Exploring the Effect of Assessment Construct Complexity on Machine Learning Scoring of Argumentation

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

Paper presented at the 2021 Virtual Annual International Conference of the National Association for Research in Science Teaching (NARST)
April 7-10, 2021

[1] Department of Biochemistry and Molecular Biology and CREATE for STEM Institute, Michigan State University
[2] Department of Mathematics, Science, and Social Science Education, University of Georgia

Author Note

Kevin C. Haudek https://orcid.org/0000-0003-1422-6038
Xiaoming Zhai http://orcid.org/0000-0003-4519-1931

This material is based upon work supported by the National Science Foundation under Grants No. 1323162 and 1561159. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation. We acknowledge the many contributions to this collaborative project by Zoe Buck-Bracey, Tina Cheuk, Brian Donovan, April Gardner, Jonathan Osborne, Molly Stuhlsatz, Marisol Santiago and Chris Wilson.

The authors have no conflicts of interest to declare.

Correspondence concerning this article should be addressed to Kevin C. Haudek, Michigan State University, 603 Wilson Rd., Rm. 219, East Lansing, MI, 48824 or email to haudekke@msu.edu

Abstract

Argumentation, a key scientific practice, requires students to construct and critique arguments, but timely and large-scale evaluation of responses depends on automated text scoring systems, which rely on machine learning algorithms. Recent work has shown the utility of these automated systems, as well as proposing to increase the use of machine learning for high complexity assessments. Therefore, in this study, we investigated whether the construct complexity of an assessment item affected machine learning model performance. We employed human experts to score student responses to 17 argumentation items aligned to 3 levels of a learning progression and randomly selected 361 responses to use as training sets to build machine learning scoring models for each item. We were able to produce scoring models with a range of scoring agreement between computers and humans, measured by Cohen's kappa (*M* = .60; range .38 - .89). Most models demonstrated good to almost perfect performance (kappa > .60). We found that scoring models for more complex constructs, such as multiple dimensions of science learning or higher levels of a learning progression, had lower performance metrics as compared to models for items at lower levels. These negative correlations were significant for three construct characteristics we examined, complexity, diversity and structure. In order to develop automated scoring models for more complex assessment items, larger training sets or additional model tuning may be required.

1. Introduction

Assessing students' argumentation, one that is among the most critical scientific practices proposed in the K-12 Framework (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). Assessing such a complex construct has gone beyond the capacity of the traditional multiple-choice items, and thus many suggest using performance-based constructed response measures. To make the proposed measures be accessible to teachers and students, we developed machine learning (ML) algorithms to automatically score students' constructed responses of argumentation. This approach has great potential to engage teachers in using constructed responses of argumentation assessments in the classroom.

However, we found it challenging to develop solid ML models that can generate accurate scores that are highly consistent with those assigned by human experts. Among those reasons that challenge the development of ML models, the underlying assessment construct of the assessment, as well as the training sources, may be particularly critical (Zhai, Haudek, Shi, et al., 2020). A meta-analysis recently reported that assessment internal features, such as the complexity of constructs which the assessments tap, are critical factors that may moderate machine scoring accuracy (Zhai, Shi, & Nehm, 2020). However, there is limited empirical data to test this assumption specifically for argumentation assessments, and we have limited knowledge about the fluctuation of the human-machine score agreements due to the construct within a given performance assessment. Therefore, in this study we developed and investigated 17 assessment items varying in construct characteristics. We employed the same strategy to train the computer to develop 17 ML models, one for each assessment item, holding the number of randomly sampled responses constant. We examined the resulting model performance to address the research question: *How are the item construct complexity, structure, and diversity associated with ML model performance?*

2. Assessment Construct

Cronbach and Meehl (1955) suggest that 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 performance in tests. 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. What makes assessment practice complicated is that the assessment construct is usually complicated, diverse, and contains developmental features (Zhai, Haudek, Stuhlsatz, et al., 2020). In their study, Zhai, Haudek, Shi, et al. (2020) abstracted three fundamental features of ML-based science assessments: *complexity, diversity*, and *structure*. In this study, we adopted the three features as our analytical framework.

According to Bloom's taxonomy, assessment tasks may demand varying complex cognitive abilities (Forehand, 2010). In the past few decades, scholars in science education have been focused on a range of cognitive abilities, from students' conceptual understanding of scientific ideas to reasoning ability in *Complex* tasks. The K-12 Framework (National Research

Council, 2012) further articulates that meaningful science learning integrates scientific practices, crosscutting concepts, and disciplinary core ideas. In these activities, students are required to move from simple memorizing knowledge, to analyzing, evaluating, and creating abilities to complete different tasks.

