Journal of Science Education and Technology
https://doi.org/10.1007/s10956-020-09858-0

Comparison of Machine Learning Performance Using Analytic and Holistic Coding Approaches Across Constructed Response Assessments Aligned to a Science Learning Progression

Lauren N. Jescovitch[1] · Emily E. Scott[2] · Jack A. Cerchiara[2] · John Merrill[1,3] · Mark Urban-Lurain[1] · Jennifer H. Doherty[2] · Kevin C. Haudek[1,4]

Abstract

We systematically compared two coding approaches to generate training datasets for machine learning (ML): (i) a holistic approach based on learning progression levels and (ii) a dichotomous, analytic approach of multiple concepts in student reasoning, deconstructed from holistic rubrics. We evaluated four constructed response assessment items for undergraduate physiology, each targeting five levels of a developing flux learning progression in an ion context. Human-coded datasets were used to train two ML models: (i) an 8-classification algorithm ensemble implemented in the Constructed Response Classifier (CRC), and (ii) a single classification algorithm implemented in LightSide Researcher's Workbench. Human coding agreement on approximately 700 student responses per item was high for both approaches with Cohen's kappas ranging from 0.75 to 0.87 on holistic scoring and from 0.78 to 0.89 on analytic composite scoring. ML model performance varied across items and rubric type. For two items, training sets from both coding approaches produced similarly accurate ML models, with differences in Cohen's kappa between machine and human scores of 0.002 and 0.041. For the other items, ML models trained with analytic coded responses and used for a composite score, achieved better performance as compared to using holistic scores for training, with increases in Cohen's kappa of 0.043 and 0.117. These items used a more complex scenario involving movement of two ions. It may be that analytic coding is beneficial to unpacking this additional complexity.

Keywords Automated analysis · Machine learning · Learning progressions · Holistic rubrics · Analytic rubrics · Constructed response

Machine learning (ML) has potential for influencing education and has been increasingly applied specifically to science assessments (Kotsiantis 2012). Specifically, one promising aspect of ML is the potential for assessing more complex constructs that are difficult to capture with more traditional assessments (Zhai et al. 2020). However, little research has examined how certain approaches for, or technical features of, ML influence the success of assessing more complex constructs.

This study addresses the challenges of applying ML in a complex, science context to advance automated scoring model performance and automaticity of constructed response (CR) assessments aligned to a learning progression (LP). In previous studies, our research group focused on dichotomous analytic coding approaches of conceptual components within a student's response to capture the breadth of ideas in student writing. LPs, however, are holistic in nature and require an overall characterization of a response often containing several conceptual components. This study examines the ML and rubric scoring approaches best suited to score CR assessment items aligned to a LP in order to inform future research efforts. Specifically, we applied both an analytic and holistic coding

Electronic supplementary material The online version of this article (https://doi.org/10.1007/s10956-020-09858-0) contains supplementary material, which is available to authorized users.

✉ Lauren N. Jescovitch
jescovit@msu.edu

[1] CREATE for STEM Institute, Michigan State University, 620 Farm Lane, Room 115, East Lansing, MI, USA

[2] Department of Biology, University of Washington, Seattle, WA, USA

[3] Department of Microbiology and Molecular Genetics, Michigan State University, East Lansing, MI, USA

[4] Department of Biochemistry and Molecular Biology, Michigan State University, East Lansing, MI, USA

Published online: 26 September 2020

scheme to the same set of undergraduate written CR to determine whether complex constructs could be represented via both approaches. Further, we used these responses as training sets to compare ML models developed with these distinct coding approaches.

Background

Machine Learning of Constructed Response Assessments

Developing and employing CR assessments in a classroom context can provide information to help instructors to make educational decisions for student learning. This is in part because CR assessment items, which require students to answer a question in their own words, allow for a more in-depth analysis of students' content understanding, and elicit students' higher order thinking (Allen and Tanner 2006; Jönsson and Svingby 2007; Montgomery 2002). A common assessment format used in large enrollment, introductory Science, Education, Engineering, and Mathematics (STEM) higher education courses is multiple-choice (fixed response) (Nicol 2007). Multiple-choice questions are fast, simple, and easy to evaluate; however, these assessments inform the instructor very little about the heterogeneity of students' thinking. More complex scientific practices such as argumentation, explanation, and integration of core ideas may be difficult to measure in a multiple-choice assessment (Allen and Tanner 2006; Nehm et al. 2012). However, CR answers can be difficult to evaluate, interpret, and return feedback in a timely manner for both instructors and students (Gerard et al. 2019). Over the last few decades, many studies have reported on the utility of ML in education and assessments as a way of quickly and accurately evaluating written text (Zhai et al. 2020).

Machine learning is described by Mitchell (1997) as a "computer program that improves its performance at some task through experience" (p. 2). "Experience" in this context refers to information (e.g., labels in a data set, outcomes of previous trials) available to the machine. ML "programs" are usually computational methods and algorithms, as opposed to programmed "if-then" statements in traditional computer programming. ML has been used in various text-based applications, including natural language processing and text analysis with expert-coded data to predict classifications for new data, or more specifically to predict scores of student CR. Scoring in this context does not mean assigning points to a response, but classifying text responses into groups (e.g., responses which include scientific ideas about photosynthesis). Recent work on automated scoring of student responses has focused on *supervised ML approaches*, or processes using responses along with human codes to develop a predictive scoring model, rather than *unsupervised ML*, which does not use human codes as input variables, but rather attempts to identify patterns in the responses (Kotsiantis 2007). ML approaches typically include two phases: training and testing.

Typically, in supervised ML, human-labeled data are used to "train" the machine in order to generate a scoring model based on a set of attributes extracted from the data. Once the scoring model is established, it is "tested" by comparing the consistency of human labels and the machine labels on subsets of the same (or new) data (Jordan and Mitchell 2015). This is also the primary measure used to validate the ML scoring model: how closely the machine-predicted scores match human-assigned scores (Williamson et al. 2012).

Developing a scoring model for one assessment question using a training set of data requires an iterative process including development, use, and refinement of an expert validated scoring rubric and expert scores (or codes) (Nehm et al. 2010). Then, using the predictive models, instructors can submit data collected from a developed CR item and generate quick formative results of their students' thinking about the targeted construct. Thus, these models applied to new data sets can be used to characterize student thinking or ability to inform instruction and learning in the classroom.

Various automated scoring tools and approaches have been developed for evaluation of student CR in science, over a range of grade levels and topics (Liu et al. 2016; Mitchell et al. 2002; Mayfield and Penstein-Rose 2010; Nehm and Haertig 2012; Sieke et al. 2019; Sripathi et al. 2019). For example, automated scoring systems have been applied to conceptual assessments aligned with key disciplinary ideas in undergraduate biology (Nehm et al. 2012; Sieke et al. 2019). These scoring systems identify important conceptual components in student responses, which then can be used to classify responses. Other efforts have applied automated scoring systems to assess scientific practices in middle school science, such as argumentation (Mao et al. 2018; Haudek et al. 2019). These efforts identified key components of the practice of argumentation (e.g., claim) via automated scoring. Such scores were then used as feedback to individual students to help in revising their arguments (Lee et al. 2019), thereby demonstrating the potential of automated scoring systems. Finally, recent efforts have attempted to create automated scoring systems that align with cognitive models of learning, such as LPs (Anderson et al. 2018). These cognitive models represent a complex construct for automated scoring, in that they attempt to classify student responses into levels based on the sophistication of the response. For example, Anderson et al. (2018) and Thomas et al. (2019) used automated scoring to evaluate student responses on a LP for carbon transforming processes.

Although these works successfully applied ML to complex constructs like LPs, there are remaining challenges to unravel in order to facilitate broader application of ML techniques to

complex assessment constructs outside the immediate context of the study. For example, some challenges that occur during development of ML models for complex constructs include lexical diversity of student language, synonyms and abbreviations, infrequent or incorrect ideas, and overlapping qualifiers in rubric descriptions (Jescovitch et al. 2019a; Liu et al. 2014, 2016). These challenges are largely dependent on variables such as the content being assessed, the complexity of the coding rubric applied, and the sample of collected data; therefore, not all challenges while developing particular models can be anticipated.

Challenges in Applying Coding Approaches to Machine Learning

Some previous work in automated scoring has focused on conceptual (or analytic) scoring of student responses (Liu et al. 2014; Moharreri et al. 2014; Sieke et al. 2019). These studies have employed coding schemes to identify the presence or absence of specific ideas within students' text responses. Additionally, we noticed that analytic rubrics reduce coding complexity for humans, which may expedite ML model development and improve overall performance (Jescovitch et al. 2019a, 2019b). Analytic rubrics are defined in this experiment as using multiple bins to capture conceptual components, which are not mutually exclusive and are dichotomously scored (0, 1), where 1 indicates the presence of the given concept. Each analytic rubric bin is designed to represent a single concept and each response must be scored for each analytic rubric bin. Multiple concepts can be present within the same response, and therefore, the response could be scored as 1 in multiple analytic rubric bins. Some studies have found that analytic rubrics are more reliable in that they check key content components of reasoning and provide specific feedback to students (Jönsson and Svingby 2007; Jescovitch et al. 2019b; Yune et al. 2018).

Other previous work has used multi-leveled coding schemes to try to characterize the quality of complete student responses (Liu et al. 2016; Mao et al. 2018; Wiley et al. 2017). Such holistic coding schemes may be developed based on "correctness" of an explanation; however, not all such schemes may be closely tied to an underlying framework. Such frameworks are much-needed tools for organizing and executing specific biology education research agendas to uncover student thinking and guide instruction (Nehm 2019). One such possible framework to examine student performance is a LP. A LP describes "successively more sophisticated ways of reasoning within a content domain" for students (Smith et al. 2006, p. 1). Therefore, LPs may represent a unique application of a holistic coding approach, in that LPs are cognitive models of student learning and are structured to be an ordered classification of the development of student reasoning (Wilson 2009). Accordingly, rubrics aligned to LPs do not just identify a general measure of quality or completeness of a response but, instead, they identify specific developmental levels of reasoning within a content domain (Smith et al. 2006). Thus, LPs represent a more complex construct than just providing a correct explanation in a given scenario (Zhai et al. n.d.; Gotwals et al. 2012). Accordingly, there are challenges to be addressed when using automated scoring of LP associated assessments. One such challenge is whether the sophistication of reasoning in higher LP levels can be accurately and reliably identified by ML scoring models. This is particularly true in the case of undergraduate science, in which reasoning about phenomena can entail connecting many content pieces. Another challenge to be investigated is a key question in the current study: can the complex reasoning in discrete LPs levels be "unpacked" from holistic rubrics and identified by a series of analytic, conceptual bins?

