Journal of Science Education and Technology (2021) 30:239–254
https://doi.org/10.1007/s10956-020-09895-9

Using Machine Learning to Score Multi-Dimensional Assessments of Chemistry and Physics

Sarah Maestrales[1] · Xiaoming Zhai[2] · Israel Touitou[3] · Quinton Baker[1] · Barbara Schneider[1] · Joseph Krajcik[3]

Accepted: 21 December 2020 / Published online: 26 March 2021
© This is a U.S. government work and not under copyright protection in the U.S.; foreign copyright protection may apply 2021

Abstract

In response to the call for promoting three-dimensional science learning (NRC, 2012), researchers argue for developing assessment items that go beyond rote memorization tasks to ones that require deeper understanding and the use of reasoning that can improve science literacy. Such assessment items are usually performance-based constructed responses and need technology involvement to ease the burden of scoring placed on teachers. This study responds to this call by examining the use and accuracy of a machine learning text analysis protocol as an alternative to human scoring of constructed response items. The items we employed represent multiple dimensions of science learning as articulated in the 2012 NRC report. Using a sample of over 26,000 constructed responses taken by 6700 students in chemistry and physics, we trained human raters and compiled a robust training set to develop machine algorithmic models and cross-validate the machine scores. Results show that human raters yielded good (Cohen's $k = .40–.75$) to excellent (Cohen's $k > .75$) interrater reliability on the assessment items with varied numbers of dimensions. A comparison reveals that the machine scoring algorithms achieved comparable scoring accuracy to human raters on these same items. Results also show that responses with formal vocabulary (e.g., velocity) were likely to yield lower machine-human agreements, which may be associated with the fact that fewer students employed formal phrases compared with the informal alternatives.

Keywords Three-dimensional science learning · Machine learning · Automatic scoring

Measuring science knowledge and achievement has long been an important topic in science education research. The National Research Council ([NRC], 2012) has spelled out what they call three-dimensional learning to better facilitate student knowledge development and meet the demands of a modern STEM workforce. Three-dimensional learning encourages knowledge-in-use that can be generalized and used across multiple scientific fields to successfully meet the rapidly changing demands of science and technology careers adapted to the emerging issues of the twenty-first century (Harris et al., 2019; Haudek et al., 2019). This concept of knowledge-in-use occurs when students apply disciplinary core ideas (DCIs) in tandem with science and engineering practices (SEPs) and crosscutting concepts (CCs) to solve problems or make sense of phenomena. While the Framework for K-12 Science Education (NRC, 2012) presents a promising new vision of science learning, assessing the three-dimensionality of science learning is challenging (see NRC assessment report [2014] and the National Academies of Sciences, Engineering, and Medicine report [2019]).

Multiple-choice questions are used ubiquitously in national, state, and classroom assessments of science achievement. However, these multiple-choice assessments typically rely on memorization of key concepts and thus have difficulty meeting the needs for assessing knowledge-in-use learning. Instead of strictly using multiple-choice questions, assessments should incorporate items that use the three dimensions of learning through a variety of task formats including constructed response (CR), which requires students to use their knowledge to solve problems with scientific practices (e.g., Harris et al., 2019).

✉ Sarah Maestrales
maestral@msu.edu

[1] College of Education, Michigan State University, Lansing, Michigan, USA

[2] Department of Mathematics and Science Education, University of Georgia, Athens, Georgia, USA

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

Unfortunately, CR is both time and resource consuming to score compared with multiple-choice items, and thus, teachers may not be willing to implement CR items in their classrooms. Approaches that employ machine learning have shown great potential in automatically scoring CR assessments (Zhai et al., 2020a). As indicated in a recent review study (Zhai et al., 2020c), machine learning has been adopted in many science assessment practices using CRs, essays, educational games, and interdisciplinary assessments (e.g., Lee et al., 2019a; Nehm et al., 2012). More importantly, machine scoring can provide automatic and immediate feedback to students and teachers, with the potential to accelerate the use of three-dimensional assessment practices in classrooms, benefiting science learning (Lee et al., 2019b).

While the potential of machine learning has been recognized, few studies have tackled the true challenge of scoring CR items on multi-dimensional science assessments. There are relatively few studies applying machine learning to analyze assessment items in which students perform tasks that require the use of multiple dimensions of scientific knowledge to make sense of phenomena (Zhai et al., 2020a). In addition, none of the studies explicitly document whether and how these assessments measure the dimensionalities of science learning.

We examine the capacity of machine learning to automatically score multi-dimensional science assessments, contrasted with human scorers, using a large database of student CRs. We highlight how the machine agreement changes as we increase the size of the training set as compared with human rater agreement. Additionally, we discuss some of the complex challenges for achieving high agreement between human and machine algorithms when scoring multi-dimensional assessments, including clarifying rubrics for improved agreement between human and machine scoring and treatment of missing and outlier responses and scores. This study answers three questions: (1) How reliable were human raters in scoring multi-dimensional responses? (2) Could machine learning algorithms score multi-dimensional assessments as accurately as humans? and (3) How are key phrases in student responses associated with machine scoring of the multi-dimensional assessments?

Defining the Dimensions of Science Learning

Building on the NRC reports, science researchers are calling for more comprehensive assessments that gauge students' abilities to use knowledge to explain phenomena, solve real-world problems, engage in creative and critical thinking, and analyze and interpret scientific data (Haudek et al., 2019). According to Pellegrino (2013), assessments that include CR should reflect the principles spelled out in *A Framework for K-12 Science Education* (NRC, 2012) and the Next Generation Science Standards (NGSS Lead States [NGSS], 2013), by carefully considering and identifying which of the three dimensions of learning each question is being designed to measure. In contrast to multiple choice, CR items are often difficult to develop and to score. Most studies that include them have not met the challenge of specifying and connecting the three dimensions of learning with their assessment items.

The NRC Framework for K-12 Science Education (2012) recommends that the learning and instruction of science throughout K-12 should integrate SEPs, DCIs, and CCs to make sense of phenomena. The NGSS puts this recommendation into practice by creating three-dimensional performance standards or expectations explaining the key concepts and skills students should be able to use at a given grade level (NGSS Lead States, 2013). Each dimension of learning has its own grade-based expectations. Measuring SEPs offers insight into how students use practices employed by scientists and engineers in the field, such as gathering and obtaining information and using it in argumentation (NRC, 2012). The CCs are considered "crosscutting" because they are concepts used to aid in examining phenomena and solving problems across all fields of science and engineering (NRC, 2012). In turn, assessments should measure how students apply disciplinary knowledge through SEPs and CCs.

Creating and Scoring Multi-dimensional Science Assessments

Despite the obvious advantages of CR in gaining more in-depth insights into student understanding, they are used somewhat less frequently than multiple choice because scoring CR requires considerable time and effort. However, it is important to create assessments that capture students' use of the three dimensions of science learning, and machine learning offers the potential to facilitate scoring, making these assessments more tenable for classroom and research purposes. Yet, three-dimensional items can be complicated to score as they must evaluate students' knowledge of the subject matter through the DCIs and provide insights about how students develop, understand, and use that knowledge. To score them, one must be familiar, not only with core content, but also with how the dimensions of learning work together and are being measured. These constraints prevent the broad use of three-dimensional tests that rely on CRs for science learning.

Because scoring CRs requires considerable time and effort on behalf of human raters, this can potentially elongate the period before students receive feedback (Ha & Nehm, 2016). However, studies (e.g., Lottridge et al., 2018) suggest that it should be possible to decrease human rater costs through automated machine learning algorithms. This would allow teachers and researchers to collect more detailed data on students' science knowledge with scoring costs comparable with those of multiple-choice assessments.