Diversity reflects the combinations of different cognitive demands in performing a task. In dealing with three-dimensional science learning, the assessment is multifaceted. The cognitive demands during solving such a science problem may include multiple components (e.g., practices, disciplinary core ideas). The number and combinations of components of the cognitive demands that the assessment task requires to perform on the task feature the complexity of the construct. The more components required, the more diverse of the assessment construct is. Students are required to perform on science tasks by showing their ability to conduct practices and understanding of scientific knowledge. Compared to assessing such three-dimensional science learning, some assessments may only focus on one dimension of the construct, which is less diverse.

Zhai, Haudek, Shi, et al. (2020) further argue that the assessment constructs used with ML reflects cognitive developmental features and denoted this feature as *Structure*. That is, students' competence of and proficiency with science ideas and practices progresses 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).

3. Scientific Argumentation

Scientific argumentation is the practice of reasoning within a domain by constructing and critiquing links between scientific claims and evidence (Osborne et al., 2004). Argumentation is also foundational to scientific inquiry, as it requires the evaluation of evidence and claims (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). Over the last decades, many studies have examined how to implement scientific argumentation in the classroom to improve student practice (Cavagnetto, 2010; Driver et al., 2000). Concurrent with these efforts are a number of studies which investigate assessing students' written scientific arguments, since writing is one of the prominent forms of engaging in argument (Lee et al., 2014; McNeill, 2009).

Many of the current analytic frameworks for scientific argumentation rely on Toulmin's (1958) foundational perspective on argumentation (Sampson & Clark, 2008). Toulmin (1958) allowed for domain-specific elements within an argument, while recognizing some elements of the argument are universal across disciplines. Following this framework, different statements within an argument hold different functions. Claims are statements that assert a perspective, while data are used to support a claim and warrants justify the use of data for a claim. Further work has expounded that argument activities include both construction of one's own argument as well as considering arguments made by others (Berland & Reiser, 2011; Osborne, 2010). Osborne and colleagues (2016) proposed a learning progression for how middle school students develop in scientific argumentation practice. This progression identifies three different levels based largely on coordination of argument elements and includes both construction and critique of arguments as activities of scientific argumentation.

4. Methods

4.1. Study context and participants

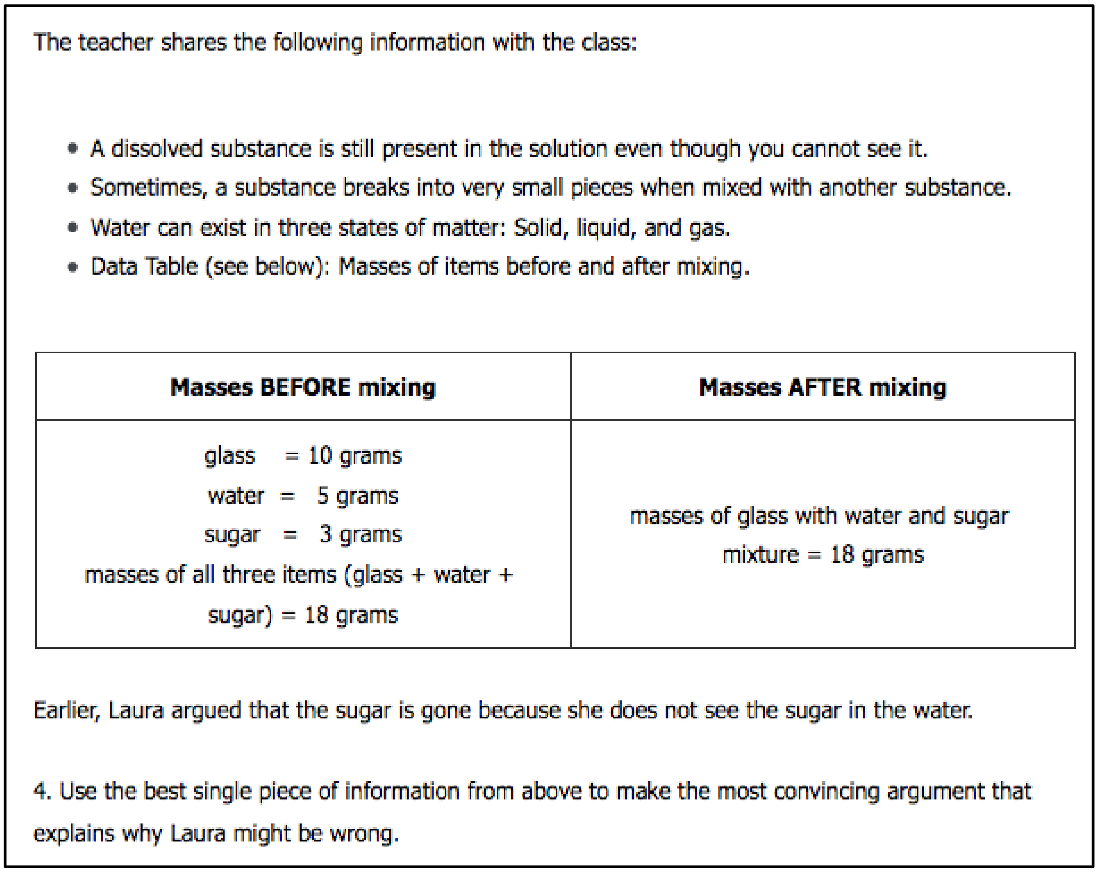
Assessing scientific argumentation in STEM education is critical in that it is one of the identified scientific practices proposed in the K-12 Framework (NRC, 2012) and the NGSS (NGSS Lead States, 2013). Assessing such a complex construct has gone beyond the capacity of traditional multiple-choice items, and the use of performance-based constructed response items may help better assess how students generate arguments (Sampson & Clark, 2008). However, evaluating large numbers of student-written responses can be challenging for an individual teacher. As such, we were interested in developing ML scoring models for middle school argumentation responses, aligned to a learning progression for the development of argumentation skills at those levels. We collected responses from 931students from science classrooms in grades 5-8 from two school districts in California. Student responses were collected electronically via Qualtrics. Sets of items were sequenced differently in assessments given to different school districts, thus leading to different numbers of responses collected for each item.