Holistic scoring rubrics are generally used with construct maps in LPs to assign a single score to a student response based on the type of reasoning used (Wilson 2009). Holistic rubrics are defined in this experiment as being polytomous with each scoring-level being mutually exclusive to the others and which captures a unique and often complex set of response characteristics. Capturing complex content interactions is claimed to be easier with holistic rubrics than analytic rubrics (Tomas et al. 2019). Holistic rubrics provide a comprehensive evaluation of a response and are thought of as easy to use (Jescovitch et al. 2019b; Yune et al. 2018).

A key consideration in the application of either type of rubric is inter-rater reliability (IRR), or how closely scores assigned by multiple raters to a single response agree. Comparisons of studies employing holistic or analytic rubrics approaches to human coding are summarized by Tomas et al. (2019) and Brookhart (2018), and find there is little evidence to support one approach as universally "better" than the other for human coding. This is an important consideration for developing ML models, as the reliability of human coding in the training dataset is a key requirement for achieving high performing scoring models (Liu et al. 2016).

Since there are different approaches to human coding of student responses, it is no surprise that successful ML classification models have been developed using either approach. Previous ML studies have developed models for short, content-rich responses, using holistic rubrics and multi-level automated scoring models (Anderson et al. 2018; Prevost et al. 2016; Thomas et al. 2019). However, other studies have used analytic rubrics, focused on key conceptual ideas in a domain, to optimize model performances (Haudek et al. 2012; Moharreri et al. 2014; Sieke et al. 2019). Even though both types of coding approaches have been successfully employed to train ML models, little work has been done to directly compare these approaches for generating ML models for the same construct.

A recent review has shown that there is a wide variety of ML algorithms employed in automated scoring of text

responses (Zhai et al. 2020). Further, the study points out that technical features of ML are often not reported. One such technical feature of supervised ML is the general type(s) of classification algorithm(s), which can be either binary or multiclass (Bishop 2006). Although there are technical ways to allow different types of algorithms to function using either binary or multiclass labels (Hastie et al. 2009), some studies have looked at the accuracy of the different classification approaches using the same set of data (Balyan et al. 2018). However, there has been no work to see if one of these approaches has advantages for the training set used by ML models, in the context of science CR assessment. Often, studies have produced training sets for ML that utilize the human-assigned score consistent with the assessment rubric type. This leaves an open question of whether ML algorithms can be developed to identify the same underlying, complex construct using a coding approach that is different from what was employed by humans in traditional qualitative work. If so, are there any advantages to one coding approach over another for ML model development?

In a key study, Liu et al. (2014) deconstructed a holistic scoring rubric into a series of conceptual (analytic) rubrics for middle school science, which were then used to generate a human-coded data set. These scores for the conceptual rubrics were subsequently recombined into a holistic score. When comparing human-computer agreement, they found moderate to high Cohen's kappa over several items. However, the results provided were only at the level of the holistic, not conceptual, codes which makes understanding the limitations of these approaches difficult. This represents a remaining challenge in applying ML in CR assessments: to examine the potential benefits of different coding approaches and whether reliability of training set data from both these approaches remain high.

Research Questions

We systematically compared two coding approaches to generate training datasets for ML: (i) a holistic approach based on LP levels and (ii) a dichotomous, analytic approach of multiple concepts, or bins, of student reasoning. Specifically, we investigated developing ML models to evaluate student responses to our LP assessments in an undergraduate STEM context. Our three questions for this research study are to determine:

1. What human-human IRR trends can be achieved using analytic and holistic rubrics?
2. To what extent can we achieve agreement between holistic and composite analytic classifications of student reasoning at a large scale?
3. Which coding approach, analytic or holistic, better supports development of an accurate ML scoring model?

Study Context: Principle-Based Reasoning Using Learning Progressions as a Complex Construct

Scientific principles represent "rules" or constraints that mediate how multiple phenomena occur, such as matter and energy being conserved during chemical reactions, and are a hallmark of expert scientific thinking (AAAS 2011; NRC 2012). Consequently, principle-based reasoning supports intellectual coherence across a diversity of physiology concepts by focusing students' attention on the deep features and relationships of systems that have broad explanatory power (Goldstone and Day 2012). When students attend to these deep features and relationships, they begin to connect multiple phenomena that may appear superficially distinct but are fundamentally related (Mohan et al. 2009; Chi and VanLehn 2012; Modell 2000). For example, the principle of flux is ubiquitous across all physiological systems (Modell 2000; Michael and McFarland 2011). Flux is the rate of movement of a substance (ions, air, water, glucose, etc.), that is directly proportional to the magnitude of gradient and inversely proportional to the magnitude of resistance ($F = G/R$). Flux is captured in the equations for Ohm's Law, Fick's Law of Diffusion and Poiseuille's Law.

We have begun development of a flux LP that posits students' principle-based reasoning develops in stages. Our preliminary qualitative analysis of student interviews and written explanatory answers in this domain suggest students first focus on surface features and irrelevant relationships (i.e., level 1). Then, they begin to identify some deep features of flux (e.g., for ion flux concentration gradients, membrane potential, resistance) and how they relate in a principled way (levels 2 and 3). At the highest levels (levels 4 and 5), students consistently identify deep features and how those features relate in principle-based ways to frame their reasoning about phenomena. From this work, we see that students' first challenge in tackling physiology problems is learning to differentiate between surface features that are specific to a particular scenario (e.g., comparing irrelevant traits among species) with the deep features relevant to the problem (e.g., comparing pressure gradients among species; Schwartz and Martin 2004; Doherty et al. 2019; Scott et al. 2019). Once students differentiate surface and deep features (the transition between Levels 1 and 2), their next set of struggles focus on accounting for all the meaningful ways the deep features interact (transitions between levels 3, 4 and 5). In flux problems, this entails accounting for multiple, potentially opposing gradients (e.g., concentration and electrical gradients, osmotic and pressure gradients) and various forms of resistance (e.g., diffusion distance, number of channels on a membrane) depending on the nature of the problem. Occasionally, holistic rubrics seem additive in nature (e.g., level 5 represents two ideas from level 4 used together correctly); however, this is not an essential feature of the LP only that reasoning at a specific level requires attending to multiple features.

Methods

Constructed Response Items

As part of a larger project, we developed a set of assessment items aligned to a developing flux LP. The set of items targeted the key principle of flux in various physiological concepts (e.g., diffusion and bulk flow) and contexts (e.g., cardiovascular and respiratory systems). The set of items was designed to elicit student reasoning across all 5 proposed LP levels. For this study, we evaluated four CR assessment items (Fig. 1), each targeting five levels of the flux LP in an ion context. We enacted several criteria for the selection of the four items for the study: (1) all items shared the context under investigation (using flux to explain movement of ions into or out of cells due to concentration and/or electrical gradients); (2) we selected two items with a simpler context (movement of one ion; Nernst potential (NP1) and action potential peak (APP1)); the other two items had a more difficult context (movement of two ions; glutamate receptor (GR2) and resting membrane potential (RMP2)); (3) the cellular contexts were either generic or familiar to students in undergraduate