Scientists are building rubrics to measure aspects of three-dimensional learning related to human and machine scoring, but what seems to have received less attention is attributing the outcomes of the algorithm to the dimensions being scored. This is an important concept that needs to be examined in greater depth as it requires students to develop knowledge across subject domains. Outlining the dimensions associated with each item is one of the key contributions of our study. Further exploration of human and machine scoring for three-dimensional learning CRs needs to incorporate DCIs, SEPs, and CCs while recognizing the complexity and difficulty of this particular task.

Applications of Machine Learning in K-12 Science Assessment

The number of studies on the use of machine learning to score science assessments is increasing as the technology becomes more accessible and recent studies show considerable promise for machine scoring for science CRs across multiple age groups. A prior review (Zhai et al., 2020c) suggests that various researchers have used more than 20 programs or platforms to study the automatic scoring of science learning assessments. For example, researchers from ETS (Liu et al., 2014) applied machine learning for scoring student CRs that explained science phenomena through multi-dimensional reasoning with "c-rater," an automated machine algorithmic program. The Liu team from ETS developed multi-level rubrics and found that the machine was capable of automatically scoring student responses, achieving moderate to large Cohen's kappa (k) values between the machine and human scorers. They found that "c-rater" could capture valid ideas in students' responses and provide nuanced information about their performance.

Other scientists collaborated with ETS and explored the classroom applications of c-rater-ML in several projects. Mao et al. (2018) applied the c-rater-ML to automatically evaluate students' written argumentation to provide automatic feedback to students. The machine feedback could assist students to revise their arguments. In another study, Zhu et al. (2017) reported that more than 77% of students made revisions after receiving machine feedback and those who revised their responses received higher scores than the others in the final test.

The earliest development of programs besides c-rater is the SPSS Text Analysis (Nehm & Haertig, 2012), which required humans to develop word libraries manually. This program is costly and labor-intensive for users. The Summarization Integrated Development Environment (SIDE) developed by the TELEDIA lab at Carnegie Mellon University is the first free machine-learning-based confirmatory analysis program used in science education (Mayfield & Rosé, 2010, 2013). Its successor, Light Summarization Integrated Development Environment (LightSIDE), is more user-friendly, flexible, and accessible to public users and has been frequently used in studies. Other open sources such as RapidMiner or Weka are all popular in automatically scored science assessments. However, most of these programs applied individual algorithms each time. If one algorithm does not work well, users can choose another.

Instead of a single algorithm, this study employed the Automated Analysis of Constructed Response (AACR) Web Portal (AACR, 2020) to automatically score students' CRs using multiple algorithms. Different from most commercial automatic scoring programs, which fit one type of algorithm at a time, the AACR Web Portal developed eight algorithms that can be employed simultaneously. The AACR scoring Web Portal was developed to serve classroom needs for formative assessment purposes. Currently, it is used for exploratory and confirmatory factor analysis in item development. For exploratory analysis purposes, AACR can be applied with unsupervised machine learning to identify patterns and lexical features of student responses. Based on the findings, researchers can revise their items and rubrics iteratively. A confirmatory analysis is used in the late stage of item development to develop and validate the machine algorithms.

Based on the predictions of each algorithm, derived from the cross-validation, the machine assigns weights toward the algorithm that best optimizes the algorithmic parameters. The ensemble approach has been tested and compared with other general classifiers. Extensive experimentation by Large, Lines, and Bagnall (2019) reveals that the ensemble approach has measurable benefits over other classifiers such as alternative weighting, selection, or meta-classifier approaches. More importantly, the ensemble approach outperforms other classifiers with small training datasets. While machine learning is being used more frequently and continued research leads to improved reliability, a question remains about what steps researchers should take to move from human to machine scoring of multi-dimensional assessments.

Methods

Sample

To begin this process, for new and untested items, a large database of student responses is necessary. "Crafting Engaging Science Environments (CESE)," an ongoing science intervention with 6700 high school students in California and the Midwest, provided such a database. Within the CESE sample, 48.5% were male students and 51.5% were female students, 29.2% of students identified their race as white, 47.5 identified their race as Hispanic, 11.9% as black, and 5.0% as Asian. Almost three quarters of students (74%) reported speaking Spanish in the home. To measure baseline science understanding at the beginning of the intervention, CESE relied on an assessment developed using National Assessment of Education Progress (NAEP) open-source science test-bank items. All 6700 students took the test, and this yielded 26,800 constructed responses to four CR items. With 26,800 responses requiring classification, it would be possible to learn if machine scoring was a viable option. To conserve resources, the team decided to learn if AACR could score the CRs with the same reliability as human raters.

Instruments and Measures

CESE adopted four NAEP questions in this project. The questions tapped phenomena in chemistry or physics using everyday scenarios. Though these test questions were not initially designed to be three-dimensional as those illustrated in the *Framework* (NRC, 2012), multiple dimensions of science learning were detected in the responses to these items. As shown in Fig. 1, when reviewing the responses to each question, between 14 and 55% of students responded using multi-dimensional reasoning (i.e., the use of DCIs, SEPs, and CCs associated with the NGSS performance expectations) without being prompted. To accommodate this, three response classifications including incorrect, correct, and "multi-dimensional correct (MDC)" were adapted from the original binary rubrics. The MDC rating was awarded only if a student was able to demonstrate reasoning with regards to associated DCIs, CCs, or SEPs. Figure 1 shows the distribution of incorrect, correct, and MDC responses used by students on the assessment, as identified by human scorers.

The research team partnered with third party scientists to further verify the two- and three-dimensional rubrics for the CRs. The newly developed rubrics allowed the project team to use the existing items to probe students' learning of DCIs, CCs, and SEPs. Explanations of items and their rubrics are given below, while more complete details of the items, rubrics, sample responses, and associated dimensions of learning are shown in Appendices 1—4.

Item 1: Experimental design, shown in Appendix 1, asks students to identify the error in an experiment where a student tests three different shoes, each on a different floor, to determine which had the highest coefficient of friction. Students' responses included middle school level reasoning associated with NGSS performance expectations for DCI ETS1.A Defining and Delimiting Engineering Problems, and the grades 3–5 level SEP of Planning and Carrying Out Investigations (2013). To adapt this rubric to the NGSS, the MDC score for this item meant students correctly identified an error in the experimental setup (DCI) and explained that she could not compare the frictional force of different shoes on different floors due to the failure to isolate a variable while holding the floor constant (SEP). A correct response was given for students who correctly identified the error without explaining how the error affected the experiment.

Item 2: Relative motion focused on relative motion between two vehicles traveling on the highway (see Appendix 2). This item asked students to explain why a truck driving alongside them on the highway appeared to not be moving. Students engaged in middle school level reasoning associated with NGSS performance expectations for DCI [PS2.A] Forces and Motion, and CC Scale and Proportion (2013). To align the rubric with the NGSS performance expectations, the MDC classification was reserved for students who connected the truck's speed to that of the observer inside the vehicle (DCI) and stated the equal relative speeds would cause the phenomenon (CC). Responses that discussed only the speed without referencing how this related to the visual event were considered correct, but not MDC.