4.2. Assessment tasks

We developed a total of nineteen constructed response assessment items aligned to a learning progression for argumentation for middle school students (see Wilson et al., under review). These items targeted different levels of the learning progression and engaged students in constructing and/or critiquing arguments (see example items; Figures 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 which was later dropped from analysis (see section 4.4. below). The items were divided into 3 item sets based on science contexts: sugar dissolving in water (S), the kinetic motion of gases (G) and bacterial growth (B). Coding rubrics for the student responses were developed with each item and each rubric had a different number of possible codes intended to capture the different quality of student performance in argumentation within a given item.

Here we present a sample assessment task item from the sugar dissolving in water context (Figure 1). Preceding items in the sugar set, set up a scenario using arguments from competing 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 one of the fictitious character's argument. This item is aligned to level 2a in a learning progression (Osborne et al., 2016), since it requires students to provide a counter-critique to another person's argument. The response coding rubric for this item had four possible levels to rank students' responses based on if they were able to provide a valid critique based on evidence.

Figure 1. *Sugar 4 item*



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.

4.3. Coding of the construct characteristics of assessment tasks

Coding scheme. We developed a coding scheme for characteristics of the nineteen assessment items. We borrowed from three existing frameworks (Osborne et al., 2016; NGSS Lead States, 2013; Zhai, Haudek, Shi, et al., 2020) to capture different characteristics relevant to machine learning, science learning and scientific argumentation. Using these frameworks, we examined three components (*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 loosely approximates Bloom's taxonomy categorization (Bloom et al., 1956). Items ranged from low-level tasks like identifying provided information in the item to high-level tasks like evaluating multiple pieces of information. 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. We reasoned that items that engaged multiple dimensions of science learning (e.g. cross cutting concept and argumentation) would be more difficult. Since all items were designed to assess argumentation, all items 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 in 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.