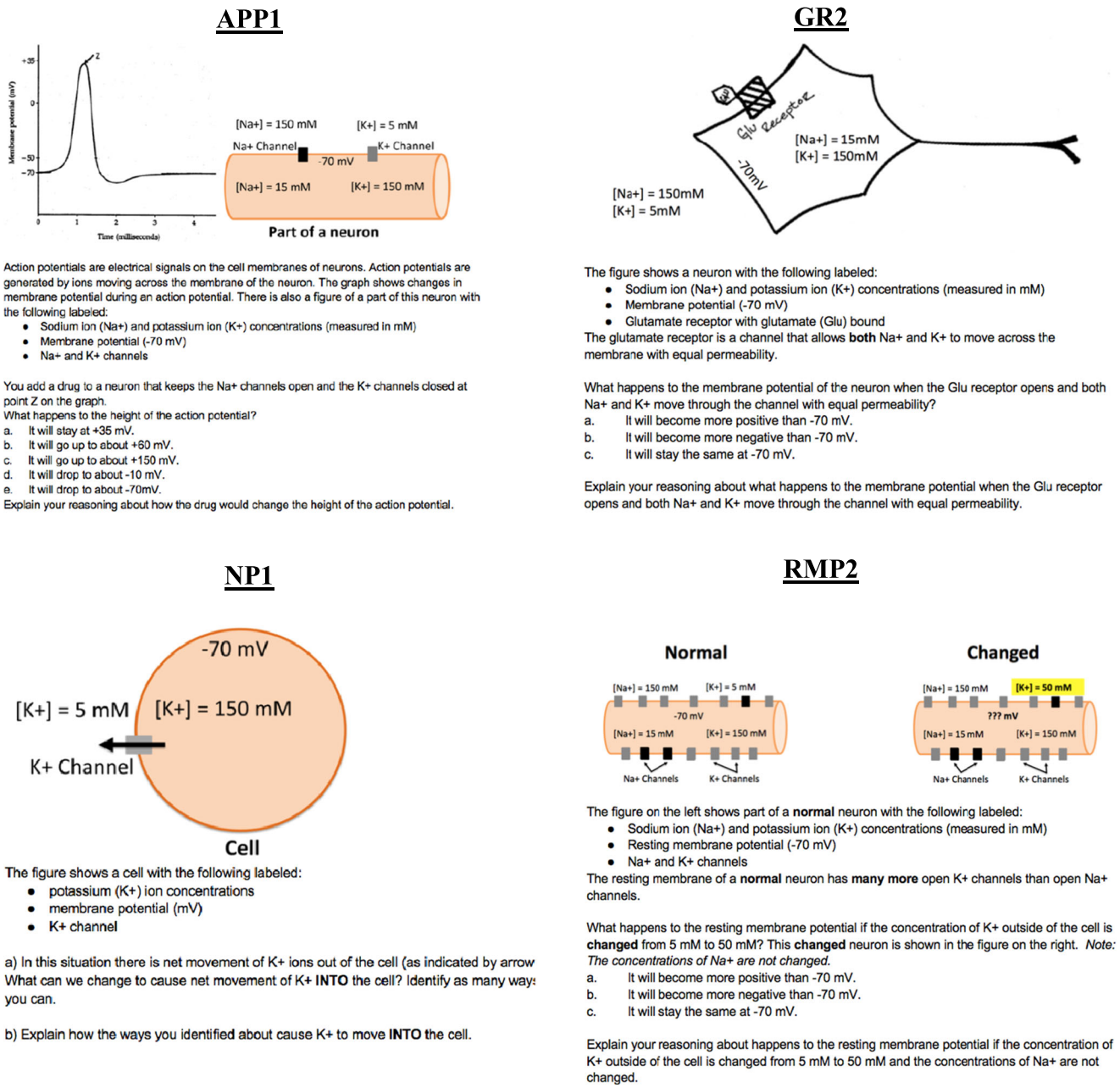

Fig. 1 Flux constructed response assessment items: action potential peak (APP1), Nernst potential (NP1), glutamate receptor (GR2), and resting membrane potential (RMP2)

physiology courses; and (4) the items had to elicit a wide range of student reasoning across all five levels of the developing LP. Each of these items was administered to undergraduate students as online homework or bonus assessments in physiology courses from several institutions in the USA. In order to get a variety of student responses, we collected responses at two public, research intensive universities and two public, community colleges. A total of ten different courses administered the items, ranging from introductory to advanced levels. For one item, NP1, some responses were also collected as part of another study for which we did not collect information about institution type or course. From the collected responses, we randomly selected 700 responses to each item in order to have an equal set of responses for rubric development, coding, and computer model development. This study was determined to be exempt by the Michigan State University Institutional Review Board because student data collected was part of normal educational testing methods and all responses used in this study were de-identified. Each item was analyzed with the workflow outlined in Fig. 2 and described in detail below. RMP2 will be used as the main example for the rest of the paper to illustrate various patterns in student thinking and challenges to the holistic and analytic rubric development process.

Rubric Development

Holistic rubrics for each item were already developed as part of the refinement procedure for the ion flux LP and aligned to the LP framework (e.g., RMP2 holistic rubric is shown in Table 1). Each item's holistic rubric was also deconstructed into a series of analytic bins. Because holistic rubrics for the flux LP are not necessarily additive, but also depend on a students' conceptual change in thinking, each item's holistic rubric underwent a deconstruction process to illustrate discrete patterns of reasoning. We followed the deconstruction process outlined in Jescovitch et al. (2019b). In summary, analytic coding rubrics were developed by careful parsing of the holistic rubric definitions into smaller, conceptual "pieces" by two experts in automated scoring and/or physiology. Each analytic bin was intended to capture a single, key concept in student responses that was relevant to the item context. Experts also proposed Boolean logic operators to recombine, or reconstruct, the analytic bins back into a single composite score, matching the student reasoning captured by the holistic rubric aligned with the flux LP. An example of RMP2 deconstruction with the alignment between holistic and analytic codes is illustrated in Table 2. For example, a student response that was coded in both analytic bins "It will become more positive than -70 mV" and "Electrical gradient is stronger than concentration gradient" would be assigned the composite score of level 5, since these concepts aligned with reasoning classified in the indicator 5.2 in the holistic rubric.

Thus, two rubrics per item (one holistic and one analytic) were developed for a total of eight rubrics. NP1 was the exception to this process, as this item already had a developed analytic rubric from previous work. Once the analytic rubrics were agreed upon by two experts, the rubrics were moved to human coding.

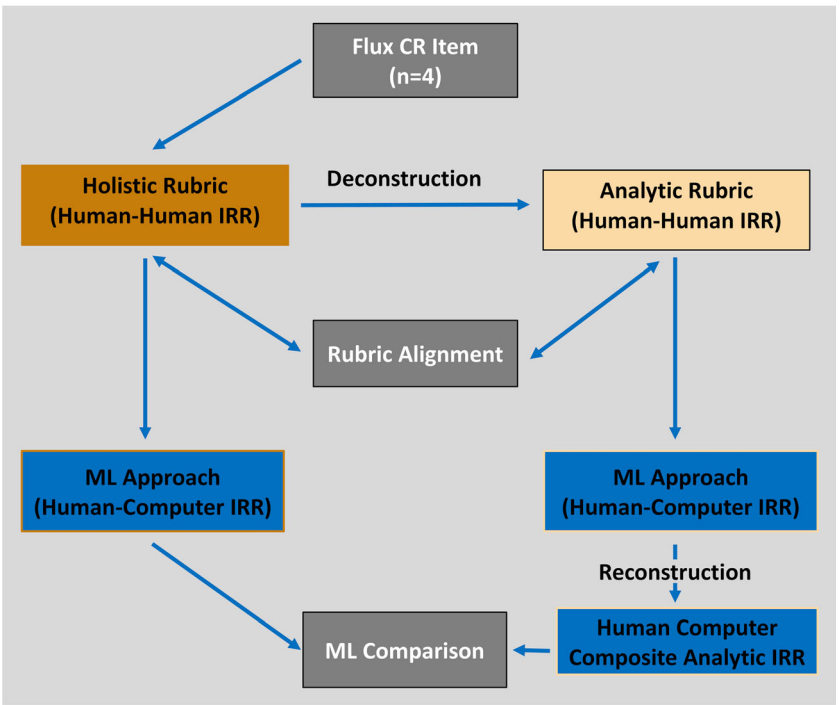

Fig. 2 Workflow of rubric experiment: student responses to a flux constructed response assessment item were used to develop a holistic rubric, the holistic rubric was deconstructed into an analytic rubric and realigned to holistic rubric, both rubrics trained machine learning models, and then analytic codes were reconstructed into composite analytic codes for model comparisons

Human Coding

Student responses were coded by two independent, expert coders using both holistic and analytic rubrics, after training on the rubrics. Briefly, coders were trained with a set of 100 responses for calibration. We used Cohen's kappa (κ) as a measure of IRR (Cohen 1960) and set a threshold of $\kappa \geq 0.8$ for calibration between human-human (HH) coders, as this represents strong agreement between scorers (McHugh 2012). If $\kappa < 0.8$ for any analytic bin or holistic rubric, discussion occurred to achieve consensus and refine rubrics. This training process of independent coding and discussion continued using additional subsets of 50 responses until $\kappa \geq 0.8$. In a few cases, where $\kappa < 0.8$ after multiple training rounds, we used 0.6 as a threshold. If a bin did not meet a $\kappa \geq 0.6$, then this bin was consensus-scored and noted as difficult to code. Frequently, these rubric bins did not have enough positive cases for coders to readily identify those concepts in agreement. Between rounds of coding, we revised the rubrics as possible to clarify coding criteria; however, since the rubrics were aligned to the developing LP, any revisions still needed to align with the underlying framework. Once calibrated, coders would move to full, independent coding until a total of 700 responses were coded. A third coder, involved in rubric development, was used to reconcile any responses with discrepant scores in the full coded data set. These final scores were used for the ML training set.

To reduce potential coder bias (Bierema et al. 2020), coders scored a subset of collected student responses ($n = 700$ per rubric; two rubrics per item) by alternating full coding between holistic and analytic rubrics, and by items (ESM 1). Responses for the same item were also randomized to appear in different order between rubrics.

Machine Learning Model Development

Once consensus coding was completed, the data were used to supervise training of ML models. Because previous research shows success in both holistic and analytic coding with different ML algorithms (Prevost et al. 2016; Thomas et al. 2019), we used the human coded data to train two ML models: (i) an 8-classification algorithm ensemble implemented in R, or the Constructed Response Classifier (CRC) (Jurka et al. 2012; Sieke et al. 2019), and (ii) a single classification algorithm implemented in LightSide Researcher's Workbench (Carnegie Mellon University's Language Technologies Institute; http://ankara.lti.cs.cmu.edu/side/). We choose these two ML platforms to test because they both rely on open-source code and thereby are potentially usable or adaptable for any researcher. Secondly, both these platforms have been successfully used to develop ML models to classify undergraduate CR in biology domains previously (see Sieke et al. 2019; Moharreri et al. 2014). During analysis for one item (APP1), we noticed duplication of some responses in the coded dataset and removed these responses to keep only unique student responses ($n = 669$) for model training.

CRC treats the task of assigning scores to student writing as a ML text classification problem (Aggarwal and Zhai 2012). Each individual student response is treated as a document and the bins in the scoring rubric are treated as classes. Text features of each document are extracted as *n-grams* and used as input variables in classification algorithms. Expert-assigned codes (i.e., labels) to each response are used as the target variable to train the classification algorithms. CRC then generates predictions on whether each given document is a member of each class. The eight algorithms used include support vector machines (Hearst et al. 1998), supervised latent dirichlet allocation (Blei and McAuliffe 2007), logitboost (Friedman et al. 2000), classification trees (Breiman et al. 1984), bagging classification trees (Hothorn and Lausen 2005), random forests (Breiman 2001), penalized generalized linear models (Friedman et al. 2010), and maximum entropy models (Kazama and Tsujii 2005). Algorithms vote independently on the categorization, or code, but those individual votes are combined to make a final categorization prediction. The predictions are combined using a stacking scheme that includes categorization designation based on highest aggregate voting, votes weighted by prediction probability, and votes weighted on algorithm performance. Thus, the predictions of the set of individual algorithms are then combined to produce a single class membership prediction for each student response and rubric bin (Large et al. 2019).

