within each coding rubric, where evenness for an item can vary between 1 (indicating responses are equally distributed across all levels of the rubric) to 0 (responses are only in one level of the rubric). By using such a measure, we expected to quickly identify models built on imbalanced data sets (Chawla et al., 2004). In this study, the mean evenness of training sets is 0.80, which indicates that most items had responses close to equally distributed across levels of the rubrics.

Machine Learning Algorithmic Model Development

Since each item had a different number of total student responses collected, we randomly selected a subset of 361 responses for each item in order to have an equal number of responses for training individual ML models. We used a supervised ML text classification approach to assign scores to student-written responses (Aggarwal & Zhai, 2012). During our ML development process, each student response was treated as a document, and the coding rubric indicates a multi-level class. We used natural language processing to extract text features from responses, which were then used as inputs for an ensemble of eight individual algorithms, common in ML classification applications, to generate predicted scores (Jurka et al., 2013). To develop the ML text scoring models, we used the Constructed Response Classifier (CRC) tool, which has been used to produce accurate ML models for short, concept-based constructed responses in science assessment (Jescovitch et al., 2020; Uhl et al., 2021). This CRC tool is part of a set of web-based applications that perform the necessary text parsing and supervised ML algorithms. The R code and documentation are available at https://github.com/BeyondMultipleChoice/AACRAutoReport. The CRC tool allows a user to specify some feature engineering settings, for example, changing the n-gram length, and how numbers in responses are extracted. The ML scoring system then generated predictions on whether each given document was a member of each class. The ML model was generated and validated using a 10-fold cross-validation approach. We chose to use a bag-of-words approach that allowed feature engineering for several reasons, including that such an approach is common in short, concept-based scoring in science assessment. Additionally, in a larger study, this allowed us to iterate on ML model development to finetune models for the specifics of the assessment item and elements of argumentation in a coding rubric. Further, previous studies had used the same approach to develop ML models for scientific argumentation tasks with *very good* performance metrics (see Wilson et al., 2023; Zhai et al., 2023).

During model development, we decided to drop two items in the item context of bacterial growth, since these items were in a different format (i.e., multiple-choice

followed by explaining your answer) than all other items. Since we did not know how the multiple-choice selection would influence student responses, we did not generate ML models for these two items and dropped them from subsequent analyses. We generated an ML model for each of the remaining 17 items individually, using the same set of text processing procedures and input model parameters for each model, including stemming text, removal of stopwords and numbers, and using unigrams and bigrams. In usual practice for developing scoring models, one would optimize ML model performance by tweaking tuning parameters or text processing strategies (Madnani et al., 2017). However, this complicates comparing model performance across assessment items, as the model performance would be dependent on assessment items (and collected responses) as well as the technical parameters used to build the model (Madnani et al., 2017; Shermis, 2015). By using a consistent set of parameters for all models in this study, we aimed to focus on the role of the internal characteristics of the assessment items on ML model performance.

The ML models produced a predicted score for each response that can be compared to the human-assigned score. To evaluate the overall performance of each computer model, we calculated accuracy, Cohen's kappa (k), and Spearman rank-order correlations (r_s, $N = 361$) between the computer-predicted and human-assigned scores (Table 3). We also calculated precision, recall and F1 scores for all models, using macro-averaged precision, recall and F_1 values for multi-class models for items with more than two possible levels in the response rubric (see Table 2). We examined the accuracy of the ML models built using equal numbers of responses in the training set for all models. We used benchmarks of Cohen's kappa of moderate ($k > 0.40$), substantial ($k > 0.60$) to near perfect ($k > 0.80$) agreement between human and computer assigned scores as used by Nehm et al. (2012) for evaluation of the overall human-machine agreements. We conducted all statistical tests, including Kruskal-Wallis H test, Fischer Exact test, and Spearman's correlation, using IBM SPSS Statistics for Windows, version 27 (IBM Corp, Armonk, NY, USA).

Results

Our coding results show a good range of item characteristics and overall levels of the item components (see Table 2). *Structure* showed the most balance, with nearly equal amounts of items over the three levels of scientific argumentation practice. We found that most items engaged students in multiple dimensions of science learning (*Diversity*), with three-dimensional items being the most common. Finally, we found that most items displayed high or low level *Complexity*, with fewer items in the intermediate levels. Below, we report the machining scoring model performance and how machine performance is associated with item features.

Machine Learning Model Performance Across Items

The machine scoring models showed a range of overall performance metrics (Table 3). The mean Cohen's kappa was substantial ($M = 0.60$, $SD = 0.15$), and

Table 3 Machine learning model performance measures by argumentation items

Item	*k*	Accuracy	r_s*	Precision[a]	Recall[a]	F_1[a]
S1	0.62	0.77	0.788	0.68	0.54	0.57
S2	0.75	0.87	0.740	0.86	0.88	0.87
S3	0.80	0.90	0.811	0.91	0.91	0.91
S4	0.55	0.73	0.715	0.71	0.57	0.58
S5	0.55	0.69	0.708	0.62	0.56	0.56
B1	0.89	0.94	0.947	0.85	0.75	0.78
B2	0.80	0.89	0.841	0.83	0.88	0.85
B3	0.46	0.65	0.685	0.61	0.51	0.51
B6	0.61	0.91	0.724	0.78	0.62	0.68
G1	0.62	0.91	0.619	0.71	0.63	0.67
G2	0.65	0.89	0.653	0.78	0.67	0.72
G3	0.46	0.65	0.650	0.61	0.62	0.61
G4	0.51	0.69	0.641	0.69	0.66	0.67
G5	0.62	0.83	0.641	0.84^	0.53	0.81^
G6	0.40	0.65	0.556	0.59	0.56	0.56
G7	0.49	0.73	0.470	0.55	0.42	0.43
G8	0.38	0.75	0.488	0.74^	0.36	0.72^

k = Cohen's kappa; r_s = Spearman's rho

*All r_s values $p < .01$

[a] For items with > 2 levels in response rubric, the macro-averaged value for a metric is reported