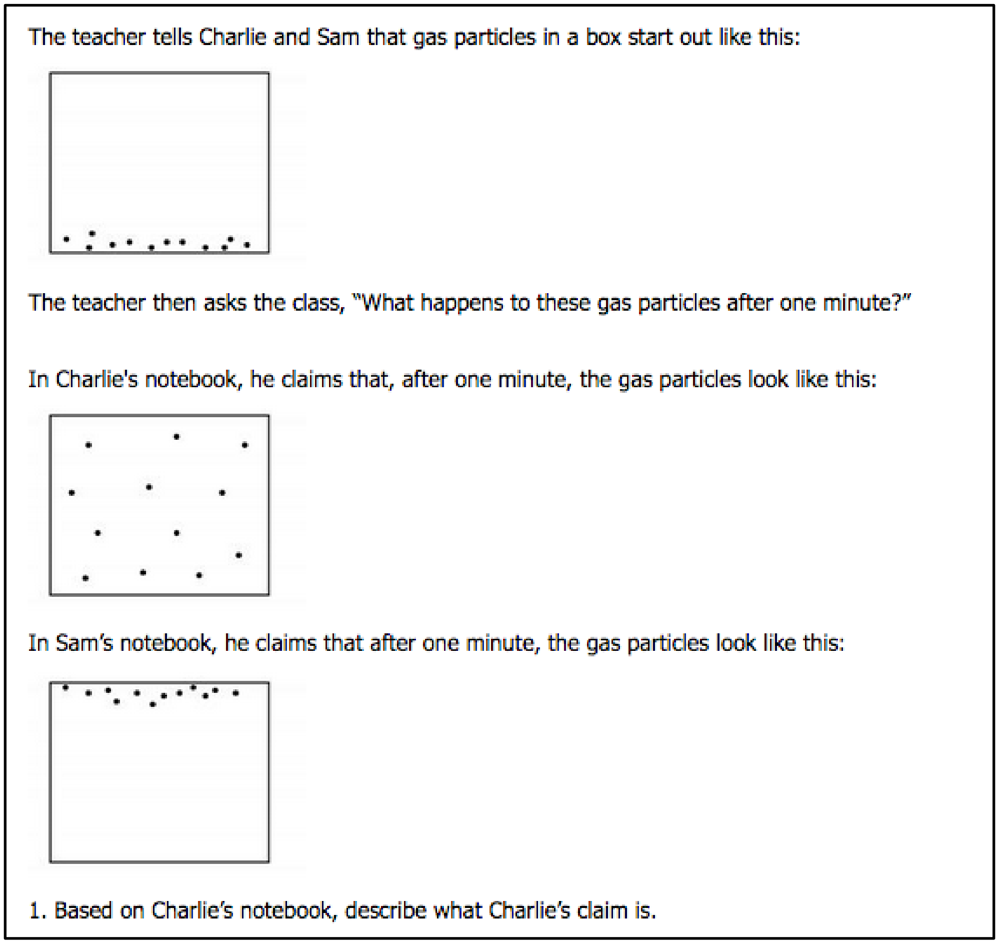
Table 1. *Developed coding scheme for three item characteristics relevant to machine-learning scored assessments.*

Characteristic	Level 1	Level 2	Level 3	Level 4
Complexity	Memorization Item provides which knowledge to use	Apply Item requires students to possess the knowledge to use	Evaluate Students use data or information to reach a conclusion	Analyze Students use or integrate multiple data or information to reach conclusion
Diversity	Only engages 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 counter-claim	N/A

Example items and codes. Here we present two examples of item coding, to illustrate high- and low- level item characteristics. First, we present the coding characteristics for item S4 (see Figure 1), as an item with 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 incorporate the evidence into a critique, which 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 an item in the context of gas diffusion (G1; Figure 2). This is the first item in this context subset, and students are introduced to two fictitious characters and their differing ideas on how gases move in a given space. These characters and their ideas will be used repeatedly through the context subset, but this first item displays relatively low complexity of constructs. We coded this item for *Structure* as a level 1, since it asks students to identify another person's (character's) claim. We coded the item as a level 1 in *Complexity*, since it requires students to re-iterate information that is presented to them as part of the set-up of the item. For this item, students must only describe Charlie's model; the item does not require students to analyze or do anything else with 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-quality response. The item is aligned with the practice of Engaging in Argument but does not require explicit use of disciplinary ideas or cross cutting 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. Thus, other science dimensions are not present for this specific item.

Figure 2. *Gases 1 item*



The teacher tells Charlie and Sam that gas particles in a box start out like this:

The teacher then asks the class, "What happens to these gas particles after one minute?"

In Charlie's notebook, he claims that, after one minute, the gas particles look like this:

In Sam's notebook, he claims that after one minute, the gas particles look like this:

1. Based on Charlie's notebook, describe what Charlie's claim is.

Coding of assessment tasks. After developing the assessment task coding scheme and defining generic examples for each characteristic at each level, two coders independently coded the nineteen items on all three components. We compared codes on all characteristics and coders met to discuss any codes in disagreement. During these consensus discussions, coders reviewed both the assessment items and response coding rubrics to see which elements were necessary for a complete and full answer. We used the consensus codes of the items as the final codes for item complexity.

4.4. Machine learning algorithmic model development.

Human coding of responses. For each set of items, two coders were trained on the associated coding rubrics for each item. The two 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; k) between the coders was > 0.6 on each rubric component or until three rounds of training were completed. Cohen's kappa for the last training round is indicated in Table 1. Disagreements on scores in the training round were resolved by consensus discussion between coders. For some items in the gases contexts, we only have after discussion consensus scores for the training round and cannot provide Cohen's kappa (N/A in Table 2). The remaining data set of student responses was split into two subsets and each coder scored one subset independently.