LightSide Researcher's Workbench was used to extract feature information at the holistic LP sub-level. The following individual algorithms were used in an initial round of testing: logistic regression: L2 regularization (ridge regression), L2 logistic regularization (dual), and decision tree modeling. One best performing algorithm from the initial testing was used for round iteration and modeling improvement. Both CRC and LightSide ML models used a 10-fold cross validation process for model validation (Sieke et al. 2019).

We calculated model performance metrics between human and computer predicted codes within the training set and used κ as a primary metric for evaluation. If $\kappa \geq 0.8$ for human-computer (HC) agreement, the ML model was determined as "matured" and no additional steps were taken. If HC $\kappa < 0.8$, then we explored issues in model performance to improve the model. We continued revisions until the model had either achieved $\kappa \geq 0.8$ or we completed five rounds of model revision. The number of rounds of revisions were dictated based on exploratory findings.

The most common activities in model revision included, but were not limited to using software for feature selection in models, applying different preprocessing rules to the training

Table 1 RMP2 holistic rubric indicators and exemplars across 5 learning progression levels

Learning progression level	Indicator	Exemplars
5	1. Explain that membrane potential will reflect EK, which is now − 29/more positive/in between EK and ENa, BECAUSE there are more K channels [than Na] OR RMP is between EK and ENa 2. Explain that the concentration driving force out for K is reduced and the electrical driving force in for K+ is unchanged (at first) so the net driving force for K+ is into the cell causing the membrane potential will be more positive or closer to new EK.	5.1) It will be more positive than − 70 mV. / Because there are many more K+ channels, they have a larger influence in setting the membrane potential, which in a normal neuron is close to − 70 mV because of its proximity to the K+ equilibrium potential of − 91 mV. Under the new circumstances, the K+ equilibrium potential is − 29 mV; thus, resting potential will be more positive than − 70 mV. 5.2) As the [Na+] is not changed, Na+ would not cause any change in membrane potential. An increased [K+] on the outside means a slightly decreased concentration drive for K+ to move to the outside, and as the electric drive of K+ does not change, it makes sense for the neuron to be slightly more positive.
4	1. Explain that membrane potential will reflect EK, which is now − 29/more positive/less negative/in between EK and ENa OR Explain a weaker electrical gradient will be created due to a weaker concentration gradient as the cell goes to equilibrium (use definition of equilibrium potential)	4.1) It will be more positive than − 70 mV. / EK+ for the cell on the right is − 29 mV, which is more positive than − 70 mV (RMP), so the new RMP will be more positive than − 70.
3	1. Use Ek or the definition of equilibrium potential but make mistakes in reasoning, not Ek calculation mistakes (e.g., EK is more positive causing K to move into the cell and make membrane potential more negative, EK is positive making the membrane potential positive) 2. Explain that the membrane potential will become more positive because the lower concentration gradient reduces the driving force out on K OR the rate of K flowing out (without discussion of electrical gradient or EK)	3.1) It will be more positive than − 70 mV. / Less K+ will need to exit the cell to reach equilibrium. When you calculate the equilibrium for K+ using the nernst equation, Ek+ calculates to be positive. Since the ENa+ is also positive, the resting potential of the cell will be more positive. 3.2) It will be more positive than − 70 mV. / With a higher concentration of potassium on the outside of the cell, there is less driving force for potassium to leave, hence a weaker gradient. With less potassium leaving, this will result in a more positive membrane potential since potassium has a positive charge.
2	1. Explain that the higher external K+ concentration will result in K+ moving into the cell, making the membrane potential more positive 2. Explain that outside the cell is more positive which directly impacts the charge difference across the membrane so membrane potential must be more negative/positive 3. Attempt to relate ion movement with membrane potential but make mistakes: i) K+ will flow in so the MP becomes more negative ii) K+ and Na+ have equal flow so no MP change iii) Still a concentration gradient so K+ leaves and MP becomes more negative 4. Explain the RMP will stay the same because still concentration gradients acting on K+ and Na+ in a similar way as before so no MP change OR the number of channels has not changed	2.1) It will be more positive than − 70 mV. / A higher concentration of K+ outside the cell will lead to more ions entering the cell so the membrane potential will increase. 2.2) It will be more negative than − 70 mV. / Since there is a higher concentration of K+ outside of the cell the electric potential difference is greater causing a more negative resting membrane potential. 2.3) It will be more negative than − 70 mV. / The concentration gradient will cause K+ ions to leave the neuron and since K+ is positive, the neuron will have less positive ions and become more negative. 2.4) It will stay the same at − 70 mV. / The RMP is regulated by the proportion of open K+ channels to open Na+ channels. Since the number of channels did not change, the RMP should stay the same.
1	1. Explain that RMP will stay the same because it is a set point OR that ions will move to achieve equilibrium OR make a general or irrelevant statement about the system or describe an irrelevant (e.g., cell "wants") OR incorrect (K+ moves from low to high gradient) process.	1.1) It will stay the same at − 70 mV. / I do not you can change the resting membrane potential, I think it is constant in every system. 1.1) It will stay the same at − 70 mV. / If the outside of the cell is more positive, then the K+ ions will diffuse in order to create an electric equilibrium and nothing will change

Each rubric had 5 levels, each level representing a major pattern of student reasoning. Each learning progression level was further divided into sublevels called indicators. Indicators represented common ways students reasoned in a particular level. Coders could use indicators to help during coding, although all results reported here are at the higher 5-level classification. Exemplars are actual student responses coded for each indicator

set (e.g., synonyms), using different ensembling methods, finding patterns in discrepancies, scoring additional data (by initial coders) to append to the training set, creating dummy responses, and using software tools to identify important linguistic features in responses.

Comparison of Holistic and Analytic Approaches

Both holistic and composite analytic scores are reported using 5-level LP designations. Because analytic codes are binary, these codes had to be combined using the validated Boolean

logic into the 5-level LP designations, which we refer to as the composite analytic score, in order to compare holistic scores generated via holistic and analytic coding. To check alignment of the two different coding rubrics for each item, the HH holistic codes and HH composite analytic codes for each response were compared for validation and Cohen's kappa was calculated. If the overall alignment showed $\kappa < 0.8$, discrepancies between human-assigned codes in the two approaches were evaluated for human mis-scores, acceptable changes to Boolean logic in the composite score or definitions for either rubric in order to achieve $\kappa \geq 0.8$. To evaluate our research questions 1 and 3, we examined IRR among coders (both HH and HC IRR) on both analytic and holistic rubrics. We then examined HH and HC agreement and trends per item and across items. For our research question 2, we report on Boolean logic, validation, the number of discrepancies during HH analytic and HH holistic alignments, an explanation of those discrepancies, and the amount of time during each phase of coding and rounds of deconstruction.

Results

Human coding agreement and ML model performance results varied across items and rubric type. We report on human coding, holistic and composite analytic alignment, and ML performance and then summarize our findings. All results are reported as a 5-level LP code unless otherwise noted.

Human-Human Inter-rater Reliability

Our first research question was to determine what HH IRR trends can be achieved using analytic and holistic rubrics. All items elicited acceptable HH IRRs, as measured by Cohen's kappa (κ), but these trends varied across assessment items and rubric approach (Table 3). Higher IRR between human coders was achieved using the holistic rubric than for the corresponding composite analytic code for APP1 and GR2. However, higher HH IRR was achieved using an analytic approach to calculate a composite holistic value for NP1 and RMP2. The distribution of responses in each LP level for RMP2, as determined by human-human codes, is supplied in ESM 2.

Our second research question was to determine if student reasoning identified in holistic rubrics could be reliably identified and computed from analytic rubrics with high agreement at a large scale across contexts. We extended our previous research methods (Jescovitch et al. 2019b) to monitor agreement between holistic and composite analytic codes using Cohen's kappa to four CR items. We achieved near-perfect agreement between codes assigned in both approaches for all four items, after rubric and/or Boolean logic revision (see last column of Table 3). During revision, we often found that one or two analytic bins had to be broadened to focus more on a concept than a "keyword" (also see below). These results allowed us to conclude that human codes assigned to the ML training set were reliable and that both approaches were aligned to the construct under investigation, or LP framework.

Analytic Human-Human Inter-rater Reliability Unpacked

To continue our investigation into research question 2, we wanted to investigate if coders achieved similar levels of IRR for all concepts. Therefore, we evaluated each item's analytic rubrics in terms of number of analytic bins, mean and individual bins' HH Cohen's kappa (with standard error), and the number of positive cases (n) that humans agreed are present in that bin when humans coded independently (Table 4). The number of analytic bins necessary to capture the reasoning in the four items ranged from 8 to 16 bins. Unsurprisingly, the items that were in a context of 2 ions moving required more analytic bins than the items focused on 1 ion moving.

Mean κ for HH IRR of each item's bins was generally strong, with APP1 being slightly below our target threshold. However, the range of κ varied greatly across the bins within an item. Each item had at least one analytic bin that did not meet our target of $\kappa \geq 0.8$.

One of the issues we encountered with analytic coding was coders adopting too narrow a definition for a given analytic bin; this was rarely observed in holistic coding. For example, in RMP2, bin 5, coders achieved very good agreement $\kappa = 0.77$ with 131 positive cases (i.e., ~19% of the data set). However, during revision and alignment of rubrics, it became apparent the scorers had coded for the exact bin concept "K+ moves because of a concentration gradient" but not the concept of "higher external K+ concentration will result in K+ moving into the cell," which was also included in the holistic rubric and coding. Thus, coders were narrowing in on the specific language rather than identifying the underlying concept.

Human-Computer Inter-rater Reliability

Having established the reliability of the human codes assigned by both holistic and analytic approaches in our training set, we used these data to develop ML models to investigate our third research question. We found minimal difference between the two ML approaches (LightSide and CRC) we employed across the four items (Table 5).