^ A coding level was dropped in the macro average since the machine made zero predictions in a level for this item

the mean accuracy was fairly high ($M = 0.79$, $SD = 0.11$). We developed five ML models for items in the sugar context with an average *k* of 0.65 and a range of 0.55 to 0.80. For the four bacteria items in the analysis, we developed an ML model for each with an average *k* of 0.69 and a range of 0.46 to 0.89. The eight gas items show decreased model performance, with a moderate *k* of 0.52 and a range of 0.38 to 0.65. Other performance metrics for the eight gas items (precision, recall and F_1 score) were also less than models for other item contexts. The gas items proved the most challenging to develop scoring models for, as model performance ranged from moderate to substantial (Nehm et al., 2012). We also found that the highest performing models in the gas context had items with only 2 or 3 scoring levels.

Since the coding scheme was ordinal in nature, we also calculated Spearman's rho, a non-parametric measure of correlation, as a measure of model accuracy to account for "near misses" in the computer scoring. We found that all but one model in the sugar and bacteria contexts had a rho over 0.70, which has been suggested as a threshold for acceptable ML model performance. As expected, this one model with low correlation (B3) also had the lowest accuracy and human-machine agreement measures for these contexts. Interestingly, this item had fairly

high human-human IRR measures and was at level 1c in the learning progression. We note that other items in these two contexts showed better model performance even when at a higher learning progression level (e.g., S4) or exhibited lower human-human IRR (e.g., S2) for the training set. Surprisingly, none of the models for items in the gas context produced a rho greater than 0.70, although five of these models had rho greater than 0.60. Despite these challenges, correlations between human and machine scores were significant at the $p < .01$ level. From these findings, along with the human-machine Cohen's kappa results, we conclude that we have well-performing models for items in the sugar and bacteria contexts, with a range of model performance for items in the gases context. For all remaining analyses, we used Cohen's kappa as the measure of model performance, since it corrects for "chance" agreement between coders (i.e., human & computer; McHugh, 2012). We performed a Kruskal-Wallis H Test to examine if the context of the item influenced the ML model performance, as measured by Cohen's kappa. No significant differences (Chi-square = 3.55, p = .169, $df = 2$) were found among the three item contexts. Therefore, we collapsed items and models from all contexts into a single data set for further analysis.

Further, we looked at the relationship between evenness or the distribution of responses in the levels of an item rubric and model performance. Surprisingly, there was no correlation between the balance of responses across rubric levels and Cohen's kappa of the resulting model (Pearson's $r = .08$, $p = .765$, $N = 17$). Similarly, there were no significant correlations between the evenness of responses and recall, precision or F_1 of the resulting models, although we found a weak, positive correlation between recall and evenness (Pearson's $r = .39$, $p = .118$, $N = 17$). The finding indicates that machine identification of more positive cases in the data set may be a little more sensitive to data balance for this argumentation coding scheme.

Association of Item Complexity with Model Performance

To examine if there was any association between the different construct characteristics of the items, we conducted a series of pairwise Fisher's exact tests. All three pairwise tests (Complexity by Diversity; Complexity by Structure and Diversity by Structure) returned significant results (Fisher's exact test 2-sided, $p < .01$), suggesting these three characteristics are not truly independent in this sample of items. This is unsurprising, in that items that ask students to critique arguments (i.e., high level of *Structure*) generally require students to analyze or evaluate (i.e., higher level of *Complexity*) as opposed to using memorized information, for example. Since the three construct characteristics were not independent, we could not combine all item variables (i.e., characteristics and levels) into a single statistical model. Instead, to answer our research question, we examined how the different characteristics of construct complexity are associated with model performance, respectively.

First, we examined the model performance using Cohen's kappa by the level of *Complexity* (Fig. 3). We found that models for items at level 1 of *Complexity* had better performance than models for items at all other levels. Our finding suggests that all models at level 1 of *Complexity* achieved substantial agreement as measured

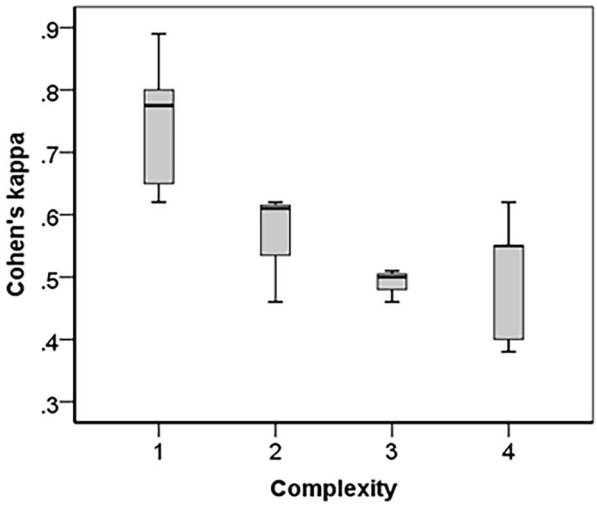


Fig. 3 A box plot of model performance for items exhibiting different levels of Complexity. The shaded box represents middle 50% values, and whiskers extend to maximum and minimum values. The thick black line represents the median value

by Cohen's kappa ($k > 0.60$). On the other hand, we found that a majority of models for higher *Complexity* items (levels 2, 3, and 4) did not achieve even this substantial performance threshold. The median model performance indicated by Cohen's *k* for level 2 items was slightly higher than model performance for levels 3 and 4. For level 2 items, we found that most models had a performance in the range of 0.5 to 0.6, which is in moderate to substantial agreement. We found that models for level 3 items had the narrowest range of performance while level 4 items showed a much larger range of performance. Spearman's rho correlation coefficient was used to examine the relationship between Cohen's kappa and the level of item *Complexity*. There was a significant negative correlation between the two variables ($r_s = -0.752$, $p < .001$, $N = 17$), indicating that the more complex the construct, the lower amount of agreement between machine and human scores.