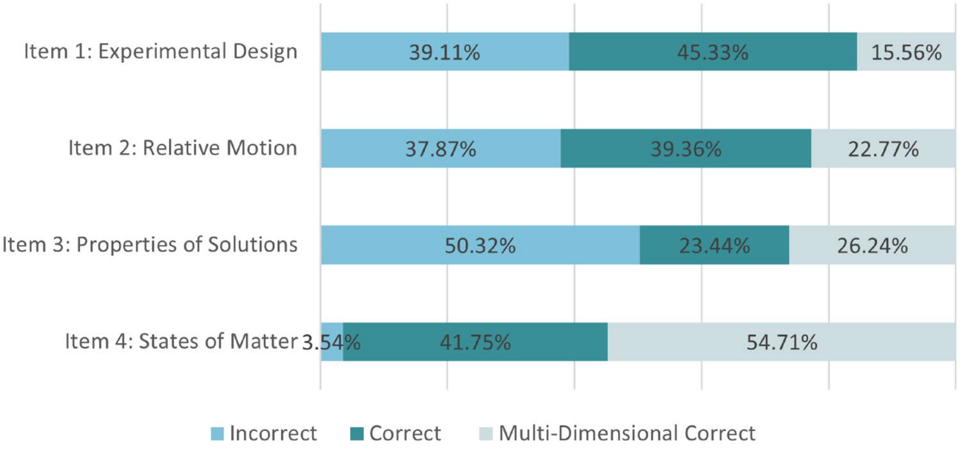


Fig. 1 Dimensionality of student responses

In the third constructed response, item 3: properties of solutions, students engaged elementary to middle school level reasoning for the DCI associated with PS1.A Structures and Property of Matter. As shown in Appendix 3, this was a fully three-dimensional item and students showed middle school level reasoning in cause and effect (CC) and planning and carrying out investigations (SEP) as outlined by the NGSS (2013). This question asked students to design an effective experiment to differentiate between the contents in two identical glasses. One glass contained saltwater while the other contained fresh water. Students could not suggest tasting the contents of either glass. To achieve the MDC classification, students would describe an experiment that controlled for relevant variables (SEP) to differentiate between fresh water and a solution (DCI) and correctly attribute causality (CC) to the chosen experiment by explaining the outcome. For a correct score, the DCI and SEP were considered inseparable. For example, a response that stated "test the density" was incorrect, because it explained neither an experiment that would do so, nor the expected results.

The fourth and final constructed response, item 4: states of matter, asked students to demonstrate their understanding of what causes matter to change states (see Appendix 4). The question asked students to explain why water in a hot pot would evaporate more quickly than in a pot on the counter. Students used reasoning related to NGSS performance expectations for PS1.A Structure and properties of matter (DCI) and an understanding of energy flow (CC) related to performance expectations of Energy and Matter (2013). For the score to be classified as MDC, students would first demonstrate an understanding that the water evaporated (DCI). Second, students would attribute causality (CC) to the heat transferred from the stove to the water. Measured somewhat differently than other items, correctly reporting either the CC or the DCI was enough for a correct score.

Scoring occurred in a two-cycle process. The first cycle involved human scoring while the second cycle used human scores to train the machine algorithm to score. Figure 2 shows the flow of the two-cycle process which began with rubric development and ended in a completed algorithm which could instantly score remaining responses.

Cycle 1: Human Training and Scoring

Ten undergraduates were recruited to score students' responses. These undergraduates were in their junior or senior years of a natural science major (nine physics students and one biology student) and had completed at least two college-level courses in both physics and chemistry. The raters participated in training sessions and scored in an iterative process that included a cycle of rigorous training sessions, calibration, and revisions for clarification. This was followed by a second cycle that involved calibrating the machine's scores by providing the human scores and responses.

As described above, the process shown in Fig. 2 began with construction of the multidimensional rubrics. When the rubrics were completed, raters were trained, and then proceeded to practice scoring. As training, raters scored a randomized sample of student CRs. The randomization procedure considered students' knowledge level as well as ethnic, racial, and geographic factors. AACR recommends a minimum human agreement of $k = 0.80$ before compiling a training set for the machine, so low inter-rater reliability (IRR) meant continued rater training. When the raters achieved a high IRR for all raters from the small practice

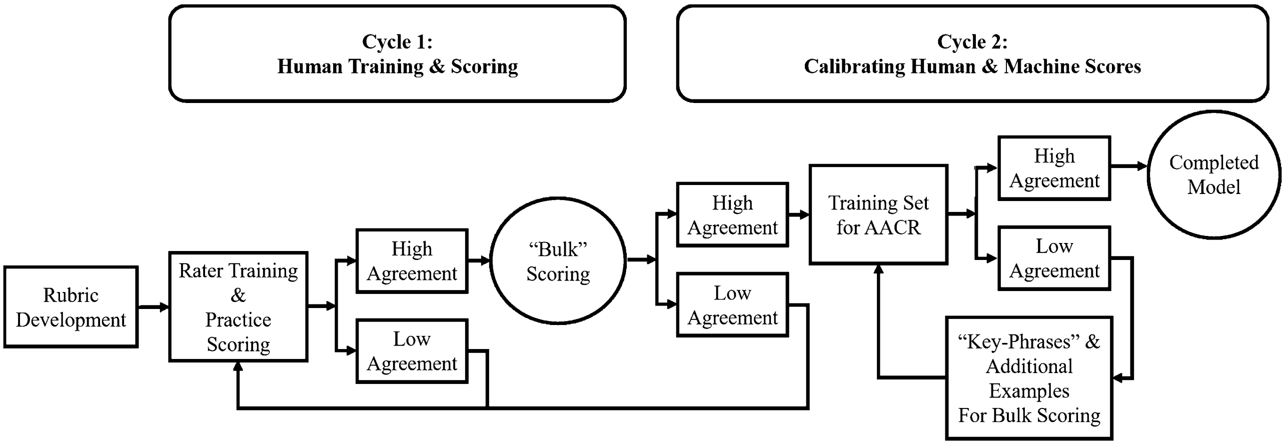


Fig. 2 Training and algorithm development

sets, they next completed "bulk" scoring sets of various sizes.

After bulk scoring, additional IRR testing ensured sustained agreement. If the IRR was low on the bulk scoring, raters returned to training. With high agreement for human raters from the bulk scoring, human scores were compiled to create a training set for the AACR algorithm. Because the success of the algorithm may depend on the quality of the data, it is imperative to create a quality training set with high inter-rater agreement. Therefore, it is very important to address issues suspected to reduce reliability in human scoring of multi-dimensional open-ended CR items.

The Challenges for Human Raters and How We Addressed Them

It became apparent that the rubrics needed to be explicit in what was expected of the student. The first consideration is to list all possible solutions. From item 3: properties of solutions, we learned that some issues in scoring arose, unexpectedly, from raters' advanced knowledge of science and that rubrics needed to state nearly all possible correct experiments. Agreement typically improved with each round of scoring, as new experiments were identified and included in the rubric.

The second consideration was to create a hierarchy outlining the importance of each dimension's contribution to the score. Disagreements arose when rubrics did not explicitly state which dimension was being measured. In scoring item 1: experimental design, for example, raters disagreed on whether stating a direct claim was as important as knowing how to control for variables in the experiment. After the training session where this issue emerged, the human agreement fell from $k = 0.71$ to $k = 0.38$ (shown later in Table 1). To correct for this in later scoring sessions, we revised the rubric and created a "dimensional hierarchy" for our raters, explicitly stating which dimensions and which specific aspects of each dimension were being measured. This process took too long, however, and due to subsequent changes to the scoring team during the process, the item did not make it to machine scoring during our collaboration with AACR.

The third consideration was to carefully weigh the choice to make changes to the scoring team. When the agreement is high, changes can drop that agreement rapidly, but the low agreement can be improved by training new raters on an item. Consistent training and calibration sessions helped to bring new raters into agreement, but new scoring teams did not easily come to agreement with past raters' scores. Often, this meant re-scoring assignments done by the previous rating team because the machine needed consistent agreement across the entire training set. When possible, maintaining the same scoring team until completion of the training set can help sustain a high inter-rater agreement. Conversely, the low agreement can be improved by introducing the item to a new team if the original team showed a lower than desirable agreement.

Bulk Scoring Criteria

AACR requested large training sets, with at least one hundred classified responses of each scoring proficiency category for each item. To meet this requirement, raters began to score in larger "bulk" sets. When their agreement was high, raters scored independently, but continued to score some content overlapping between raters. This allowed for continued checks for inter-rater agreement, which was calibrated after each wave of scoring so that raters could discuss disagreements. Scoring continued in this way until raters successfully scored at least 100 responses representing each of the three proficiency categories. Generally, two or three raters were selected to score a given item based on their shared availability and the scoring was assigned to these pairs or triads. The number of responses scored in a wave was determined by the raters' availability and the number of responses needed to obtain 100 examples of each proficiency category.