Overall, we were able to generate ML models using CRC with acceptable levels of performance using either holistic or composite analytic approaches (Table 6). Kappas for models trained with holistic coded data (HC IRR) range from 0.603 to 0.696 and κ for composite analytic models (HC IRR) range from 0.649 to 0.737. All models achieved greater than 70%

Table 2 RMP2 final deconstruction of holistic rubric to analytic components (figure based on Jescovitch et al. 2019b). An "X" represents that a concept is essential for a response to be scored at a particular indicator

LP	It will become more positive than − 70 mV	It will become more negative than −70 mV	RMP is a set point/ fixed value	K+ moves because of a concen. gradient	K+ moves out of the cell	K+ moves into the cell	Increased concentration/ amount/ positives outside the cell changes the charge difference	K+ and Na+ move in opposite directions	Stay the same because gradient did not change directions or concen. change has little to no effect	Mistakes on the calculation or applying of EK+ or Nernst equation	Electrical gradient is stronger than concentration gradient	Decreased concen. gradient OR decreased K+ moving out	Decreased electrical gradient on K+	MP is in between EK and ENa OR reflect EK which is now − 29 OR mention equilibrium potential for EK.	There are more K+ channels (than Na+ channels)	Concen. gradient is stronger than the electrical gradient
5.2	X										X					
5.1	X													X	X	
4.1												X		X		
4.1	X											X	X			
4.1	X													X		
3.2												X				
3.1										X						
3.1											X					
3.1																X
3.1		X												X		
2.4									X							
2.3		X		X												
2.3		X			X											
2.3		X				X										
2.3								X								
2.2							X									
2.1				X		X										
1.2																

Table 3 Human-human inter-rater reliability ($\kappa \pm$ standard error) across the 5-level flux learning progression

Item	Holistic	Composite analytic	Human holistic-human composite analytic
APP1	0.872 ± 0.015	0.844 ± 0.016	0.964 ± 0.008
NP1	0.812 ± 0.018	0.892 ± 0.014	0.862 ± 0.016
GR2	0.847 ± 0.016	0.780 ± 0.018	0.910 ± 0.013
RMP2	0.758 ± 0.019	0.880 ± 0.014	0.817 ± 0.017

accuracy, with the best performing model over 80% accurate. Each item's HC IRR matrices for both approaches are provided in ESM 3.

Examining the confusion matrices shows the CRC model performance was less accurate on levels with fewer number of responses, for example, level 5 on APP1 and NP1 (panels A and B; ESM 3). For all 4 items, model accuracy was slightly lower for one of the higher levels (4 or 5) in the LP, suggesting that more complex reasoning in these levels may be more difficult for ML to detect in our training set. We note for levels that had lower accuracy due to one of the issues above that models using the composite analytic scores were generally slightly better at predicting responses at these levels than models trained using the holistic scores (e.g., level 4, panel C; ESM 3).

Similarly, we were able to use LightSide to produce ML models with mostly acceptable model performance. Kappas for models trained with holistic coded data (HC IRR) range from 0.586 to 0.693 and κ for composite analytic models (HC IRR) range from 0.664 to 0.685. All models achieved greater than 69% accuracy. We note that two of the LightSide developed models did not perform quite as well as CRC using the composite analytic scores. Further, one holistic model with LightSide had a Cohen's kappa < 0.6, although this item, GR2, had the lowest performing models regardless of coding or ML approach. This suggests that the student reasoning elicited by this item is difficult to capture by ML techniques. To simplify the rest of the results, we will focus on reporting outcomes from using CRC for building ML models.

To examine whether a given coding approach led to better performing ML models, we examined performance for each approach on each of four items. There was substantial improvement in analytic composite scores for GR2 and RMP2 compared to the holistic models for these items using CRC (Table 6). Models for NP1 had only very slight differences in κ and accuracy between approaches; while models for APP1 showed some small gain in performance using the analytic composite score, it is within one standard error. Interestingly, GR2 had better agreement between human coders (HH IRR; Table 3) holistically, but the HC IRR shows better agreement via the composite analytic model approach (Table 6). RMP2 has substantially better agreement for HH and HC analytic rubrics as compared to holistic codes. Overall, HH IRR across items was higher than the HC IRR for the same items, regardless of coding approach. This is typical for supervised ML approaches as the accuracy of the HH IRR sets an upper limit for the ML based on those data.

To further investigate the performances of the ML models, we leveraged the ordinal nature of the LP levels to calculate quadratic weighted kappas (QWK), which is one way to account for the distance between levels of mis-scores (Fleiss and Cohen 1973). The QWKs were calculated for HC agreement on each item and coding approach using CRC; these metrics were greater than their unweighted κ counterparts. Williamson et al. (2012) suggested a minimum threshold for QWK of 0.7 for evaluating automated scoring performance. All four of the models built using analytic composite scores exceeded this threshold, while two of the models using

Table 4 Analytic human-human inter-rater reliability (κ) and positive cases (n) of a response by analytic bin

Item	Number of analytic bins	Analytic bin mean ± SD		Analytic bins															
				1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
APP1 ($n = 700$)	8	0.763 ± 0.196	κ	0.50	0.48	0.97	0.86	0.65	0.94	0.79	0.92								
			n	25	82	151	334	60	223	24	88								
NP1 ($n = 700$)	9	0.883 ± 0.092	κ	0.70	0.87	0.92	0.77	0.92	0.91	0.95	0.94	0.98							
			n	9	38	82	157	37	511	169	388	98							
GR2 ($n = 700$)	13	0.857 ± 0.099	κ	0.59	0.83	0.81	0.86	0.88	0.80	0.80	0.91	0.88	0.90	0.95	0.95	0.97			
			n	25	346	159	59	77	203	111	96	62	146	118	354	226			
RMP2 ($n = 700$)	16	0.856 ± 0.085	κ	0.75	0.71	0.75	0.86	0.77	0.80	0.88	0.86	0.80	0.87	0.89	0.96	0.94	0.95	0.97	0.96
			n	4	7	35	4	131	79	213	98	6	35	32	309	180	30	106	537

Table 5 Mean human-computer inter-rater reliability and standard errors, represented by holistic learning progression levels 1–5 matrices, compared by ML tool

	Holistic (Cohen's kappa)		Composite analytic (Cohen's kappa)	
Item	CRC	LightSide	CRC	LightSide
APP1	0.696 ± 0.022	0.693 ± 0.019	0.737 ± 0.021	0.685 ± 0.022
NP1	0.656 ± 0.023	0.681 ± 0.022	0.658 ± 0.023	0.664 ± 0.022
GR2	0.606 ± 0.023	0.586 ± 0.023	0.649 ± 0.023	0.667 ± 0.021
RMP2	0.603 ± 0.023	0.614 ± 0.023	0.720 ± 0.020	0.669 ± 0.021

holistic scores exceeded this target. The two models that do not meet this target are for items in the more complex scenario of two ion species moving. We also calculated the QWK for HH IRRs for each item for each coding method (ESM 4). The degradation of means between QWK for HH and HC holistic scores ranged from − 0.120 to − 0.197 with an average of − 0.145 while the degradation between QWK for HH and HC for the composite analytic scores ranged from − 0.088 to − 0.164 with an average of − 0.118. Although the degradation for a number of models was above the suggested threshold (Williamson et al. 2012), we note that the HH IRR as measured by QWKs for both coding methods were extremely high (0.818–0.920). Therefore, the observed degradation may be due to robust human agreement during scoring as opposed to poor model performance, as nearly all models met the minimum QWK threshold. The most interesting finding is that while NP1 was the only item to have a larger degradation (− 0.164) for composite analytic than holistic, the QWK for the final HC model performed slightly better as the composite analytic than the holistic counterpart.

Analytic Human-Computer Inter-rater Reliability Unpacked

To continue our investigation into research question 3, we unpacked the analytic rubrics to determine if critical components of student reasoning were being predicted accurately by the ML models. We examined each item's individual analytic bins by calculating mean HC κ (with standard error) across all bins, HC κ for each individual bin, number of positive cases (n)

Table 6 Human-computer inter-rater reliability (κ ± standard error) using CRC across the 5-level flux learning progression

	Holistic		Composite analytic	
Item	Cohen's kappa	Accuracy (%)	Cohen's kappa	Accuracy (%)
APP1	0.696 ± 0.022	77.7	0.737 ± 0.021	80.7
NP1	0.656 ± 0.023	75.4	0.658 ± 0.023	75.3
GR2	0.606 ± 0.023	70.6	0.649 ± 0.023	74.3
RMP2	0.603 ± 0.023	70.3	0.720 ± 0.020	78.4

that the computer predicted correctly, and precision (percentage of true predicted positives out of all predicted positive) of cases in that bin (Table 7). Higher precision values indicate that the model is good at detecting only true positive cases without identifying false positive cases. Overall, analytic mean κs were lower and more variable than composite analytic κ (range 0.574–0.726). However, these average measures include bins that the model was completely unable to predict, or $\kappa = 0$. For example, RMP2 had four bins with $\kappa = 0$, and if these were removed, the mean would increase from 0.574 to 0.766.

Models for analytic bins trained with low frequency positive cases but with less lexical diversity, or fewer unique words used, performed better than models trained with higher frequency positive cases but exhibiting very diverse language. Tomas et al. (2019) identified CR assessments as *divergent*, or exhibiting rich lexical diversity, in that criteria can be met by a broad range of words and phrases within student individual responses. For example, in RMP2, bin 11, or "there are more K+ channels (than Na+ channels)," predicted 21 positive cases (human identified 32 positive cases) with a HC κ = 0.79. However, bin 6, or "increased concentration/amount/positives outside the cell changes the charge difference," predicted 89 positive cases (humans identified 79 positive cases) but only achieved a HC $\kappa = 0.52$. Thus, even though bin 6 had more positive examples, this bin also had more lexical diversity than bin 11 because of the increased ways that students can write about that concept. Specifically, students in bin 11 reasoned fairly consistently with "more K+ channels" appearing in their response. For example, a student wrote, "Plugging it into the nernst equation with the new outside K+ concentration, the Ek is -29. And since there are many more K+ channels open, the RMP will be more closer to Ek, so RMP should be more positive."