Association of Item Diversity with Model Performance

Next, we examined the model performance using Cohen's kappa by levels of *Diversity*, or the number of science learning dimensions (Fig. 4). We found that models for items at level 1 of *Diversity* had better performance than models for items that engaged multiple dimensions of science learning (levels 2 and 3). We found that all models at level 1 of this characteristic were above substantial agreement ($k > 0.60$). Conversely, we found that only a few models for higher *Diversity* items achieved even the substantial performance benchmark. The median model performance for level 2 items was slightly lower than median model performance for level 3; otherwise, the model performance for levels 2 and 3 items was very similar. Models for level 3 *Diversity* items showed a slightly narrower range of performance, with nearly all models exhibiting moderate to substantial agreements. Spearman's rho correlation coefficient was used to examine the relationship between Cohen's kappa and the level of item *Diversity*. There was a significant negative correlation between these

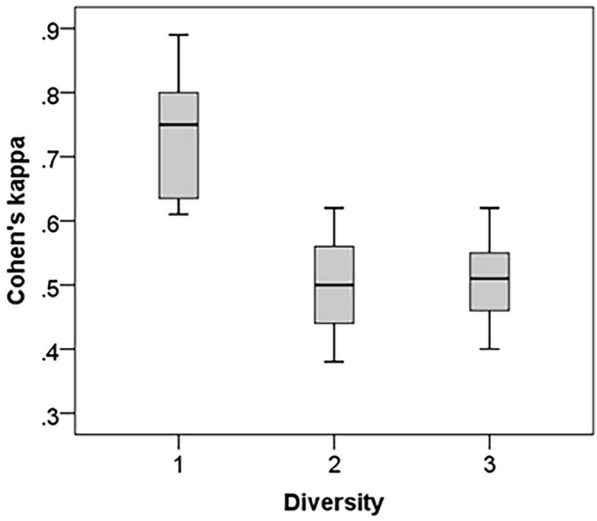


Fig. 4 A box plot of model performance for items exhibiting different levels of Diversity

two variables ($r_s = -0.718$, $p < .01$, $N = 17$), indicating that the more diverse the construct, the less human-machine scoring accuracy.

Association of Item Structure on Model Performance

We examined how the *Structure* of the constructs, aligned to cognitive levels of a learning progression, was associated with model performance (Fig. 5). We found that, in general, models for items at structure levels 2 and 3 had lower Cohen's kappa than for items at level 1. All level 1 items for structure exceeded the substantial Cohen's kappa benchmark. We found that the measures of the average of

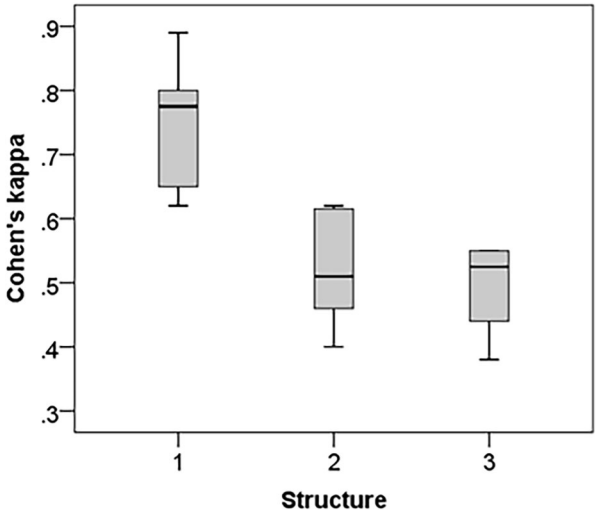


Fig. 5 A box plot of model performance for items exhibiting different levels of Structure

Cohen's kappa for models at levels 2 and 3 are nearly the same, but level 2 models show a larger range of performance. For level 2 items, nearly all models exceeded the moderate performance benchmark, with a few models exceeding the substantial agreement benchmark. The level 3 items showed a lower range of performance with no model exceeding the substantial performance threshold. Spearman's rho correlation coefficient was used to examine the relationship between Cohen's kappa and the level of item diversity. There was a significant negative correlation between the two variables ($r_s = -0.764$, $p < .001$, $N = 17$), indicating the more complicated structure of the construct, the lower agreement between machine and human scores.

Discussion

Though ML has great potential to be widely applied in science assessments, the accuracy of machine scoring of short, concept-based, text responses remains largely a "black box." (Rupp, 2018). Previous work had shown that assessment developers could predict with some accuracy whether a science assessment item would have a good ML scoring model for a small set of written science assessment items (Lottridge et al., 2018). A recent critical review suggested that examining factors in a systematic way that impact machine scoring and developing generalizable algorithms is critical to further increase the usability of ML in the science assessments (Zhai et al., 2020). With this in mind, the current study examined a critical internal feature of assessments (Zhai et al., 2021b), characteristics of the target construct, and how these features are associated with machine scoring accuracy. Aligned to an ML-based assessment framework (Zhai et al., 2020b), we identified three critical characteristics of the assessment constructs: complexity, diversity, and structure. Our empirical findings suggest that construct characteristics do associate with ML model performance. We found negative and significant correlations between assessment item characteristics and ML model performance for all three components we examined. The difference in performance is most pronounced at the lowest level or least sophisticated for each characteristic when compared to all other higher levels. That is, for a given construct characteristic, we observed large differences in model performance when comparing level 1 to any of the higher levels but less difference in performance between higher levels (e.g., level 2 vs. level 3 items). It could be that tuning model parameters, as is usual practice in developing ML models (Madnani et al., 2017), may improve model performance for mid-level constructs more so than higher-level constructs.