We calculated a diversity index, evenness, for the human assigned scores to responses to each item (Pielou, 1966). This measure represents the distribution of scores at the various levels of coding rubrics, where evenness for an item can vary between 1 (indicating responses are equally distributed, or balanced, across all levels of the rubric) to 0 (responses are only in one level of the rubric). In this study, the mean evenness is .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 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 machine learning models. We used a supervised ML text classification approach to assign student written responses a score (Aggarwal & Zhai, 2012). During our ML process, each student response is treated as a document and the coding rubric is treated as a multi-level class. The computerized scoring system then generates predictions on whether each given document is a member of each class. We use text processing based on natural language processing to extract text features from responses which are then used as inputs for an ensemble of eight individual algorithms, common in ML classification applications, to generate predicted scores (Jurka et al., 2013). The computer model is generated and validated using a 10-fold cross-validation approach.

During model development, we decided to drop two items in the bacteria context, 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 explanations, we did not generate ML models for these two items and dropped them from subsequent analysis. We generated a 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, using unigrams and bigrams. In usual practice for developing scoring models, one would optimize model performance by tweaking tuning parameters or text processing strategies (Madnani et al., 2017). However, this complicates comparing model performance across items, as the model performance is 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 hoped to investigate the role of the assessment item characteristics in model performance.

Table 2. *The characteristics of the item tasks and the interrater reliability of human coders on student responses.*

Item	Complexity	Diversity	Structure	# of levels in 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	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	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	N/A	0.771
G8	4	2	3	4	548	N/A	0.635

The computer models produce a predicted classification for each response that can be compared to the human assigned holistic score. For each computer model, we calculated accuracy, Cohen's kappa (k), a measure of interrater reliability, and Spearman rank-order correlations (r_s, N=361) between the computer predicted score and human assigned score. We examined the accuracy of the ML models built using equal numbers of responses in the training set for all models and used a consistent set of text processing procedures. We used benchmarks of moderate (k>.4), substantial (k > .6) to near perfect (k >.8) agreement between human and computer assigned scores as defined by Landis & Koch (1977) for overall evaluation of model performance.

5. Results

Overall, we had a good range of item characteristics over all levels of the item components (see Table 2). *Structure* showed the most balanced distribution of items across levels, 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 most common. Finally, we found that most items displayed high (evaluate) or low (memorization) level *Complexity*, with fewer items in the intermediate levels.

5.1. Machine learning model performance across items

First, we developed a total of 17 ML models using a consistent set of model parameters and an equal number of student responses. The models showed a range of performance metrics (Table 3). The mean Cohen's kappa was substantial (*M*= .60, *SD*= .15) and the mean accuracy was fairly high (*M*= .79, *SD*= .11). We developed five ML models for items in the sugar context with an average Cohen's kappa of .65 and a range of .55 to .80. For the four bacteria items, we developed a ML model for each with an average k= .69 and a range of .46 to .89. The eight gas items proved the most challenging to develop scoring models for, as model performance ranged from moderate to substantial (Landis & Koch, 1977). The eight gas items show decreased model performance, with a moderate average k= .52 and a range of .38 to .65. We 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> 0.7, which has been suggested as a threshold for quality model performance (Williamson et al., 2012). As expected, this one model (B3) also had the lowest accuracy and 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 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 gases context produced a rho greater than 0.7, although five of these models had rho > 0.6. Despite these challenges, correlations between human and model scores were significant at the $p<0.01$ level. From these findings, along with the model Cohen's kappa results, we conclude that we have well-performing models for items in the sugar and bacteria context, 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 agreement between coders (i.e. human & computer) for chance agreement (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.

Table 3. *Machine-Human agreement measures on argumentation tasks*

Item	k	Accuracy	r_s^*
S1	0.62	0.77	0.788
S2	0.75	0.87	0.740
S3	0.80	0.90	0.811
S4	0.55	0.73	0.715
S5	0.55	0.69	0.708
B1	0.89	0.94	0.947
B2	0.80	0.89	0.841
B3	0.46	0.65	0.684
B6	0.61	0.91	0.724
G1	0.62	0.91	0.619
G2	0.65	0.89	0.653
G3	0.46	0.65	0.650
G4	0.51	0.69	0.640
G5	0.62	0.83	0.640
G6	0.40	0.65	0.556
G7	0.50	0.73	0.482
G8	0.38	0.75	0.488