However, student responses classified in bin 6 varied greatly in language, using terms such as "amount," "concentration," or "positives" as these words would indicate the same conceptual idea in this context (i.e., more of a positively charged ion). For example, one student wrote, "the resting membrane potential will become more negative, because the charges on the outside of the cell are now more positive than the charges in the inside of the cell." While another student wrote, "If the extracellular K+ concentration increases, then the concentration gradient of K+ across the membrane decreases, so there is less of a

Table 7 Analytic human-computer inter-rater reliability (κ), positive cases (n), and precision of positive cases (%) in responses using CRC by analytic bin

Item	Number of analytic bins	Analytic bin mean ± SD		Analytic bins															
				1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
APP1 ($n = 669$)	8	0.649 ± 0.317	κ	0	0.44	0.54	0.70	0.81	0.86	0.88	0.97								
			n	0	200	129	313	67	196	32	93								
			%	0	36.0	76.0	86.6	76.1	94.4	81.3	94.6								
NP1 ($n = 700$)	9	0.717 ± 0.086	κ	0.58	0.65	0.65	0.67	0.68	0.77	0.80	0.81	0.83							
			n	8	74	44	116	68	530	148	378	79							
			%	62.5	51.4	97.7	86.2	54.4	92.5	90.5	92.6	94.9							
GR2 ($n = 700$)	13	0.726 ± 0.289	κ	0.07	0.14	0.71	0.75	0.76	0.78	0.80	0.83	0.84	0.85	0.95	0.97	0.99			
			n	1	3	115	53	78	188	118	78	44	124	114	363	227			
			%	100	100	83.5	88.7	88.5	87.8	93.2	91.0	95.5	96.0	97.4	97.2	99.1			
RMP2 ($n = 700$)	16	0.574 ± 0.379	κ	0	0	0	0	0.33	0.52	0.70	0.73	0.77	0.79	0.79	0.80	0.89	0.93	0.97	0.99
			n	0	0	0	0	378	89	190	154	32	54	21	289	167	30	108	536
			%	0	0	0	0	34.9	83.2	84.2	63.6	81.3	68.5	100.0	90.3	95.2	93.3	96.3	99.8

difference between the charges on both sides. A smaller difference in charges means a more positive resting membrane potential." Thus, both these responses were classified into bin 6, but used very different text in their responses.

Not all models for analytic bins showed similar trends in performance. For example, some bins had poor precision with corresponding poor κ (e.g., bin 2 in NP1; Table 7). A few bins showed acceptable κ, but still exhibited poor precision (e.g., bin 10 in RMP2), which may lead to specific types of errors in computer prediction for composite scores dependent on that bin. A good number of bins that show these poor model performances are due to low frequency of responses in these bins. One concern of using an analytic composite score is that bins with poor performance may be critical in calculating the composite or LP holistic score. For example, in RMP2, to be a level 5, a student must reason with ideas captured in either bin 14 or 11 in addition to bin 13. Even though we can meet the threshold HC $\kappa > 0.79$ for these bins individually, humans agreed that only 4% and 3% of responses included bin 14 or 11 reasoning, respectively. Although these concepts occur infrequently in student responses, they are critical to determining the holistic level 5 code.

On the contrary, there are also analytic bins that are not as critical to overall model performance. Except for NP1, at least 1 analytic bin in each item has a HC $\kappa < 0.2$ (Table 7). Bin 3 in RMP2 had 35 positive cases by human coding, but this bin captured students who reasoned with "mistakes." Thus, the different ways students made mistakes added a range of lexical diversity that the ML model could not accurately identify. Responses that included mistakes should have been at a lower level of the LP, but since the ML model could not identify this concept, it generally over-predicted the composite score for these responses.

Composite analytic κ (Table 6) may be higher than analytic mean κ values (Table 7) because individual bin predictions are reconstructed using Boolean logic to determine LP level rather than a mean average of all bins. Specifically, not all bins with $\kappa = 0$ have the same importance in distinguishing between levels and from item to item as discussed above. Because RMP2 has high performing models for bins that hold greater weight in determining higher LP levels (bins 11–15), the composite analytic κ (0.720) is substantially greater than the item's HC holistic κ (0.603).

Finally, we also looked at another important variable when considering coding approaches: time. We tracked the time required by humans for deconstruction, calibration coding, and full coding (ESM 5). Given that the holistic rubric was already developed and validated, the deconstruction process required approximately 7–9 additional hours per expert per item. Holistic coding required 5.5–8.5 h per expert. Analytic coding required approximately 7.5–13 h per expert to code the entire dataset. Analytic coding required from 4 to 18 more minutes per 50 responses than holistic human coding. It is interesting to note that increased number of analytic rubric bins and complexity of the rubric and/or item led to an increase in the amount of time to code independent from the rubric approach.

Summary

Many factors influence an accurate computer predictive scoring model. Specifically, HH IRR's are generally higher than HC IRR, which may indicate that human coders can more easily identify complex and/or naive reasoning and may set a maximum agreement level for computer predictions (since

models were trained using human consensus scores). Either coding approach, holistic or a series of analytic bins, can achieve acceptable levels of agreement by both human coders and ML models. As well, both ML platforms were able to achieve good performance on items, with only a single LightSide model failing to achieve a HC κ of at least 0.6. Model performance is also influenced by the deconstruction of holistic rubrics to analytic, the Boolean logic used to reconstruct the composite score and item context.

Discussion and Conclusions

We investigated the use of ML models to automate the scoring of CR assessments designed to elicit complex reasoning aligned to a physiology LP for undergraduate students. We compared analytic and holistic coding approaches on ML model performance for short, content-rich responses.

For our first research question, comparing human-human (HH) IRR on analytic and holistic rubrics, we found that either holistic or analytic coding approaches can be used to obtain high HH agreement (> 0.75 Cohen's kappa). We were able to achieve high HH agreement using either holistic or analytic coding through careful coder calibration and rubric deconstruction/reconstruction processes (Jescovitch et al. 2019b; Liu et al. 2014). However, there was a range of variability of HH IRR across the many analytic bins, both within a given item and across items. We found that low HH agreement in analytic rubrics was common in bins with low frequency of positive occurrences (< 5%) or that targeted ideas that exhibited high lexical diversity. These situations may make it more difficult for coders to properly calibrate using examples during the coder training phase, and therefore more likely to deviate from one another during independent coding. We suggest that if a bin is < 5% of the total training set, then researchers should accept that, even though it is an interesting concept that can be identified using human coding, the bin will most likely not be predicted accurately by ML models and thus have limited utility in automated scoring systems. Alternatively, if the bin is considered essential for the purpose of the assessment, a very large training set may need to be assembled to overcome this problem.

For our second research question about whether student reasoning identified in holistic rubrics can be reliably quantified by analytic rubrics, our findings suggest that either holistic or analytic coding approaches can be used to make sufficient ML models for complex CR science assessments. We were able to achieve good agreement between assigned scores to student responses using holistic and analytic type rubrics, with Cohen's kappa ranging from 0.817 to 0.964, which is considered near-perfect agreement. This suggests that complex reasoning in holistic levels of a LP can be identified by a series of analytic rubrics and furthers our ability to consider various coding methods for short content-based responses (Brew and Leacock 2013). Because we independently applied both holistic and analytic coding to the same set of responses, we were able to compare the performance of the resulting ML models, as well as identify the benefits and limitations of these approaches. We found that for some items there was little difference in the performance of the ML models created with these different approaches. For two items, the ML models trained with analytic coded responses and used to generate a composite score achieved slightly better performance. Both of these items used the slightly more complex item scenario of the movement of two ion species. It may be that analytic coding is beneficial to unpacking this additional complexity. This is aligned with previous work that showed differential model performance over seven open-ended science items, with the poor performing model being associated with the least-constrained (i.e., most open-ended) item (Butcher and Jordan 2010).

Finally, for our third research question, which coding approach better supports ML scoring models, our ML models trained using analytic composite scores performed similar to, or better, than holistic ML models of the same item. We also found that unpacking of analytic bins can help determine if the item elicits critical components of the holistic rubric and if the item elicits all 5 levels of the LP. This is similar to a finding reported by Tomas et al. (2019) that having humans apply both holistic and analytic codes helped define some of the criteria underlying the holistic scoring. We found that we were able to achieve good performing models with both ML platforms tested, although there is some evidence that CRC performed better than LightSide on some items using analytic composite scores, this was not true for all items or true for using holistic scores. This is likely due to the ensembling method employed by CRC, as ensembling has been shown to reduce error in ML predictions (Ali and Pazzani 1996). We also found that the CRC predicted composite analytic scores were slightly more accurate than holistic scores for some infrequent and higher LP levels. We also modified our criteria for ML model evaluation over the course of this study. A threshold of $\kappa \geq 0.8$ indicates almost perfect agreement (Landis and Koch 1977; McHugh 2012). This threshold was easier to achieve in our binary, analytic rubrics than our 5-level, holistic rubric. In practice, we adopted a threshold of $\kappa \geq 0.7$ for HC IRR due to the compounding of errors of multiple analytic bins being combined within 5 levels. We also monitored model performance with QWKs, which use different weights for the degree of disagreements. We used a target threshold value of 0.7 for QWK (Williamson et al. 2012), which we were able to achieve on all 4 composite analytic models and 2 holistic models. Although degradation measures for most models were higher than the suggested cutoff, this may be due in part to very high levels of HH agreement during coding. Additionally, the models built using composite

analytic scores tended to show less degradation than the holistic models.