We examined each item characteristic and level independently to determine the association with scoring model performance. First, we looked at four levels of item *Complexity*, the way students use knowledge in the item to generate a response. We found that higher levels of *Complexity* had decreased machine scoring performance; for the most part, the four levels of complexity showed decreasing performance in a fairly negative linear relationship, with level 4 showing a large range of different model performances. As expected, tasks with simpler activities (i.e., using the information provided in the item) showed better machine scoring performance than scoring performance for more complex tasks, likely due to less lexical diversity in the

text of responses when students use information from the item (Shermis, 2015). In their previous study, Zhai et al. (2020) found the majority of CR items scored with ML models were of relatively low complexity. In our study, we had more higher-level items, which were associated with lower ML model accuracies. This represents a continued challenge for applying AI in science assessment: the need to develop advanced machine algorithmic models better suited to accurately score higher complexity assessment items.

As current educational systems move to teaching and learning of multi-dimensional science (e.g., NGSS), one of the challenges is how to assess this learning at a large scale (Anderson et al., 2018). Following, we also studied the number of science learning dimensions, or *Diversity*, engaged by an assessment item. Again, we found that higher levels of *Diversity* had decreased model performance; this was most obvious when comparing the accuracy of models for level 1 items to models for items at levels 2 and 3. This finding provides evidence that producing ML models for single dimension science items is feasible. On the other hand, we found very little difference in model performance between items at levels 2 and 3. This suggests that the challenge of developing ML models for performance assessments is for multidimensional items, not necessarily a specific dimension or number of dimensions. Although we recognize that as the number of dimensions targeted by an item increase, it is likely the number of expected concepts or ideas that must be demonstrated will also increase, which can also affect model performance (Lottridge et al., 2018). As has been noted by others, one of the challenges of scoring three-dimensional science responses is the expected integration of knowledge across dimensions, and that a single holistic score may represent student facility with different dimensions (Gane et al., 2018; Maestrales et al., 2021). In such cases, a sufficiently large training set of responses may be required for the ML model to recognize multiple patterns for the same score (Wang et al., 2021). Or, alternative coding schemes may be necessary in order to pinpoint specific patterns and student successes and challenges in the integration of dimensions.

We chose to use a previously validated assessment instrument for this study to ensure the items themselves were of sufficient quality. Thus, we did not have an equal number of items in each characteristic level, nor every possible combination of levels within this set of 19 items. As a result, levels in item characteristics display correlation. Future work could address this limitation and further validate the proposed item characteristics and levels by purposefully developing a larger set of items, with items distributed across each characteristic level.

For our third characteristic, we examined levels of item *Structure*, or the alignment to a cognitive model of development, as documented in a learning progression for scientific argumentation. We found that higher levels of *Structure* had decreased model performance in a somewhat negative linear relationship, with level 2 showing a larger range of different model performances than level 3. As expected, items aligned to lower levels of the progression showed better model performance than models for more complicated argumentation tasks, like creating a comparative argument. This aligns with empirical evidence for the learning progression itself. Higher levels of argumentation require more components to be successful (e.g., comparative statements, additional warrants) as well as for the components to be structured

appropriately (Zhai et al., 2023). Further, our findings in this study about item characteristics complement previous work examining the validity evidence for automated scores of argumentation components contained in a coding rubric. Kaldaras and Haudek (2022) found that specific components of a coding rubric, including those intended to capture critiques and reasoning which align to higher levels of the learning progression, often have more mis-scores by automated scoring models. These findings are also consistent with previous studies in automatic scoring of argumentation. Lee et al. (2019) found that having students elaborate and connect data with their reasoning was more difficult than identifying a claim. These higher argumentation skills are not only challenging for students to master, but the additional components and structure of the text in responses to these higher-level items require larger training sets or additional syntactic features of the text. In the current study, we chose to use an ensemble, bag-of-words approach to developing machine scoring models since this approach had been successfully used before for this assessment instrument (Zhai et al., 2023; Wilson et al., 2023). Further, this approach for the current and previous studies was chosen, in part, to allow for feature engineering of models to attend to specific features of argumentation and aid construct evaluation. Although newer, approaches to text classification (e.g., transformers) could also be employed, these approaches may have limited additional benefits for simple scoring tasks given the additional computational demands (Mayfield & Black, 2020). Nevertheless, as deep learning techniques continue to advance, these newer approaches are likely to address many of the specific challenges to scoring scientific argumentation assessments.

Building from recent findings from argument mining may be a promising way forward to score argumentation tasks (Lawrence & Reed, 2020), but these general strategies must also be integrated with the dimensions of science learning and may not be generalizable to other science practices. Other technical approaches to automatic scoring to detect constructed causal relationships in student text may be a promising way to capture the quality of explanation or argument (Wiley et al., 2017). Further, a recent study used machine learning to construct cognitive models of student learning in science writing using both assessment data and cognitive attributes (Lamb et al., 2021). We are excited by this possibility of leveraging machine learning during the creation of cognitive models, as opposed only to validation efforts. Based on the findings here, it is likely that using machine learning for identifying and proposing cognitive models and constructs will require larger sets of student data, especially as the cognitive models and constructs become more complicated.

Finally, we also examined a few characteristics of the labeled data sets used to train each ML model. We note a positive association between human-human interrater reliability and human-computer interrater reliability, as has been noted before (Powers et al., 2015; Williamson et al., 2012). This confirms the importance of having a well-designed coding rubric with discrete criteria in order to maximize human-machine scoring agreement on the training set. This also raises the possibility that some of the item complexity may also affect human coding reliability. As item and response complexity increases, it may be more difficult for human coders to assign scores to responses, as the structure of language in responses becomes more complex, or meaning is inferred by readers (Zhai et al., 2020a).