Note. k = Cohen's kappa; r_s = Spearman's rho.
*All r_s values p<.01

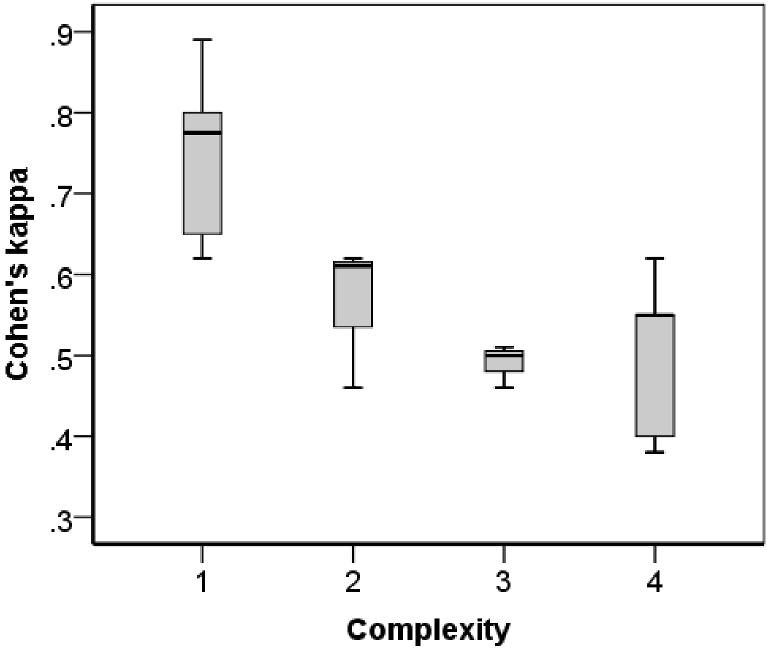
Further, we looked at the relationship between evenness, or the distribution of responses in the levels of an item rubric and model performance. Surprisingly, there was no significant correlation between the balance of response across levels and Cohen's kappa of the resulting model (Pearson's *r*= .08, p= .765, N= 17).

5.2 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 not 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 for each characteristic alone.

First, we examined the model performance using Cohen's kappa by levels of *Complexity*, or item task (Figure 3). We found that models for items at level 1 of *Complexity* had better performance than for models for items at all other levels. We see that all models at level 1 of *Complexity* achieved substantial agreement as measured by Cohen's kappa (k> .6). 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 for level 2 items was slightly higher than model performance for levels 3 and 4. For level 2 items, we see that most models had a performance in the range of .5 to .6, which is in the 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 level of item *Complexity*. There was a significant negative correlation between the two variables (r_s= -.752, p< .001, N=17).

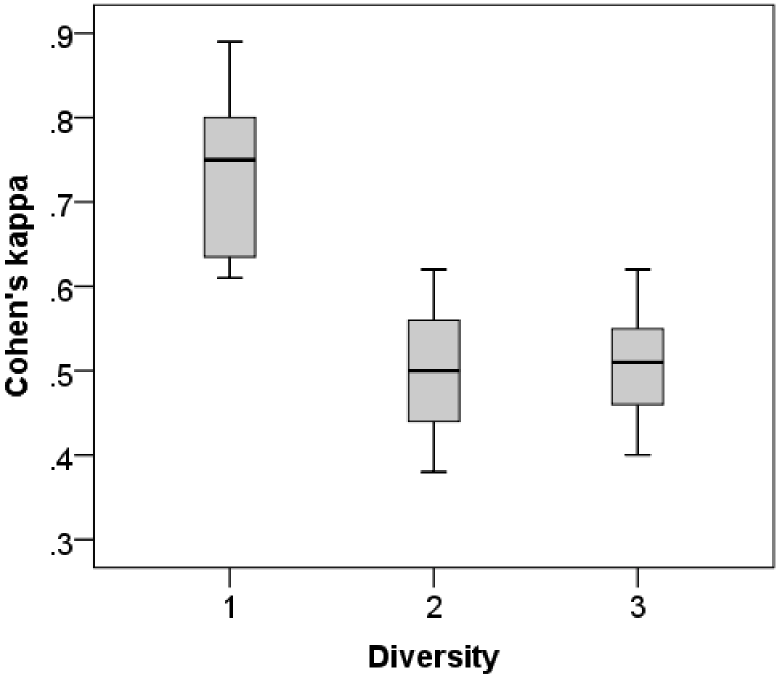
Figure 3. *A box plot of model performance for items exhibiting different levels of Complexity.* The shaded box represents middle 50% values and whiskers extend to maximum and minimum values. The thick black line represents the median value.