Cycle 2: Constructing the Training Set

In the second cycle, focused on generating a quality training set, we used the consensus, or median, score taken between raters who had scored a response in common. If a response was scored by only one rater, the individual score was considered the median. If the median score did not fall into one of the scoring categories (incorrect, correct, or MDC), the response was omitted from the training set and returned to the pool of unscored responses. Due to lower inter-rater agreement on item 3, a triad was used somewhat differently. A rater pair scored all responses in common, and the third rater in the triad acted as a tiebreaker to generate the consensus score. After the bulk scoring process, the human scores and student responses were used to develop training sets for the algorithmic models.

For each item, a robust training set was designed to examine key lexical features associated with multi-dimensional reasoning specific to the CR item. The responses and corresponding consensus scores for the three successfully scored items were given to AACR to create a predictive model. AACR developed models specific to each item using a cross-validation approach by using a portion of the scored responses to create the algorithm and reserving the rest to test the model. AACR accomplished this using feature extraction analysis. The AACR Web Portal examined each response in the training set by its lexical features, primarily combinations of a number, “*n*,” with words called “n-grams” to tune parameters for the algorithm development.

To validate the accuracy of the machine algorithms, AACR Web Portal applied a cross-validation approach which was found to be the most effective when compared with split- and self-validation methods (Zhai et al., 2020b). Using cross-validation, the machine first partitioned the data into *n* subsets, named “n-folds.” A random selection of (n-1) folds of human-scored responses was used to train the machine and develop the algorithmic model, which was then used to score the remaining one-fold student responses. The machine scores of the one-fold of responses were compared with human scores to calibrate the machine-human agreement, which was indicated by parameters such as Cohen’s kappa. The training, scoring, and comparisons were iterated *n* times so that each fold of data played a role as both the training and testing sets. The average of Cohen’s kappa generated in these processes indicated the accuracy of the algorithmic model. At the same time, the algorithmic model generated a computer confidence parameter which helped to diagnose which specific responses were difficult for the algorithm to score. After receiving the results of the bulk scoring, the scoring team reviewed cases where the human and machine scores disagreed. From this point, the process diverged for the three items for which human raters were in high enough agreement to move on to the machine scoring process. Each question provided unique challenges in developing the algorithmic model.

Challenges in the Preparation of Machine Training Data and How We Addressed Them

We first considered how to bolster low human-human agreement using tie breakers. Issues arose in reaching high human-human agreement when scoring atypical responses or CRs which were open-ended. For item 3: properties of solutions, we employed a tiebreak method. This method is similar to that used by Haudek and colleagues (2019) where raters trained until achieving a human-human agreement of $k=0.60$ or greater then scored responses individually with some responses overlapping between raters. A third rater would break ties in the disagreements. Haudek and colleagues’ results showed that the machine-human agreement was similar to human-human agreement, and machine-human agreement was higher than human-human agreement for some constructs.

The second consideration was in handling underrepresentation. Human raters examined the scoring discrepancies for lexical patterns, termed as “key phrases.” Key phrases emerged for some items, but not others. Key phrases (e.g., “velocity” or “relative”) were much more apparent in item 2: relative motion. For item 3: properties of solutions, raters coded the key phrases as the experiments used in student responses. Item 4: states of matter showed frequent use of specific terms, but those terms did not seem to impact the machine scoring.

Because underrepresentation can be problematic, researchers must remain aware of the potential for responses to be scored incorrectly. When lexical patterns emerged around these errors, it was feasible to predict future discrepancies. In this study, we selected additional responses with key phrases that were less represented in the overall sample but seemed common in the machine-human disagreements. Unscored responses containing these key phrases were then mixed with random responses and added to the next wave of bulk scoring. Where there were not enough of the potentially problematic responses to include more examples in the training set, we called upon human raters to review and score responses where errors were likely.

Data Analysis

To answer the first research question, we reported the human-human agreements indicated by Cohen’s kappa by wave of scoring for each item. To answer the second research question, we calibrated the machine-human agreements for each item, using the Cohen’s kappa, and compared the agreements with the corresponding human-human agreements. We also reported the machine scoring accuracy according to the dimensions of science learning. To answer the third research question, we calculated the frequency with which key phrases were used, the frequency with which they were scored incorrectly, and the percentage they comprised the total disagreements.

Results

Reliability of Human Raters in Scoring the Multi-dimensional Assessments

Table 1 shows the cumulative agreement for human raters over the 8-month scoring period described above. Responses were scored in successive waves. For each wave, the number of responses overlapping between raters to check IRR is shown. Human raters achieved a final Cohen’s kappa after several-wave training as $k_1=0.67$, $k_2=0.80$, $k_3=0.64$, and $k_4=0.76$ for the four items, respectively. According to a criterion proposed by Fleiss (1981), kappa values over 0.75 indicate excellent agreement while values between 0.40 and 0.75 indicate good agreement. According to this criterion, the human rater reliability is excellent for two of the items and good for the others. We also found that human scoring reliability increased with successive calibration training meetings. For the most successful item, item 2: relative motion, agreement increased from $k=0.72$ and peaked at $k=0.88$ over the successive waves of training and scoring. For three of the four items, agreement between human raters was high enough to move to machine learning during the collaboration with AACR.

For item 2: relative motion, training sets were compiled for the algorithm after wave 3 (bulk scoring), wave 5 (key phrases), and wave 7 (bulk scoring). Drops in the human rater agreements correspond to those time periods between waves of scoring, during which raters were waiting for and analyzing the results of the predictive model. For instance, agreement fell from $k=0.88$ to $k=0.80$ between waves 3 and 4 where the predictive model was tested.

Changes in agreement also sometimes corresponded to considerable changes in the composition of the scoring team. This drop can be seen in the agreement for wave 2 of item 3: properties of solutions, when new scoring members joined the existing team. Although item 1: experimental design was not scored by AACR due to the low human agreement, that agreement is shown to improve after introducing a new rubric to a new scoring team. Between waves 2 and 3, agreement increased from $k=0.38$ to $k=0.73$ when the item was reintroduced to an entirely new team.

Machine-Human Agreement vs. Human-Human Agreement

Table 2 shows the mean score awarded by humans, the mean score awarded by the machine, and the machine-human agreement for each wave of machine scoring. All rounds of scoring achieved fair to good agreement (Cohen's $k=0.64$ to $k=0.81$), even with as few as 336 responses in the smallest training set. Criteria for machine scoring, proposed by Nehm and Haertig (2012), consider Cohen's kappa between 0.41 and 0.60 as moderate, between 0.61 and 0.80 as substantial, and over 0.80 as almost perfect. According to these criteria, the machine scoring outcomes for the three items were categorized as substantial to almost perfect. As shown in

Table 1 Human agreement by wave

		Wave			Cumulative			
Wave	$N_{Overlap}$	Accuracy	k	$N_{Cummulative}$	Accuracy	k	N_{Total}	Note
Item 1: experimental design								
1	99	0.81	0.71	99	0.81	0.71	0	Practice
2	50	0.58	0.38	50	0.58	0.38	0	Practice
3*	100	0.84	0.73	100	0.84	0.73	0	Practice
4	25	0.80	0.67	25	0.80	0.67	0	Practice
Item 2: relative motion								
1	60	0.83	0.72	60	0.83	0.72	51	Practice
2	60	0.85	0.77	90	0.83	0.74	80	Practice
3	90	0.93	0.88	180	0.88	0.81	439	Bulk scoring
4	30	0.87	0.80	210	0.88	0.81	564	Key phrases
5	0			210	0.88	0.81	602	Key phrases
6*	33	0.85	0.76	243	0.87	0.80	740	Bulk scoring
7*	75	0.87	0.80	318	0.87	0.80	808	Bulk scoring
Item 3: properties of solutions								
1	30	0.77	0.63	30	0.77	0.63	0	Practice
2*	190	0.73	0.56	190	0.73	0.56	0	Practice
3**	70	0.72	0.57	70	0.72	0.57	70	Bulk scoring
4**	400	0.78	0.65	470	0.77	0.64	465	Bulk scoring
Item 4: states of matter								
1	30	0.93	0.86	30	0.93	0.86	88	Bulk scoring
2	19	0.79	0.62	49	0.88	0.77	242	Bulk scoring
3*	25	0.85	0.72	74	0.87	0.75	524	Bulk scoring
4*	25	0.89	0.80	99	0.87	0.76	594	Bulk scoring