We found that items in the Gases context had slightly poorer ML model performance overall, although not significantly different from the other contexts. This may be due in part to having the fewest level 1 items (by percentage) in the Gases context. Using more items with higher level construct characteristics may have decreased the overall average model performance in this context. Further, this finding that Gases items have lower model performance is also consistent with other reported findings. Kaldaras and Haudek (2022) report that ML training using a holistic rubric (or multi-class scheme) had several items in the Gases context (G7, G7 and G8) showing larger discrepancies between machine and human assigned scores than other items in the assessment. We also found that models for these items had overall lower performance metrics. Kaldaras and Haudek (2022) note that rubrics for these items identify student reasoning about particle level motion. Since these disciplinary ideas can be more complex and such reasoning can be expressed in several ways, this may lead to lower performance of ML models in properly identifying these integrated concepts and argument elements. Since we also adopted the holistic scoring scheme in this study and did not tune models for this issue, some of the lower performance of the models in the Gases context is likely due to the same issues. This reinforces the need for larger training sets, especially for some more complex items, in order to accurately identify this reasoning.

Further, we looked at the relationship between evenness or the distribution of responses at the levels of an item rubric and model performance. Previous studies have suggested that very infrequent occurrence of student ideas or specific score levels may decrease the accuracy of automated scoring for these levels (Ha et al., 2011; Liu et al., 2016). Surprisingly, there was no association between the balance of response across levels and model performance in this study. One possible reason for this is that all of our items displayed a distribution of responses across coding levels that were above some threshold, which serves as a "lower limit" for training scoring models. Although, we note that several of our items (e.g., B3, G5, S1) had specific levels with the coding rubric that contained < 6% of the responses and therefore provided very few examples of a given level in the training set. Another possible interpretation of this finding is that the diversity index of evenness we used was not sensitive enough to differences in code distributions.

A full examination of student performance on the set of items used in this study is reported elsewhere (Wilson et al., 2023). However, we did check to see the relationship between the difficulty of the item itself, based on the human assigned score to each response, and the model scoring accuracy. Although we found a positive association between item average scores and model accuracy, it was not statistically significant, suggesting that item characteristics influence ML model performance and not just student performance. Finally, there are numerous annotation schemes for argumentation assessments, which define various argument structures, elements and/or qualities (e.g., Stab & Gurevych, 2017; Visser et al., 2022). Here, we have chosen to use annotations in our training set derived from coding rubrics, which integrate content and practice and were originally developed and validated as part of the science assessment instrument. Future work could examine the same item characteristics but use a different argumentation annotation scheme on the responses as

part of the training set to examine how prompt characteristics influence these other evaluations of argumentation.

Conclusions

Our study has provided evidence that some internal assessment features of scientific argumentation assessment are associated with ML-model scoring performance. In this study, we extracted three characteristics of the assessment construct: complexity, diversity, and structure and found negative and significant correlations between these characteristics and the performance of associated ML scoring models. It is critical to identify and understand these effects in order to understand the possibilities and limitations of using AI-based technologies, like ML-based scoring, for the evaluation of contextual rich science assessments. As science teaching and learning move to align with multidimensional learning advocated by the NGSS, accurately classifying multiple dimensions of science learning using AI and ML models will be necessary (Maestrales et al., 2021). This also highlights a challenge to furthering the use of ML-based assessments in science by aligning these assessments with models in cognitive development (Zhai et al., 2020). Extrapolating from our findings, producing ML models for increasingly sophisticated cognitive abilities will take additional model tuning, more iterative development cycles, novel technical features of text processing, and/or larger or selected training sets of labeled responses. If we wish to develop and deploy AI-based assessments in science education at scale, then it is critical to move away from designing items and automated scoring models on a one-by-one basis aligned to scattered constructs or attempting to generate scoring models *post hoc* to existing assessments, but rather move towards integrating design theories with assessment practices and incorporate all phases of assessment into a validity process (Gane et al., 2018; Zhai et al., 2021a). Further, for assessment developers to learn what works across assessment items, constructs and contexts, we must focus not only on outcomes of the models (e.g., accuracy) but also on technical features of the model as well as item characteristics (Zhai et al., 2021b, 2020c). Finding these common features of success and challenge is likely to lead to faster development and wider implementation of ML-based assessments as part of formative assessment in science classrooms.

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/s40593-023-00385-8.

Acknowledgements We thank members of the Automated Analysis of Constructed Response project for helpful conversations about this study and comments on early drafts of this manuscript. We thank all members of the Argulex project at Michigan State University, Stanford University and BSCS Science Learning for work on developing items and coding responses.

Authors' Contributions K.H. and X.Z. conceived the research question and study design. K.H. performed the analyses and K.H. and X.Z. contributed to interpretation of results. K.H. drafted the manuscript and K.H. and X.Z. revised the manuscript. K.H. and X.Z. read and approved of the final manuscript.

Funding This material is based upon work supported by the National Science Foundation under Grants No. 1323162, 1561159 and 2200757.

Data Availability The datasets used and/or analysed during the current study are available from the corresponding author on reasonable request.

Declarations

Competing Interests The authors declare that they have no competing interests.

References

Aggarwal, C. C., & Zhai, C. (2012). *Mining text data*. Kluwer Academic Publishers.

Alonzo, A. C., & Steedle, J. T. (2009). Developing and assessing a force and motion learning progression. *Science Education, 93*(3), 389–421. https://doi.org/10.1002/sce.20303

Anderson, C. W., de los Santos, E. X., Bodbyl, S., Covitt, B. A., Edwards, K. D., Hancock, I. I., Lin, J. B., Thomas, Q. M., Penuel, C., & Welch, M. M. (2018). Designing educational systems to support enactment of the Next Generation Science standards. *Journal of Research in Science Teaching, 55*(7), 1026–1052. https://doi.org/10.1002/tea.21484

Anderson, L. W. (2005). Objectives, evaluation, and the improvement of education. *Measurement Evaluation and Statistical Analysis, 31*(2), 102–113. https://doi.org/10.1016/j.stueduc.2005.05.004

Bennett, R. E., & Ward, W. C. (Eds.). (1993). *Construction versus choice in cognitive measurement: Issues in constructed response, performance testing, and portfolio assessment*. L. Erlbaum Associates.