This work helps address the challenges of using ML in science assessment. We have attempted to advance the application of ML to evaluate CR of a complex construct (Zhai 2019): principle-based reasoning in the content-rich domain of undergraduate physiology. The responses collected in this project are rich in science content, symbols, and relationships among components (e.g., differential movement of two ions), which can be challenges to developing accurate ML models using limited training sets, as is often the case in science assessments. Therefore, correct classification relies on our underlying conceptual framework to guide the ML identification of specific ideas and connections that we were interested in finding (Nehm 2019). The ML analytic approach was able to identify the most important and frequent ideas, as has been found before (Sieke et al. 2019; Moharreri et al. 2014). We have extended this effort to combine these analytic codes into a single composite code which aligns to a holistic code. Previous efforts have found this process challenging (Liu et al. 2014), some of which we have addressed by carefully monitoring the deconstruction and reconstruction process to make sure the composite score aligns with the holistic interpretation of the response.

Additionally, we have attempted to apply ML to a complex construct of a LP (Zhai 2019). LPs represent a model of cognitive development and may have various and complex underlying construct structures (Wilson 2009). Additionally, the complexity of undergraduate reasoning itself increases over the LP, so the complexity of student responses is greater at higher levels. This can pose challenges for ML scoring since holistic levels are ordinal, not interval (i.e., small changes in a response may indicate significant advance in reasoning), and some of the language used to reason at different levels may be quite similar. Therefore, the defining feature(s) which separate(s) levels may be difficult for ML algorithms to identify. Additionally, the greater number of possible levels may deflate overall performance measures. We found that the analytic components predicted by ML could be combined into a composite score that matches the holistic code and provide the same, or slightly increased, accuracy over holistic classification ML models. This is despite the fact that some individual ML analytic bins performed quite poorly. Thus, even a complex construct may be deconstructed into finer-grained ideas in order to identify critical response features; then use these features to improve overall model performance. We think this may be useful for future efforts to develop ML models for future LP aligned assessments or other complex constructs. However, this increase in model accuracy was not uniformly achieved across all items and took significant effort and time to achieve, mainly in rubric design and additional coding time.

Harsch and Martin (2013) and Tomas et al. (2019) argue that holistic and analytic approaches can be combined to offer productive outcomes for both research and practice. One possible extension of this is to investigate the combination of both holistic and analytic coding to generate one improved ML model. For example, for other items in our broader project that only target up to level 3 in the LP, we were able to successfully develop ML models using holistic codes. The additive nature of the more complex levels (4 and 5) in the items under study here may be improved by holistically predicting levels 1–3, then subsequently predicting analytic concepts that determine levels 4 and 5.

One advantage of using ML for automated scoring systems in place of programmed scoring systems is that researchers do not need to perform any additional text parsing or keyword identification for the ML procedure. The text from student responses is automatically parsed into *n-grams* in our study, then used as features (input variables) in the ML algorithms to predict expert codes (i.e., holistic or analytic codes). This is similar to previous approaches and findings in this area (Liu et al. 2016; Nehm et al. 2012). The effort exerted by humans is in the coding of the data set via rubrics to make the training data. For our study, since we began with a holistic coding rubric aligned to the levels of the LP, we spent additional time developing and applying the set of analytic coding rubrics. Because the analytic bins are meant to identify conceptual pieces relevant to the specific item and levels, we developed a procedure to guide the process of creating these rubrics (Doherty et al. 2019; Haudek et al. 2019; Jescovitch et al. 2019a, 2019b; Scott et al. 2019; Sripathi et al. 2019). Therefore, the analytic coding rubrics in our study were unique to each item, although other reports have found such conceptual rubrics may hold across items in similar contexts (Moharreri et al. 2014; Weston et al. 2015). We suggest that researchers starting new projects in automated scoring carefully consider choosing a coding approach that meets the purpose of the assessment, the intended use of the results and possible development iterations of the assessment, since our results suggest ML models can be built to classify responses with similar accuracy to either type of rubric.

There are a few limitations to our approach for this study. HH holistic and HH composite analytic codes were not initially aligned, or representative of each other, for three of the four items. We suggest that complete Boolean logic validation should be included during deconstruction/rubric development rather than during or after full coding of responses. Tomas et al. (2019) also suggest reducing irrelevant coding criteria when converting holistic criteria to analytic. Similarly, Liu et al. (2014) and we (Jescovitch et al. 2019b) have reported that during analytic rubric development for ML, often more bins are created initially than are necessary to identify the construct of interest. The interaction of using the holistic rubric to first deconstruct to an analytic rubric would

be beneficial to intended outcomes if the holistic rubric was also revised to reflect frequency or concept importance. Our broader project has a goal of developing a LP for flux reasoning in undergraduate physiology. Thus, we caution against generalizing our findings too far from the study context, although we are hopeful some of these approaches and overarching lessons can inform our rubrics as we evaluate constructs across various contexts.

Our findings can inform rubric development, coding, and ML approaches best suited to score short, content-rich CR assessments. This work is significant in that it advances our ability to use computer-automated scoring models to assess complex constructs, such as student reasoning in rich science contexts aligned with a LP. We have explored another way researchers may pursue automated scoring of more complex science constructs, by creating composite scores of smaller analytical components. Implementing such a scoring system at scale, however, requires a scoring system which can compute and represent such composite scores, while retaining the ability to provide outputs and/or feedback at multiple score levels to the intended user.

The methods explored here could help improve large-scale ML assessment and rubric development targeted at uncovering complex scientific reasoning. An important consideration in any assessment design is the potential use and interpretation of the resulting scores (Pellegrino et al. 2016). In this study, we have shown it is possible to develop ML models to predict dichotomous analytic scores as well as holistic scores for the same underlying construct. This represents an opportunity for assessment designers to consider scoring rubrics for both automated scoring and potential instructor use during item development and thus address the pedagogical and validity perspectives of ML-based assessments simultaneously (Zhai et al. 2020). From a practitioner view, it may be advantageous to have student responses automatically scored in one way (e.g., student reasoning pattern) over another (e.g., key disciplinary concepts) for specific situations. Finally, a key promise of any automated scoring system is the ability to assess large numbers of students at scale. Developing these automated scoring tools for content-rich responses provides a mechanism to evaluate student reasoning in science across courses and institutions and leads to many interesting research questions about reasoning across disciplines.

Acknowledgments We thank the members of the Automated Analysis of Constructed Response (AACR) research group, especially Juli Uhl, Kamali Sripathi, Michael Fleming, and Sam Houchlei, for their thoughtful comments regarding challenges we encountered in this project. We thank Matthew Steele and Marisol Mercado Santiago for the development and refinement of the Constructed Response Classifier. We also thank Elena Kolpikova, Rachael Cumberland, Mary Diamond, and Mallory Jackson for their contributions. For more information about the automated scoring tools investigated here, visit: beyondmultiplechoice.org.

Funding This work was funded through grants from the National Science Foundation (NSF; DUE 1660643 and 1661263).

Compliance with Ethical Standards

Conflict of Interest The authors declare that they have no conflict of interest.

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 (Michigan State University IRB x17-196e and University of Washington IRB STUDY00001316) and with the 1964 Helsinki declaration and its later amendments or comparable ethical standards.

Informed Consent Not applicable—exempt IRB.

References

Aggarwal, C. C., & Zhai, C. (2012). A survey of text classification algorithms. In C. Aggarwal & C. Zhai (Eds.), *Mining text data*. Springer.

Ali, K. M., & Pazzani, M. J. (1996). Error reduction through learning multiple descriptions. *Mach Learn, 24*(3), 173–202.

Allen, D., & Tanner, K. (2006). Rubrics: tools for making learning goals and evaluation criteria explicit for both teachers and learners. *CBE Life Sciences Education, 5*(3), 197–203. https://doi.org/10.1187/cbe.06-06-0168.

American Association for the Advancement of Science, AAAS. (2011). Vision and change in undergraduate biology education: a call to action. Washington, DC.

Anderson, C. W., de los Santos, E. X., Bodbyl, S., Covitt, B. A., Edwards, K. D., Hancock II, J. B., Lin, Q., Thomas, C. M., Penuel, W. R., & Welch, M. M. (2018). Designing educational systems to support enactment of the next generation science standards. *J Res Sci Teach, 55*(7), 1026–1052. https://doi.org/10.1002/tea.21484.

Balyan, R., McCarthy, K. S., & McNamara, D. S. (2018, May). Comparing machine learning classification approaches for predicting expository text difficulty. Paper presented at the International Florida Artificial Intelligence Research Society Conference, Melbourne, FL.

Bierema, A., Hoskinson, A.-M., Moscarella, R., Lyford, A., Haudek, K., Merrill, J., & Urban-Lurain, M. (2020). Quantifying cognitive bias in educational researchers. International Journal of Research & Method in Education. https://doi.org/10.1080/1743727X.2020.1804541.

Bishop, C. M. (2006). *Pattern recognition and machine learning*. Springer.

Blei, D. M., & McAuliffe, J. D. (2007). Supervised topic models. In Proceedings of the 20th International Conference on Neural Information Processing Systems (NIPS'07), J. C. Platt, D. Koller, Y. Singer, & S. T. Roweis (Eds.). Curran Associates Inc., USA, 121–128.

Breiman, L. (2001). Random forests. *Mach Learn, 45*(5), 5–32. https://doi.org/10.1023/A:1010933404324.

Breiman, L., Friedman, J., Stone, C. J., & Olshen, R. A. (1984). *Classification and regression trees*. Taylor & Francis.

Brew, C., & Leacock, C. (2013). Automated short answer scoring. In M. D. Shermis & J. Burstein (Eds.), *Handbook of automated essay evaluation*. Routledge.

Brookhart, S. M. (2018). Appropriate criteria: key to effective rubrics. *Frontiers in Education, 3*(22). https://doi.org/10.3389/feduc.2018.00022.

Butcher, P. G., & Jordan, S. (2010). A comparison of human and computer marking of short free-text student responses. *Comput Educ, 55*(2), 489–499. https://doi.org/10.1016/j.compedu.2010.02.012.