5.3 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. We found that models for items at level 1 of *Diversity* had better performance than for models for items that engaged multiple dimensions of science learning (levels 2 and 3) We see that all models at level 1 of this characteristic were above substantial agreement between codes (k> .6). 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 level 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 between moderate and substantial agreements. Spearman's rho correlation coefficient was used to examine the relationship between Cohen's kappa and level of item *Diversity*. There was a significant negative correlation between these two variables (r_s= -.718, p< .01, N=17).

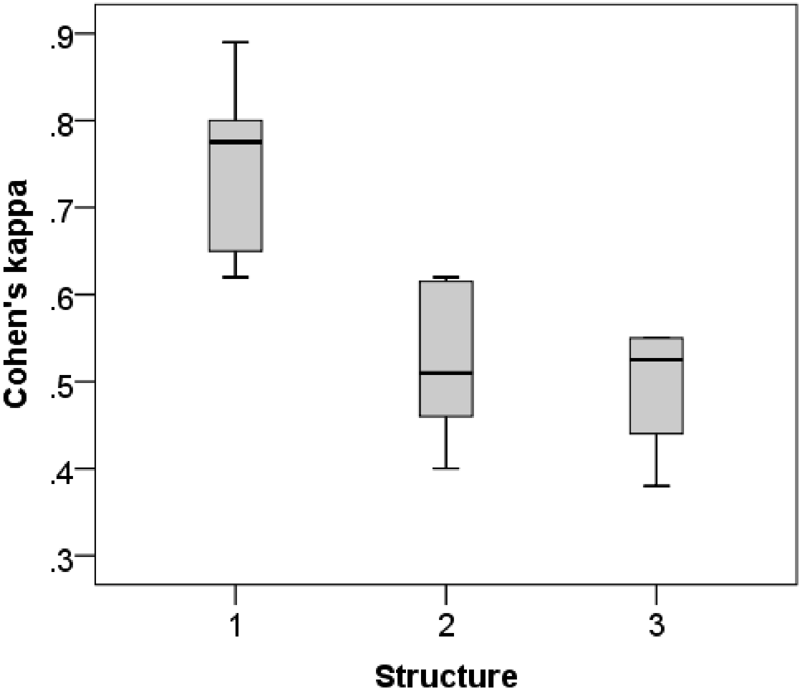
Figure 4. *A box plot of model performance for items exhibiting different levels of Diversity.*



5.4 Association of item structure on model performance

We examined how the structure of the item, aligned to cognitive levels of a learning progression, influenced model performance (Figure 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 S*tructure* exceeded the substantial Cohen's kappa benchmark. We found that the measures of the average of 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 level of item *Structure*. There was a significant negative correlation between the two variables (r_s= -.764, p< .001, N=17).

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



6. Discussion

Though ML has great potential to be widely applied in science assessments, the accuracy of machine scoring remains a "black box" and draws great concerns. A recent, critical review suggested that examining factors that impact machine scoring and developing generalizable algorithms is critical to further increase the usability of ML in science assessments (Zhai, Yin, Pellegrino, et al., 2020). With this in mind, the current study examined a critical internal feature of assessments (Zhai, Shi, & Nehm, 2020), characteristics of the target construct, and how these features are associated with machine scoring accuracy. Aligned to an ML-based assessment framework (Zhai, Haudek, Shi, et al., 2020), we identified three critical characteristics of construct: complexity, diversity, and structure. Our empirical findings suggest that construct characteristics do associate with ML model performance. We found negative and significant correlations between item characteristics and ML model performance for all three item characteristics 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 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 C*omplexity* showed decreasing performance in a fairly negative linear relationship, with level 4 showing a large range of different model performance. 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 more similar text in the responses when students use information from the item. In their previous study, Zhai, Haudek, Shi et al. (2020) found the majority of constructed response items scored with ML models were of relatively low *Complexity*. In our study, we had more higher-level items, which exhibited lower model performance. This represents a continued challenge for applying ML in science assessment: the need to develop advanced machine algorithmic models to better suited to accurately score higher complexity assessment items.

We also studied the number of science learning dimensions, or *Diversity*, engaged by an 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 any multidimensional item, not necessarily a specific dimension or number of dimensions. 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 single holistic score may represent student facility with different dimensions (Gane et al., 2018; Zhai 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).

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 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 performance 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 be structured appropriately. This is 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. 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 dimensions of science learning and may not be generalizable to other science practices.

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 coder agreement on the training set. This also raises the possibility that some of the item complexity may impact human scoring 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.

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 association between the balance of response across levels and model performance. One possible reason for this result is that all of our items displayed a distribution of responses across coding levels that were above some frequency 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 code levels that contained < 6% of the responses and therefore provide 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., under review). However, we did check to see if item difficulty, based on the human assigned score to each response, was correlated with model accuracy. Although we found a positive association, it was not significant, suggesting that item characteristics influence ML model performance and not just student performance.