Berland, L. K., & Reiser, B. J. (2011). Classroom communities' adaptations of the practice of scientific argumentation. *Science Education, 95*(2), 191–216. https://doi.org/10.1002/sce.20420

Brew, C., & Leacock, C. (2013). Automated short answer scoring. In M. D. Shermis, & J. Burstein (Eds.), *Handbook of automated essay evaluation*. Routledge. https://doi.org/10.4324/9780203122761.ch9

Cavagnetto, A. (2010). Argument to Foster Scientific Literacy. *In Review of Educational Research* 80,(3), 336–371).

Chawla, N. V., Japkowicz, N., & Kotcz, A. (2004). Special issue on learning from Imbalanced Data sets. *SIGKDD Explor Newsl, 6*(1), 1–6. https://doi.org/10.1145/1007730.1007733

Chernodub, A., Oliynyk, O., Heidenreich, P., Bondarenko, A., Hagen, M., Biemann, C., & Panchenko, A. (2019). TARGER: Neural Argument Mining at Your Fingertips. *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, 195–200. https://doi.org/10.18653/v1/P19-3031

Chinh, B., Zade, H., Ganji, A., & Aragon, C. (2019). Ways of Qualitative Coding: A Case Study of Four Strategies for Resolving Disagreements. *Extended Abstracts of the 2019 CHI Conference on Human Factors in Computing Systems*, 1–6. https://doi.org/10.1145/3290607.3312879

Cronbach, L. J., & Meehl, P. E. (1955). Construct validity in psychological tests. *Psychological Bulletin, 52*(4), 281–302. https://doi.org/10.1037/h0040957

Driver, R., Newton, P., & Osborne, J. F. (2000). Establishing the norms of scientific argumentation in classrooms. *Science Education, 84*(3), 287–312.

Gane, B. D., Zaidi, S. Z., & Pellegrino, J. W. (2018). Measuring what matters: Using technology to assess multidimensional learning. *European Journal of Education, 53*(2), 176–187. https://doi.org/10.1111/ejed.12269

Ha, M., Nehm, R. H., Urban-Lurain, M., & Merrill, J. E. (2011). Applying computerized-scoring models of written biological explanations across courses and colleges: Prospects and limitations. In *CBE Life Sci Educ* (Vol. 10, Issue 4, pp. 379–393). https://doi.org/10.1187/cbe.11-08-0081

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. https://doi.org/10.1111/emip.12253

Jescovitch, L. N., Scott, E. E., Cerchiara, J. A., Merrill, J., Urban-Lurain, M., Doherty, J. H., & Haudek, K. C. (2020). 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*, 150–167. https://doi.org/10.1007/s10956-020-09858-0

Jordan, S. (2012). Student engagement with assessment and feedback: Some lessons from short-answer free-text e-assessment questions. *Computers & Education, 58*(2), 818–834. https://doi.org/10.1016/j.compedu.2011.10.007

Jurka, T. P., Collingwood, L., Boydstun, A. E., Grossman, E., & Van Atteveldt, W. (2013). RTextTools: A supervised learning package for text classification. *The R Journal, 5*(1), 6–12.

Kaldaras, L., & Haudek, K. C. (2022). Validation of automated scoring for learning progression-aligned next generation science standards performance assessments. *Frontiers in Education, 7*. https://www.frontiersin.org/articles/10.3389/feduc.2022.968289. Accessed 8 Nov 2023.

Krajcik, J. S. (2021). Commentary—applying machine learning in science assessment: Opportunity and challenges. *Journal of Science Education and Technology, 30*(2), 313–318. https://doi.org/10.1007/s10956-021-09902-7

Lamb, R., Hand, B., & Kavner, A. (2021). Computational modeling of the effects of the Science writing heuristic on student critical thinking in Science using machine learning. *Journal of Science Education and Technology, 30*(2), 283–297. https://doi.org/10.1007/s10956-020-09871-3

Lawrence, J., & Reed, C. (2020). Argument mining: A Survey. *Computational Linguistics, 45*(4), 765–818. https://doi.org/10.1162/coli_a_00364

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. https://doi.org/10.1007/s10956-020-09889-7

Lee, H. S., Liu, O. L., Pallant, A., Roohr, K. C., Pryputniewicz, S., & Buck, Z. E. (2014). Assessment of uncertainty-infused scientific argumentation. *Journal of Research in Science Teaching, 51*(5), 581–605. https://doi.org/10.1002/tea.21147

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

Lippi, M., & Torroni, P. (2015). Argument mining: A machine learning perspective. In E. Black, S. Modgil, & N. Oren (Eds.), *Theory and applications of formal argumentation* (pp. 163–176). Springer International Publishing.