$N_{Overlap}$ is the number of items raters scored in common. $N_{Cumulative}$ is the total number of jointly scored responses in all combined waves. N_{Total} is the number of total responses scored which can be sent to the machine. A wave number followed by * indicates that this wave, and those following, were scored by a new team, while ** indicates a third rater was added to tiebreak

Table 2 Descriptions of both human and machine scores

Wave (sample)	Mean (*SE*)		k (*SE*)
	Human	Machine	
Item 2: relative motion			
1 (484)	1.83 (0.03)	1.76 (0.03)	0.78 (0.02)
2 (662)	1.89 (0.03)	1.83 (0.03)	0.78 (0.02)
3 (808)	1.85 (0.03)	1.81 (0.03)	0.81 (0.02)
Item 3: properties of solutions			
1 (468)	1.76 (0.04)	1.68 (0.04)	0.69 (0.03)
Item 4: states of matter			
1 (336)	2.59 (0.03)	2.60 (0.03)	0.76 (0.04)
2 (594)	2.51(0.02)	2.49 (0.02)	0.64 (0.03)

Item 1 is not shown because the item did not proceed to machine scoring

Table 2, the agreement between the machine and humans was as high or higher than the human-human agreement reported for the cumulative waves of scoring for two of the three items (for human-human agreement, see Table 1).

In human scoring for item 2: relative motion, raters returned a cumulative $k = 0.80$ after all waves of scoring. The selected key phrases did not significantly improve scoring for the second wave of items sent to AACR. Despite adding all examples of the key phrases from the full data set, they still comprised only 0.50% to 8.79% of the responses scored for the final training set. To build the final training set, we added additional responses without a focus on identifying key phrases for a total of 808 student responses. AACR returned their model with higher agreement to the human raters ($k = 0.81$) than the final cumulative agreement between humans ($k = 0.80$).

The lowest human agreement for any item sent to AACR was item 3: properties of solutions. By having two raters score all responses together and a third rater's scores used as the tiebreaker, the algorithm matched more closely to the humans ($k = 0.69$) than the humans agreed with each other ($k = 0.64$). Although further predictive models were not developed for this item, the method yielded substantial agreement between the machine and human scores, despite lower human–human agreement.

For item 4: states of matter, agreement was lower in the second round of scoring ($k = 0.64$ compared with $k = 0.76$). Waves of bulk scoring were sent to AACR after waves 2 and 4 of human scoring. Despite adding an additional 258 responses and achieving similar cumulative human-human agreement for the second set ($k = 0.77$ to $k = 0.76$), agreement between the machine and humans still fell.

Accuracy of the Machine Scoring Associated with Dimensions of Learning

To better understand the capabilities of the machine scoring algorithm, we compared the machine and human scores by the associated dimensions of learning. Table 3 shows the distribution of scoring proficiency classifications for humans in all 4 items, how the machine classified the same responses, and agreement for the three items that had sufficient human-human agreement to move to machine scoring. The accuracy, or percent agreement, ranges from 28 to 93% for the individual proficiency levels for each item regardless of the associated dimensions, but the machine scored with accuracy greater

Table 3 Human and machine percentage of score, agreement, certainty, and dimensionality

Proficiency Level	Item dimensionality DCI, CC, SEP	Human	Machine	Mean Probability (SE)	Accuracy
Item 2: relative motion ($N = 808$)					
Incorrect	Incorrect	37.87	39.98	90 (0.00)	93.14
Correct	DCI	39.36	38.74	91 (0.01)	86.79
MDC	DCI + CC	22.77	21.29	82 (0.01)	78.80
Item 3: properties of solutions ($N = 468$)					
Incorrect	Incorrect or DCI only	50.32	57.63	85 (0.01)	92.31
Correct	DCI + SEP	23.44	16.56	78 (0.01)	59.63
MDC	DCI + CC + SEP	26.24	25.81	80 (0.01)	79.51
Item 4: states of matter ($N = 594$)					
Incorrect	Incorrect	3.54	2.19	68 (0.03)	28.57
Correct	DCI or CC	41.75	46.30	83 (0.01)	83.87
MDC	DCI + CC	54.71	51.52	87 (0.01)	82.77

Item 1 is not shown because it did not proceed to machine scoring. Mean probability refers to the prediction returned from AACR that a given score was correct.

MDC multidimensional correct, *DCI* disciplinary core ideas, *CC* crosscutting concepts, *SEP* science and engineering practice than 59% for all categories which were well represented in the sample. The lowest overall accuracy (28.57%) and lowest reported certainty of score (0.68) corresponded to the least represented classification which is the incorrect category for item 4: states of matter. This proficiency level comprised less than 4% of the training sample. For each item, the lowest accuracy and lowest certainty of score correspond to the category with the least representation.

Table 3 shows that the machine classified student responses with a similar distribution to human scores but with a tendency to score a little lower than human raters. The machine awards between 0.4 and 3.2% fewer MDC proficiency classifications for each question. This is similarly reflected in Table 2, where the mean score awarded by the machine is slightly lower than for humans in nearly all cases. Table 3 also shows that for items where all proficiency levels were well represented, the machine scored the incorrect classifications with higher accuracy than other categories. This was not reflected in the machine's predicted certainty however, and it cannot be asserted that the machine reported the highest confidence in scoring incorrect responses.

For item 2: relative motion, Table 3 shows that the machine showed high accuracy in scoring both the correct and MDC proficiency classifications. The use of only the DCI for a correct answer was scored with an accuracy 86.79%, and MDC responses which combined the use of the DCI with cause and effect (CC) were scored with an accuracy 78.80% when compared with human classifications.

As shown in Table 3, scoring for item 3: properties of solutions was fully three-dimensional. The relative amount of human and machine scores for two-dimensional (correct) responses were 23.44% and 16.56%, respectively. The machine matched more closely to humans for three-dimensional (MDC) responses, which comprised 26.24% of the human proficiency classifications for this item and 25.81% of machine scores.

Item 4: states of matter was scored differently from item 2, as it allowed for the use of either of two different dimensions of reasoning for partial credit. For a correct response, students could attribute the phenomenon to the evaporation of water into smaller particles (DCI) or reason that it was caused by the heat of the stove (CC). Human scorers classified responses as correct 41.75% of the time while the machine used this proficiency level for 46.30% of responses. The machine and humans classified responses with the MDC proficiency level 51.52% and 54.71%, respectively. The correct and MDC responses were both well represented in the sample and each were scored with accuracy over 82%.

Key Phrases in Scoring Open-Ended Constructed Response

As demonstrated in Table 4, the machine can score open-ended CRs despite the varied language engaged by students. For item 2: relative motion, the majority of students (56.06%) chose to use common phrases (e.g., "speed") to describe the phenomenon while a smaller proportion (8.79%) used advanced vocabulary (e.g., "velocity"). Student responses with more advanced vocabulary use seemed to be proportionately more represented among those that were scored incorrectly by the machine. For example, 18.31% of student responses that include the formal phrase "velocity" were scored incorrectly compared with 11.70% of responses that used the word "speed." As shown in Table 4, these key phrases were often characterized by including commonly used phrases with a high prediction of a correct score, or an atypical response accompanied by a low prediction score. Examples of student responses, the human score, the machine score, and the certainty of the machine score for item 2: relative motion can be found in Appendix 5.