6.1 Implications

Our study has provided evidence that science item construct characteristics associate with ML-model performance. It is critical to identify and understand these effects in order to identify the possibilities and limitations of ML-based scoring of science assessments. As science teaching and learning moves to align with multidimensional learning advocated by the NGSS, accurately classifying multiple dimensions using ML models will be necessary (Maestrales et al., 2021). This also highlights a challenge to advancing ML-based assessments in science by aligning these assessments with models of cognitive development (Zhai, Haudek, Shi, et al., 2020). Extrapolating from our findings, producing ML models for increasingy sophisticated cognitive abilities will take additional model tuning, more iterative development cycles, novel technical features of text processing and/or larger training sets of labeled responses. If we want to develop and deploy ML-based assessments in science at scale, then it is critical to move away from designing items and models on a "one-off" basis aligned to scattered constructs, but toward integrating design theories with assessment practices and incorporate all phases of assessment into a validity process (Gane et al., 2018; Zhai et al., 2021). Further, for assessment developers to learn what works across contexts, we must not only focus on outcomes of the models (e.g. accuracy) but technical features of the model as well as item characteristics (Zhai, Shi, & Nehm, 2020; Zhai, Yin, Pellegrino, et al., 2020). 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.

7. References Cited

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

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

Bloom, B. S., Engelhart, M. D., Furst, E. J., Hill, W. H., & Krathwohl, D. R. (1956). Taxonomy of educational objectives: The classification of educational goals. In *Handbook I: Cognitive domain*. David McKay Company.

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

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. In *Science Education* (Vol. 84, Issue 3, pp. 287–312).

Forehand, M. (2010). Bloom's Taxonomy. In *Emerging Perspectives on Learning, Teaching and Technology* (1st ed., Vol. 1).

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

Jurka, T. P., Collingwood, L., Boydstun, A. E., Grossman, E., & Van Atteveldt, W. (2013). RTextTools: A Supervised Learning Package for Text Classification. In *The R Journal* (Vol. 5, Issue 1, pp. 6–12).

Landis, J. R., & Koch, G. G. (1977). The measurement of observer agreement for categorical data. In *Biometrics* (Vol. 33, pp. 159–174).

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., 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

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

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

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

NGSS Lead States.(2013). *Next Generation Science Standards: For States, By States.* [Report]. The National Academies Press.

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. (2010). Arguing to Learn in Science: The Role of Collaborative, Critical Discourse. In *Science* (Vol. 328, pp. 463–466).

Osborne, J. F., Henderson, B., MacPherson, A., Szu, E., Wild, A., & Shi-Ying, Y. (2016). The Development and Validation of a Learning Progression for Argumentation in Science. In *Journal of Research in Science Teaching* (Vol. 53, Issue 6, pp. 821–846).

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

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

Toulmin, S. (1958). *The Uses of Argument*. Cambridge University Press.

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

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

Williamson, D. M., Xi, X., & Breyer, F. J. (2012). A Framework for Evaluation and Use of Automated Scoring. In *Educational Measurement: Issues and Practice* (Vol. 31, Issue 1, pp. 2–13). https://doi.org/doi:10.1111/j.1745-3992.2011.00223.x

Wilson, C., Haudek, K. C., Osborne, J., Stuhlsatz, M., Cheuk, T., Donovan, B., Buck-Bracey, Z., Santiago, M., & Zhai, X. (under review). Using Automated Analysis to Assess Middle School Students’ Competence with Scientific Argumentation. *Journal of Research in Science Teaching*.

Zhai, X., Haudek, K. C., Stuhlsatz, M. A. M., & Wilson, C. (2020). Evaluation of construct-irrelevant variance yielded by machine and human scoring of a science teacher PCK constructed response assessment. In *Studies in Educational Evaluation* (Vol. 67, p. 100916). https://doi.org/10.1016/j.stueduc.2020.100916

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

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

Zhai, X., Shi, L., & Nehm, R. H. (2020). A Meta-Analysis of Machine Learning-Based Science Assessments: Factors Impacting Machine-Human Score Agreements. *Journal of Science Education and Technology*. https://doi.org/10.1007/s10956-020-09875-z

Zhai, X., Yin, Y., Pellegrino, J. W., Haudek, K. C., & Shi, L. (2020). Applying Machine Learning in Science Assessment: A Systematic Review. In *Studies in Science Education* (Vol. 56, Issue 1, pp. 111–151). https://doi.org/10.1080/03057267.2020.1735757