Liu, O. L., Brew, C., Blackmore, J., Gerard, L., Madhok, J., & Linn, M. C. (2014). Automated scoring of constructed-response science items: Prospects and obstacles. *Educational Measurement: Issues and Practice, 33*(2), 19–28. https://doi.org/10.1111/emip.12028

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. https://doi.org/10.1002/tea.21299

Lottridge, S., Wood, S., & Shaw, D. (2018). The effectiveness of machine score-ability ratings in predicting automated scoring performance. *Applied Measurement in Education, 31*(3), 215–232. https://doi.org/10.1080/08957347.2018.1464452

Madnani, N., Loukina, A., & Cahill, A. (2017). A Large Scale Quantitative Exploration of Modeling Strategies for Content Scoring. *Proceedings of the 12th Workshop on Innovative Use of NLP for Building Educational Applications*, 457–467. https://aclweb.org/anthology/W/W17/W17-5052.pdf

Maestrales, S., Zhai, X., Touitou, I., Baker, Q., Schneider, B., & Krajcik, J. (2021). Using machine learning to score multi-dimensional assessments of Chemistry and Physics. *Journal of Science Education and Technology, 30*(2), 239–254. https://doi.org/10.1007/s10956-020-09895-9

Mao, L., Liu, O. L., Roohr, K., Belur, V., Mulholland, M., Lee, H. S., & Pallant, A. (2018). Validation of Automated Scoring for a Formative Assessment that Employs Scientific Argumentation. In *Educational Assessment* (Vol. 23, Issue 2, pp. 121–138). https://doi.org/10.1080/10627197.2018.1427570

Mayfield, E., & Black, A. W. (2020). Should You Fine-Tune BERT for Automated Essay Scoring? *Proceedings of the Fifteenth Workshop on Innovative Use of NLP for Building Educational Applications*, 151–162. https://doi.org/10.18653/v1/2020.bea-1.15

McHugh, M. L. (2012). Interrater reliability: The kappa statistic. In *Biochemia Medica* (Vol. 22, Issue 3, pp. 276–282). https://doi.org/10.11613/BM.2012.031

McNeill, K. L. (2009). Teachers' use of curriculum to support students in writing scientific arguments to explain phenomena. *Science Education, 93*(2), 233–268. https://doi.org/10.1002/sce.20294

Mirzababaei, B., & Pammer-Schindler, V. (2021). Developing a conversational agent's capability to identify structural wrongness in arguments based on toulmin's model of arguments. *Frontiers in Artificial Intelligence*. https://doi.org/10.3389/frai.2021.645516

National Research Council. (2012). *A Framework for K-12 Science Education: Practices, crosscutting concepts, and Core ideas*. The National Academies Press. http://www.nap.edu/openbook.php?record_id=13165. Accessed 8 Nov 2023.

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. https://doi.org/10.1007/s10956-011-9300-9

Nguyen, H., & Litman, D. (2018). Argument Mining for Improving the Automated Scoring of Persuasive Essays. *Proceedings of the AAAI Conference on Artificial Intelligence*, *32*(1). https://doi.org/10.1609/aaai.v32i1.12046

NGSS Lead States. (2013). *Next Generation Science standards: For States, by States*. The National Academies Press.

Osborne, J. F. (2010). Arguing to learn in science: The role of collaborative, critical discourse. *Science, 328*, 463–466.

Osborne, J. (2013). The 21st century challenge for science education: Assessing scientific reasoning. *Thinking Skills and Creativity, 10*, 265–279. https://doi.org/10.1016/j.tsc.2013.07.006

Osborne, J., Erduran, S., & Simon, S. (2004). Enhancing the quality of argumentation in school science. *Journal of Research in Science Teaching, 41*(10), 994–1020. https://doi.org/10.1002/tea.20035

Osborne, J. F., Henderson, J. B., MacPherson, A., Szu, E., Wild, A., & Yao, S. Y. (2016). The development and validation of a learning progression for argumentation in science. *Journal of Research in Science Teaching, 53*(6), 821–846. https://doi.org/10.1002/tea.21316

Pellegrino, J. W., Chudowsky, N., & Glaser, R. (Eds.). (2001). *Knowing what students know: The science and design of educational assessment*. National Academy Press. http://books.nap.edu/books/0309072727/html/index.html. Accessed 8 Nov 2023.

Pellegrino, J. W., Gane, B. D., Zaidi, S. Z., Harris, C. J., McElhaney, K., Alozie, N., Pennock, H., Severance, P., Neumann, S., Fortus, K., Krajcik, D., Nordine, J. S., Furtak, J., Briggs, E., Chattergoon, D., Penuel, R., Wingert, W. R., K., & Van Horne, K. (2018). The challenge of assessing "knowledge in use": Examples from three-dimensional science learning and instruction. *Proceedings of the 13th International Conference of the Learning Sciences (ICLS) 2018*, *2*, 1211–1218.

Pellegrino, J. W., & Hilton, M. (2012). *Education for life and work: Developing transferable knowledge and skills in the 21st century*. National Academies Press.

Pellegrino, J. W., Wilson, M. R., Koenig, J. A., & Beatty, A. S. (2013). *Developing assessments for the next generation science standards*. National Academies Press.

Pielou, E. C. (1966). The measurement of diversity in different types of biological collections. *Journal of Theoretical Biology, 13*, 131–144. https://doi.org/10.1016/0022-5193(66)90013-0

Powers, D. E., Escoffery, D. S., & Duchnowski, M. P. (2015). Validating automated essay Scoring: A (Modest) refinement of the Gold Standard. *Applied Measurement in Education, 28*(2), 130–142. https://doi.org/10.1080/08957347.2014.1002920

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

Sampson, V., & Clark, D. B. (2008). Assessment of the ways students generate arguments in science education: Current perspectives and recommendations for future directions. In *Science Education* (Vol. 92, Issue 3, pp. 447–472). https://doi.org/10.1002/sce.20276

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. https://doi.org/10.1002/tea.20311

Shermis, M. D. (2015). Contrasting State-of-the-Art in the Machine Scoring of Short-Form Constructed Responses. In *Educational Assessment* (Vol. 20, Issue 1, pp. 46–65). https://doi.org/10.1080/10627197.2015.997617

Smith, C. L., Wiser, M., Anderson, C. W., & Krajcik, J. (2006). Implications of Research on Children's learning for standards and Assessment: A proposed learning progression for Matter and the Atomic-Molecular Theory. *Measurement: Interdisciplinary Research and Perspectives, 4*(1–2), 1–98. https://doi.org/10.1080/15366367.2006.9678570

Song, Y., Heilman, M., Klebanov, B., B., & Deane, P. (2014). Applying argumentation schemes for essay scoring. *Proceedings of the First Workshop on Argumentation Mining*, 69–78. https://doi.org/10.3115/v1/W14-2110

Stab, C., & Gurevych, I. (2017). Parsing argumentation structures in persuasive essays. *Computational Linguistics, 43*(3), 619–659. https://doi.org/10.1162/COLI_a_00295

Toulmin, S. (1958). *The uses of argument*. Cambridge University Press.