For item 3: properties of solutions, students provided numerous experiments, including boiling or evaporating the water to look for remaining residue. Human raters coded key phrases as words associated with the types of experiments used by students as shown in Table 4 (e.g., evaporate, smell). Item 3 had lower human agreement ($k=0.64$) than the other items scored with the machine ($k=0.80$ and $k= 0.76$), but still fell within the boundaries of substantial agreement ($k=0.61$–0.81). We found, generally, that the percentages of machine-human disagreements of each key phrase were consistent with the frequency with which they appeared in student responses. For instance, the word "density" was used by 13.98% of students in the sample and accounted for 10.34% of the responses where the machine and humans disagreed.

Item 4: states of matter shows a similar trend where the percentages of machine-human disagreements were also consistent with the frequency with which they appeared in student responses. For instance, the use of "into the air" was used in 15.66% of all responses and comprised 13.51% of all disagreements. The key phrases selected for this item were used in at least 15% of the responses and were scored incorrectly in similar proportions ranging from 15.08 to 16.14% of their total use.

Discussion

The performance expectations embodied in NGSS for chemistry and physics cannot be effectively measured unless meaningful and scorable three-dimensional assessments are developed (Cheuk et al., 2019; Pellegrino, 2013; NRC, 2014). Because such assessments require the other dimensions to be used together with scientific practices, they may involve a variety of performance-based tasks, such as writing short answers or drawing, to capture students' mastery of performance expectations (Pellegrino, 2013; NRC, 2014).

Table 4 Key phrases associated with the machine scoring

Key phrase(s)	% of all responses	% Scored incorrectly	% All MH disagreements	Mean probability (SD)
Item 2: relative motion ($N_{\text{Responses}}=808$ and $N_{\text{Disagreements}}=102$)				
Velocity and relative	0.50	50.00	1.96	77 (0.10)
Relative	0.87	71.43	4.90	81 (0.10)
Velocity or relative without speed	7.18	20.69	11.76	84 (0.09)
Fast	7.80	12.70	7.84	87 (0.11)
Velocity	8.79	18.31	12.75	85 (0.09)
Speed	56.06	11.70	51.96	89 (0.10)
Item 3: properties of solutions ($N_{\text{Responses}}=465$ and $N_{\text{Disagreements}}=87$)				
Taste	1.08	0.00	0.00	80 (0.07)
Freeze	2.15	10.00	1.15	83 (0.11)
pH	2.80	7.69	1.15	80 (0.08)
Dissolve	3.01	7.14	1.15	80 (0.13)
Smell	6.02	14.29	4.60	83 (0.10)
Mass or weight	7.74	16.67	6.90	83 (0.09)
Evaporate	7.96	21.62	9.20	83 (0.11)
Boil	10.97	19.61	11.49	82 (0.11)
Density	13.98	13.85	10.34	83 (0.10)
Item 4: states of matter ($N_{\text{Responses}}=594$ and $N_{\text{Disagreements}}=111$)				
Steam	37.54	16.14	32.43	86 (0.10)
Into the air	15.66	16.13	13.51	88 (0.09)
Heat and evaporation	20.37	14.05	15.32	89 (0.09)
Heat	21.21	15.08	17.12	89 (0.10)
Evaporation	37.54	15.25	30.63	86 (0.10)

Item 1 is not shown because it did not move to machine scoring. Mean Probability refers to the prediction made by AACR that the algorithm assigned the same classification as the human raters

The NRC (2014) calls for investment of time and other resources into the development of these new assessments and to facilitate the implementation of these three-dimensional assessments in the classroom; this includes “existing and emerging technologies” that support scoring. In accordance with this initiative, this study built upon prior work of developing multi-dimensional assessments and implementing machine learning approaches for scoring those assessments.

This study demonstrated the ability of automated analysis to facilitate the transition to multi-dimensional assessment by showing high agreement between computers and humans when scoring CR items. Machine learning could facilitate scoring three-dimensional assessments more quickly than human scoring alone, allowing teachers and researchers to collect detailed information on students’ knowledge-in-use as recommended by the NRC. The findings have contributed to our knowledge by building on a foundation of research focused on machine scoring, which has previously been applied for scoring key concepts (Nehm & Haertig, 2012) and argumentation (Cheuk et al., 2019), among other purposes. This study has added to the literature surrounding machine scoring of CR by providing a comparison of multiple items, showing how each item was scored by humans and the machine algorithm, and demonstrating how each item aligned to the NGSS performance expectations.

Using an objective measure, this study analyzed student responses to determine the use of two- and three-dimensional learning to describe phenomena. Through specially developed rubrics, this study has shown that the machine algorithm could score accurately when students engaged reasoning associated with CCs. All three of the items scored by the machine included a CC to differentiate between correct and incorrect classifications. In fact, the machine was able to successfully classify students’ use of a single dimension or multiple dimensions for each item, with accuracy comparable with the human raters. The machine algorithm also scored similarly to human raters when a correct response included any of multiple possible experiments.

It is important to note that it took longer than expected to develop scoring models because this was a training exercise where the items were neither designed to be three-dimensional nor to be scored with machine learning. As such, it required a great deal of thought, training, rubric development, and IRR testing in an iterative process. This has allowed for exploration of very open-ended response items showing that machine algorithms can attain high accuracy on sufficiently large data sets, and even train the machine to correctly classify responses by the associated dimensions. Once the algorithm obtained high agreement between the humans and the machine, AACR was able to instantly score the remaining responses.

While prior studies have collected evidence that supervised machine learning can be used successfully in automatic scoring, some argue that when presenting the results of machine learning, training sets are not discussed sufficiently in presenting their results (Geiger et al., 2020). Few studies explicitly describe whether and how their assessments and rubrics target the three dimensions of science learning, or how well machine learning classifies student responses into the different categories. Consequently, we have limited knowledge about how machine scoring can support three-dimensional assessment practices. In this study, we have discussed each item used in the analysis, including how the rubrics tap each dimension and the details of the training set. By doing so, we were able to examine the machine's capacity to classify and score items based on the dimensionality of scientific knowledge employed by students.

Consistent with other studies, our results suggest that rater calibration and sample size might be significant factors impacting machine performance. These issues can be mitigated with continued rater training and larger sample sizes or "training sets" for the machine to build its algorithmic models (Balfour, 2013; Cheuk et al., 2019). Additionally, given that the quality of human-scored training data might be critical to machine scoring (Balfour, 2013), we examined how to improve human scoring of multi-dimensional assessments to facilitate machine performance. In this course, successful results emerged from our study that might be valuable for future applications in machine scoring. On scoring these multi-dimensional CRs, we found that the best method to train the raters, among those we tried, was not just to explain the correct answer in the rubric, but to also inform raters which dimensions were being scored. The human scoring or coding of multi-dimensional responses should include explicit rubrics that list all possible solutions and show the hierarchy of importance for the dimensions being measured. Comprehensive rubrics helped humans to score consistently. It is also important to carefully weigh decisions to change the composition of the scoring team.

Advanced vocabulary or key phrases (e.g., "velocity") may increase the challenge for machine scoring, as compared with informal key phrases (e.g., "speed"). We suspect that this finding may be associated with the fact that fewer students used formal key phrases than those who used the informal alternatives. This concept of representation appears again where the algorithm's lowest agreement to humans coincides with the least represented scoring proficiency.

Because machine learning has a difficult time scoring more unique texts (Balfour, 2013), and students could correctly propose many experiments, researchers hypothesized poor results for item 3: properties of solutions. Despite the broad range of possible answers, we obtained good to substantial agreement ($k = 0.69$) between human raters and the machine. This was similar to the results of Haudek and his colleagues (2019) where they obtained higher agreement between humans and the machine than between human raters for some constructs. The results from this study also show that it is possible to bolster a training set with low human-human agreement through tiebreakers.