Uhl, J. D., Sripathi, K. N., Meir, E., Merrill, J., Urban-Lurain, M., & Haudek, K. C. (2021). Automated writing assessments measure undergraduate learning after completion of a computer-based Cellular respiration Tutorial. *CBE—Life Sciences Education, 20*(3), ar33. https://doi.org/10.1187/cbe.20-06-0122

Visser, J., Lawrence, J., Reed, C., Wagemans, J., & Walton, D. (2022). Annotating Argument Schemes. In C. Plantin (Ed.), *Argumentation Through Languages and Cultures* (pp. 101–139). Springer Nature Switzerland. https://doi.org/10.1007/978-3-031-19321-7_6

Walker, J. P., & Sampson, V. (2013). Learning to argue and arguing to learn: Argument-driven inquiry as a way to help undergraduate chemistry students learn how to construct arguments and engage in argumentation during a laboratory course. *Journal of Research in Science Teaching, 50*(5), 561–596. https://doi.org/10.1002/tea.21082

Wambsganss, T., Janson, A., & Leimeister, J. M. (2022). Enhancing argumentative writing with automated feedback and social comparison nudging. *Computers & Education, 191*,. https://doi.org/10.1016/j.compedu.2022.104644

Wambsganss, T., Niklaus, C., Cetto, M., Söllner, M., Handschuh, S., & Leimeister, J. M. (2020). AL: An adaptive learning support system for argumentation skills. *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, 1–14. https://doi.org/10.1145/3313831.3376732

Wang, C., Liu, X., Wang, L., Sun, Y., & Zhang, H. (2021). Automated scoring of Chinese grades 7–9 students' competence in interpreting and arguing from evidence. *Journal of Science Education and Technology, 30*(2), 269–282. https://doi.org/10.1007/s10956-020-09859-z

Wiley, J., Hastings, P., Blaum, D., Jaeger, A. J., Hughes, S., Wallace, P., Griffin, T. D., & Britt, M. A. (2017). Different approaches to assessing the quality of explanations following a multiple-document inquiry activity in science. *International Journal of Artificial Intelligence in Education, 27*(4), 758–790. https://doi.org/10.1007/s40593-017-0138-z

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

Wilson, C. D., Haudek, K. C., Osborne, J. F., Buck Bracey, Z. E., Cheuk, T., Donovan, B. M., Stuhlsatz, M. A. M., Santiago, M. M., & Zhai, X. (2023). Using automated analysis to assess middle school students' competence with scientific argumentation. *Journal of Research in Science Teaching*, *n/a* (n/a). https://doi.org/10.1002/tea.21864

Wilson, M. (2009). Measuring progressions: Assessment structures underlying a learning progression. *Journal of Research in Science Teaching, 46*(6), 716–730. https://doi.org/10.1002/tea.20318

Zhai, X., Haudek, K. C., & Ma, W. (2023). Assessing argumentation using machine learning and cognitive diagnostic modeling. *Research in Science Education, 53*(2), 405–424. https://doi.org/10.1007/s11165-022-10062-w

Zhai, X., Haudek, K. C., Stuhlsatz, M. A. M., & Wilson, C. (2020a). Evaluation of construct-irrelevant variance yielded by machine and human scoring of a science teacher PCK constructed response assessment. *Studies in Educational Evaluation, 67*,. https://doi.org/10.1016/j.stueduc.2020.100916

Zhai, X., Haudek, K., Shi, L., Nehm, H., & Urban-Lurain, M. (2020b). From substitution to redefinition: A framework of machine learning-based science assessment. *Journal of Research in Science Teaching, 57*(9), 1430–1459. https://doi.org/10.1002/tea.21658

Zhai, X., Yin, Y., Pellegrino, J. W., Haudek, K. C., & Shi, L. (2020c). Applying machine learning in science assessment: A systematic review. *Studies in Science Education, 56*(1), 111–151. https://doi.org/10.1080/03057267.2020.1735757

Zhai, X., Krajcik, J., & Pellegrino, J. W. (2021a). 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. https://doi.org/10.1007/s10956-020-09879-9

Zhai, X., Shi, L., & Nehm, R. H. (2021b). 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. https://doi.org/10.1007/s10956-020-09875-z

Zhu, M., Liu, O. L., & Lee, H. S. (2020). The effect of automated feedback on revision behavior and learning gains in formative assessment of scientific argument writing. *Computers & Education, 143*,. https://doi.org/10.1016/j.compedu.2019.103668

Authors and Affiliations

Kevin C. Haudek[1] · Xiaoming Zhai[2,3]

✉ Kevin C. Haudek
haudekke@msu.edu

Xiaoming Zhai
xiaoming.zhai@uga.edu

[1] Department of Biochemistry and Molecular Biology and CREATE for STEM Institute, Michigan State University, 620 N. Shaw Lane, Rm 115, East Lansing, MI 48823, USA

[2] Department of Mathematics, Science, and Social Studies Education, University of Georgia, 125M Aderhold Hall, 110 Carlton St, Athens, GA 30602, USA

[3] AI4STEM Education Center, University of Georgia, 125M Aderhold Hall, 110 Carlton St, Athens, GA 30602, USA