Addressing the complications in building an accurate algorithmic model was beneficial at multiple levels. The search for lexical patterns in scoring discrepancies not only facilitated construction of a more robust model for scoring, but also identified human raters' errors even after sufficient agreement was reached. Even with high human-human agreement, we found that some discrepancies corresponded to human errors. We were then able to address these errors with raters directly to prevent recurrence. Outside the context of this paper, reviewing discrepancies between humans and the machine provided insight into broader vocabulary use and creativity in responses.

Limitations

As exploratory research, this study adopted existing items in a national test and developed multi-dimensional rubrics according to NGSS to score students' responses. While this study collected evidence indicating sufficient machine capacity in terms of scoring responses according to the dimensionalities, there were limitations to this study. Given that the assessment tasks were adopted from a national test that was not originally designed to be three-dimensional, we only have one three-dimensional item while the other three are two-dimensional. Though this study achieved high machine-human agreements for these items, future studies should develop more three-dimensional items and test the machine capacity to automatically score three-dimensional assessments. Because three-dimensional assessments are more complex than two-dimensional assessments, this could be more challenging for the machine to score.

Conclusions and Implications

The benefits of applying the three dimensions of science learning, by incorporating SEPs, DCIs, and CCs, have been proposed in both the NRC Framework (NRC, 2012) and the NGSS (NGSS Lead States, 2013). This concern is particularly pronounced because most multi-dimensional assessments contain CRs and scoring of CRs is both time and labor intensive. This implies notable efforts for human scoring on behalf of educators, state departments, and researchers. This study found that human experts were able to reliably (Cohen's $k > 0.60$) score student responses to assessment items with varied dimensions. This study shows that machine scoring was capable of classifying student responses when measured by the use of the dimensions of learning spelled out by the NGSS, with accuracy that was comparable with human experts. This shows promise for the use of machine learning to facilitate the measurement of in-depth science understanding and meeting the recommendations for science assessment from the NRC. Once assessments are constructed, and a number scored, the remainder of a large sample can be scored almost instantly.

This study indicates that the automated analysis of two- and three-dimensional CR items may be a viable solution in reducing both financial and time costs associated with measuring in-depth science knowledge to facilitate the gradual shift to follow NRC guidelines. Despite the labor involved in constructing the rubrics, training raters, and developing algorithms, once the models were complete, the machine algorithm could continue to score the remaining students' CRs rapidly. Machine scoring of three-dimensional assessments could have several meaningful impacts on state or national standardized testing, and the monitoring of student performance in the classroom. With coordination, quality three-dimensional assessments could be delivered online and CRs scored almost instantly. However, given what we have learned, this is a complex process both in terms of identifying dimensionality and building rubrics which can be reliably scored. If it is possible to develop items that are three-dimensional and create rubrics for them, this would be an important contribution to meeting new science education reform efforts. Machine scoring could facilitate the use of more robust measures of students' understanding through knowledge-in-use assessments.

Acknowledgments We are grateful to the team members: Dan Oleynik, Jacob Crosby, Taryn Stefanski, Cullen Hudson, Eleonora Baker, and Rachel Marias-Dezendorf.

Funding This material is based upon work supported by the National Science Foundation under Grant No. OISE: 1545684.

Compliance with Ethical Standards

Conflict of Interest The authors of this paper have no conflict of interest in the publication of this paper.

Informed Consent was obtained from all individual participants involved in the study.

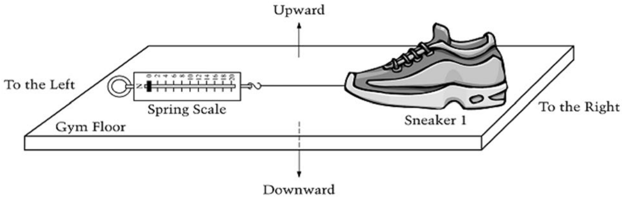
Ethical Approval All procedures performed in studies involving human participants were in accordance with the ethical standards of the institutional and/or national research committee and with the 1964 Helsinki Declaration and its later amendments or comparable ethical standards.

Appendix 1

Item 1: Experimental Design Text and Rubric

Question text

Meg designs an experiment to see which of three types of sneakers provides the most friction. She uses the equipment listed below. 1. Sneaker 1 2. Sneaker 2 3. Sneaker 3 4. Spring scale. She uses the setup illustrated below and pulls the spring scale to the left.



Meg tests one type of sneaker on a gym floor, a second type of sneaker on a grass field, and a third type of sneaker on a cement sidewalk. Her teacher is not satisfied with the way Meg designed her experiment.

A. Describe one error in Meg's experiment.

Alignment to the NGSS (2013) Performance Expectations

Dimension	Grade-band	Performance expectation
DCI	6–8	ETS1.A Defining and Delimiting Engineering Problems
CC	(N/A)	(N/A)
SEP	3–5	Planning and Carrying Out Investigations

Student example and multi-dimensional rubric

Multi-dimensional correct	Correct	Incorrect
"Meg's error is that she is testing three experiments in separate and different settings, allowing the experiments to have different outcomes. This stops her from knowing if her other shoes work on a gym floor or grass field or a cement sidewalk."	"Meg should have tested the sneakers in the same location for each test."	"Meg should've used different types of sneakers, not the same."
DCI: Student correctly identifies the error in the experimental setup. & SEP: Student explains this is a failure to control for variables or that the results cannot be compared.	DCI: Student correctly identifies an error in the experimental setup. & No SEP: Student does not explain that it controls for relevant variables.	Provides an incorrect response or irrelevant error in the experimental setup.

Appendix 2

Item 2: Relative Motion Text and Rubric

Question text

Suppose you are riding in a car along the highway at 55 miles per hour when a truck pulls up along the side of your car. This truck seems to stand still for a moment, and then it seems to be moving backward.

A. Tell how the truck can look as if it is standing still when it is really moving forward.

Alignment to the NGSS (2013) Performance Expectations

Dimension	Grade-band	Performance expectation
DCI	6–8	PS2.A Forces and Motion
CC	6–8	Scale and Proportion
SEP	(N/A)	(N/A)

Student example and multi-dimensional rubric

Multi-dimensional correct	Correct	Incorrect
“The truck looks as if it is standing still as both your car and the truck are moving at 55 mph in the same direction.”	"It is going 55 miles per hour, which is as fast as the car is going."	“the truck looks like it is still because it is losing speed."
DCI: Student relates the truck’s speed to the speed of the observer.	DCI: Student relates the truck’s speed to the speed of the observer.	Student provides an incorrect/irrelevant explanation for the phenomena OR only restates the question.
&	&	
CC: Student states that equal relative speeds would cause the truck to appear as though it is standing still.	No CC: Student does not discuss the visual phenomenon being caused by the relative speeds.	

Appendix 3

Item 3: Properties of Solutions Text and Rubrics

Question text

Maria has one glass of pure water and one glass of salt water, which look exactly alike. Explain what Maria could do, without tasting the water, to find out which glass contains the salt water.

Alignment to the NGSS (2013) Performance Expectations

Dimension	Grade-band	Performance expectation
DCI	3–5, 6–8	PS1.A Structure and Properties of Matter
CC	6–8	Cause and effect
SEP	6–8	Planning and Carrying Out Investigations

Student example and multi-dimensional rubric

Multi-dimensional correct	Correct	Incorrect
"Maria could use two similar cups and weigh them both and the heavier one is saltwater."	"Maria can weigh the cups that hold the water."	"Your body floats easier in salt water."
SEP: Student response describes an experiment that controls for relevant variables.	SEP: Student response describes an experiment that controls for relevant variables.	Student response does not describe an experiment that will differentiate fresh water from salt water.
DCI: The experiment isolates a measurement that will differentiate fresh water from salt water.	DCI: The experiment isolates a measurement that will differentiate fresh water from salt water.	
CC: Student indicates the expected result that will allow them to differentiate the fresh water and salt water.	No CC: Student does not indicate the expected result that will allow them to differentiate the fresh water and salt water.	

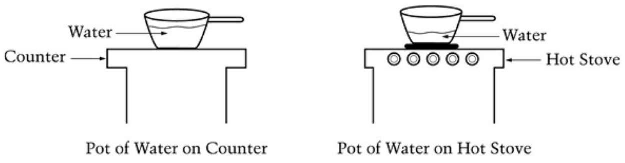
Appendix 4

Item 4: States of Matter Text and Rubrics

Question text

Anita puts the same amount of water in two pots of the same size and type. She places one pot of water on the counter and one pot of water on a hot stove.

After ten minutes, Anita observes that there is less water in the pot on the hot stove than in the pot on the counter, as shown below.



A. Why is there less water in the pot on the hot stove?

B. Where did the water go?

Alignment to the NGSS (2013) Performance Expectations

Dimension	Grade-band	Performance expectation
DCI	6–8	PS1.A Structure and Properties of Matter
CC	6–8	Energy and Matter
SEP	(N/A)	(N/A)

Student example and multi-dimensional rubric

Multi-dimensional correct	Correct	Incorrect
"The heat caused it to evaporate."	"The water evaporated."	"it dried up."
DCI: Student says the water evaporated.	DCI: Student says the water evaporated.	Provides an incorrect/irrelevant explanation.
&	OR	
CC: Attributes this to the heat from the stove.	CC: Attributes this to the heat from the stove.	

Appendix 5

Machine Errors and Certainty of Score for Student Responses to Item 2: Relative Motion

Student response	Predictive model	Human score	Machine score	Machine probability
“You share the same velocity and thus from your relative position moving alongside you, it doesn’t appear to move."	1st	MDC	Incorrect	0.65
	2nd	MDC	Correct	0.66
	3rd	MDC	Correct	0.62
"The speed of the truck in relation to your car is the same, not changing the distance between the two and creating the illusion that there is no movement of the vehicles."	1st	N/A	N/A	N/A
	2nd	MDC	Incorrect	0.80
	3rd	MDC	Incorrect	0.83
"You and the truck are moving at very similar speeds, creating the illusion that it is still."	1st	MDC	Incorrect	0.86
	2nd	MDC	Incorrect	0.77
	3rd	MDC	Incorrect	0.82
"its going 55"	1st	Correct	Incorrect	0.94
	2nd	Correct	Incorrect	0.93
	3rd	Correct	Incorrect	0.95
"The truck can seem to be looking as if it were standing still when the car is moving at a slower velocity than 55 miles per hour."	1st	N/A	N/A	N/A
	2nd	Incorrect	MDC	0.90
	3rd	Incorrect	MDC	0.90
"If the truck slows down and you are still going at the same speed it would appear that it stopped"	1st	Incorrect	MDC	0.72
	2nd	Incorrect	MDC	0.85
	3rd	Incorrect	MDC	0.86

References

AACR. (2020). September 4, 2020, Retrieved from https://apps.beyondmultiplechoice.org

Balfour, S. P. (2013). Assessing writing in MOOCs: Automated Essay Scoring and Calibrated Peer Review$^{\text{TM}}$. *Research & Practice in Assessment, 8,* 40–48.

Cheuk, T., Osborne, J., Cunningham, K., Haudek, K., Santiago, M., Urban-Lurain, M., Merril, J., Wilson,C., Stuhlsatz, M.,Donovan, B., Bracey, Z., & Gardner, A. (2019). *Towards an Equitable Design Framework of Developing Argumentation in Science tasks and Rubrics for Machine Learning. Presented at the Annual meeting of the National Association for Research in Science Teaching (NARST).* Baltimore, MD.

Fleiss, J.L. (1981). Statistical methods for rates and proportions (2nd ed.). New York: John Wiley. ISBN 978–0–471–26370–8.

Geiger, R. S., Yu, K., Yang, Y., Dai, M., Qiu, J., Tang, R., & Huang, J. (2020, January). Garbage in, garbage out? do machine learning application papers in social computing report where human-labeled training data comes from?. *In Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency* (pp. 325–336).

Ha, M., & Nehm, R. H. (2016). The impact of misspelled words on automated computer scoring: a case study of scientific explanations. *Journal of Science Education and Technology, 25*(3), 358–374.

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.

Haudek, K., Santiago, M., Wilson, C., Stuhlsatz, M.,Donovan, B., Bracey, Z., Gardner, A., Osborne, J., & Cheuk, T. (2019). *Using Automated Analysis to Assess Middle School Students' Competence with Scientific Argumentation, presented at the Annual Meeting of the National Council on Measurement in Education (NCME).* Toronto, ON.

Large, J., Lines, J., & Bagnall, A. (2019). A probabilistic classifier ensemble weighting scheme based on cross-validated accuracy estimates. *Data mining and knowledge discovery, 33*(6), 1674–1709.

Lee, H. S., McNamara, D., Bracey, Z. B., Liu, O. L., Gerard, L., Sherin, B., Wilson, C., Pallant, A., Linn, M., Haudek, K., & Osborne, J. (2019a). *Computerized text analysis: Assessment and research potentials for promoting learning*.

Lee, H. S., Pallant, A., Pryputniewicz, S., Lord, T., Mulholland, M., & Liu, O. L. (2019b). Automated text scoring and real-time adjustable feedback: Supporting revision of scientific arguments involving uncertainty. *Science Education, 103*(3), 590–622.

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

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.

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. *Educational Assessment, 23*(2), 121–138.

Mayfield, E., & Rosé, C. (2010, June). An interactive tool for supporting error analysis for text mining. *In Proceedings of the NAACL HLT 2010 Demonstration Session* (pp. 25–28).

Mayfield, E., & Rosé, C. P. (2013). *Open source machine learning for text*. Handbook of automated essay evaluation: Current applications and new directions.

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

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

National Research Council. (2014). *Developing assessments for the next generation science standards.* National Academies Press.

Nehm, R. H., & Haertig, H. (2012). Human vs. computer diagnosis of students' natural selection knowledge: testing the efficacy of text analytic software. *Journal of Science Education and Technology, 21*(1), 56–73.

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

Pellegrino, J. W. (2013). Proficiency in science: Assessment challenges and opportunities. *Science, 340*(6130), 320–323.

Zhai, X., Haudek, K., Shi, L., Nehm, R., Urban-Lurain, M. (2020a). 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., Haudek, K., Stuhlsatz, M., Wilson, C. (2020b). Evaluation of construct-irrelevant variance yielded by machine and human scoring of a science teacher PCK constructed response assessment. *Studies in Educational Evaluation*, *67*, 1-12. https://doi.org/10.1016/j.stueduc.2020.100916.

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

Zhu, M., Lee, H.-S., Wang, T., Liu, O. L., Belur, V., & Pallant, A. (2017). Investigating the impact of automated feedback on students' scientific argumentation. *International Journal of Science Education, 39*(12), 1648–1668.

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Sarah Maestrales is a graduate student in Measurements and Quantitative Methods at Michigan State University.

Xiaoming Zhai is an Assistant Professor in the Department of Mathematics and Science Education at University of Georgia.

Israel Touitou is a research associate in the CREATE for STEM Institute at Michigan State University.

Quinton Baker is a graduate student in the Department of Economics at Michigan State University.

Barbara Schneider is John A. Hannah University Distinguished Professor in the College of Education and the Department of Sociology at Michigan State University.

Joseph Krajcik is director of the CREATE for STEM Institute and the Lappan-Phillips Professor of Science Education at Michigan State University.