Chi, M. T. H., & VanLehn, K. A. (2012). Seeing deep structure from the interactions of surface features. *Educ Psychol, 47*(3), 177–188.

Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educ Psychol Meas, 20*(1), 37–46. https://doi.org/10.1177/001316446002000104.

Doherty, J. H., Scott, E. E., Cerchiara, J. A., McFarland, J., & Wenderoth, M. P. (2019). A learning progression characterizing how students in biology understand ion movement. Paper presented at the Annual International Meeting of the National Association for Research in Science Teaching (NARST). Baltimore, MD Mar 31-Apr 3.

Fleiss, J. L., & Cohen, J. (1973). The equivalence of weighted kappa and the intraclass correlation coefficient as measures of reliability. *Educ Psychol Meas, 33*(3), 613–619.

Friedman, J., Hastie, T., & Tibshirani, R. (2000). Additive logistic regression: a statistical view of boosting. *The Annals of Statistics, 2,* 337–407. 10/1214/aos/1016218223.

Friedman, J., Hastie, T., & Tibshirani, R. (2010). Regularization paths for generalized linear models via coordinate descent. *J Stat Softw, 33*(1), 1–22.

Gerard, L., Kidron, A., & Linn, M. C. (2019). Guiding collaborative revision of science explanations. *Int J Comput-Support Collab Learn, 14*(3), 291–324. https://doi.org/10.1007/s11412-019-09298-y.

Goldstone, R. L., & Day, S. B. (2012). Introduction to "new conceptualizations of transfer of learning". *Educ Psychol, 47*(3), 149–152. https://doi.org/10.1080/00461520.2012.695710.

Gotwals, A. W., Songer, N. B., & Bullard, L. (2012). Assessing students' progressing abilities to construct scientific explanations. In A. C. Alonzo & A. W. Gotwals (Eds.), *Learning progressions in science (pp. 183–210)*. Sense Publishing.

Harsch, C., & Martin, G. (2013). Comparing holistic and analytic scoring methods: issues of validity and reliability. *Assessment in Education: Principles, Policy & Practice, 20*(3), 281–307. https://doi.org/10.1080/0969594X.2012.742422.

Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The elements of statistical learning: data mining, inference, and prediction*. Springer.

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

Haudek, K. C., Santiago, M., Wilson, C. D., Stuhlsatz, M., Donovan, B., Buck-Bracey, Z., Gardner, A., Osborne, J. & Cheuk, T. (2019). Using Automated Analysis to Assess Middle School Students' Competence with Scientific Argumentation. Paper presented at the National Conference on Measurement in Education. Annual Conference, Toronto, ON. April 4-8, 2019.

Hearst, M. A., Dumais, S. T., Osuna, E., Platt, J., & Scholkopf, B. (1998). Support vector machines. *IEEE Intelligent Systems and their Applications, 13*(4), 18–28. https://doi.org/10.1109/5254.708428.

Hothorn, T., & Lausen, B. (2005). Bundling classifiers by bagging trees. *Computational Statistics & Data Analysis, 49*(4), 1068–1078. https://doi.org/10.1016/j.csda.2004.06.019.

Jescovitch, L. N., Doherty, J. H., Scott, E. E., Cerchiara, J. A., Wenderoth, M. P., Urban-Lurain, M., Merrill, J., & Haudek, K. C. (2019a). Challenges in developing computerized scoring models for principle-based reasoning in a physiology context. Paper Set: Measuring complex constructs in science education: Applications of automated analysis. Paper presented at the Annual International Meeting of the National Association for Research in Science Teaching (NARST). Baltimore, MD Mar 31-Apr 3. https://www.create4stem.msu.edu/publication/6728.

Jescovitch, L. N., Scott, E. E., Cerchiara, J. A., Doherty, J. H., Wenderoth, M. P., Merrill, J., Urban-Lurain, M., & Haudek, K. C. (2019b). Deconstruction of holistic rubrics into analytic rubrics for large-scale assessments of students' reasoning of complex science concepts. Practical Assessment, Research & Evaluation, 24(7). https://doi.org/10.7275/9h7f-mp76.

Jönsson, A., & Svingby, G. (2007). The use of scoring rubrics: reliability, validity and educational consequences. *Educational Research Review, 22*(2), 130–144. https://doi.org/10.1016/j.edurev.2007.05.002.

Jordan, M. I., & Mitchell, T. M. (2015). Machine learning: trends, perspectives, and prospects. *Science, 349*(6245), 255–260. https://doi.org/10.1126/science.aaa8415.

Jurka, T. P., Collingwood, L., Boydstun, A. E., Grossman, E., & Van Atteveldt, W. (2012). RTextTools: automatic text classification via supervised learning. *R package version, 1*(3), 9 http://CRAN.R-project.org/package=RTextTools.

Kazama, J., & Tsujii, J. (2005). Maximum entropy models with inequality constraints: a case study on text categorization. *Mach Learn, 60*(159), 159–194. https://doi.org/10.1007/s10994-005-0911-3.

Kotsiantis, S. B. (2007). Supervised machine learning: a review of classification techniques. *Informatica, 31*, 249–268.

Kotsiantis, S. B. (2012). Use of machine learning techniques for educational proposes: a decision support system for forecasting students' grades. *Artif Intell Rev, 37*(4), 331–344.

Landis, J. R., & Koch, G. G. (1977). The measurement of observer agreement for categorical data. *Biometrics, 33*(1), 159–174. https://doi.org/10.2307/2529310.

Large, J., Lines, J., & Bagnall, A. (2019). A probabilistic classifier ensemble weighting scheme based on cross-validated accuracy estimates. *Data Min Knowl Disc, 33*(6), 1674–1709. https://doi.org/10.1007/s10618-019-00638-y.

Lee, H., Pallant, A., Pryputniewicz, S., Lord, T., Mulholland, M., & Liu, O. L. (2019). Automated text scoring and real-time adjustable feedback: supporting revision of scientific arguments involving uncertainty. *Sci Educ, 103*(3), 590–622. https://doi.org/10.1002/sce.21504.

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

Liu, O. L., Rios, J. A., Heilman, M., Gerard, L., & Linn, M. C. (2016). Validation of automated scoring science assessments. *J Res Sci Teach, 53*(2), 215–233. https://doi.org/10.1002/tea.21299.

McHugh, M. L. (2012). Interrater reliability: the kappa statistic. *Biochemia Medica, 22*(3), 276–282 https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3900052/.

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. *Educ Assess, 23*(2), 121–138. https://doi.org/10.1080/10627197.2018.1427570.

Mayfield, E., & Penstein-Rose, C. (2010). An interactive tool for supporting error analysis for text mining. Proceedings of the NAACL, pp 25–28. https://www.aclweb.org/anthology/N10-2007.pdf

Michael, J., & McFarland, J. (2011). The core principles ("big ideas") of physiology: results of faculty surveys. *Adv Physiol Educ, 35*(4), 336–341.

Mitchell, T. (1997). *Machine learning*. McGraw Hill.

Mitchell, T., Russell, T., Broomhead, P., & Aldridge, N. (2002). Towards robust computerized marking of free-text responses. In *Proceedings of the sixth international computer assisted assessment conference* (pp. 233–249). Loughborough: Loughborough University.

Modell, H. I. (2000). How to help students understand physiology? Emphasize general models. *Adv Physiol Educ, 23*(1), S101–S107.

Mohan, L., Chen, J., & Anderson, C. W. (2009). Developing a multi-year learning progression for carbon cycling in socio-ecological systems. *J Res Sci Teach, 46*(6), 675–698.

Moharreri, K. M., Ha, M., & Nehm, R. H. (2014). EvoGrader: an online formative assessment tool for automatically evaluating written evolutionary explanations. *Evolution: Education and Outreach, 7*, 15.

Montgomery, K. (2002). Authentic tasks and rubrics: going beyond traditional assessment in college teaching. *Coll Teach, 50*(1), 34–40.

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

Nehm, R. H. (2019). Biology education research: Building integrative frameworks for teaching and learning about living systems. *Disciplinary and Interdisciplinary Science Education Research, 1*(15). https://doi.org/10.1186/s43031-019-0017-6.

Nehm, R. H., Ha, M., Rector, M., Opfer, J. E., Perrin, L., Ridgway, J., & Mollohan, K. (2010). Scoring guide for the open response instrument (ORI) and evolutionary gain and loss test (ACORNS). *Technical Report of National Science Foundation REESE Project, 0909999*.

Nehm, R. H., Ha, M., & Mayfield, E. (2012). Transforming biology assessment with machine learning: automated scoring of written evolutionary explanations. *J Sci Educ Technol, 21*(1), 183–196.

Nehm, R. H., & Haertig, H. (2012). Human vs. computer diagnosis of students' natural selection knowledge: testing the efficacy of text analytic software. *J Sci Educ Technol, 21*(1), 56–73.

Nicol, D. (2007). E-assessment by design: using multiple-choice tests to good effect. *J Furth High Educ, 31*(1), 53–64. https://doi.org/10.1080/03098770601167922.

Pellegrino, J. W., DiBello, L. V., & Goldman, S. R. (2016). Framework for conceptualizing and evaluating the validity of instructionally relevant assessments. *Educ Psychol, 51*(1), 59–81. https://doi.org/10.1080/00461520.2016.1145550.

Prevost, L. B., Smith, M. K., & Knight, J. K. (2016). Using student writing and lexical analysis to reveal student thinking about the role of stop codons in the central dogma. *CBE—Life Sciences Education, 15*(4), ar65. https://doi.org/10.1187/cbe.15-12-0267.

Schwartz, D. L., & Martin, T. (2004). Inventing to prepare for future learning: the hidden efficiency of encouraging original student production in statistics instruction. *Cogn Instr, 22*(2), 129–184.

Scott, E. E., Cerchiara, J. A., Jescovitch, L. N., Wenderoth, M. P., & Doherty, J. H. (2019). An emerging learning progression characterizing how students use mass balance reasoning to understand physiology. Paper presented at the Annual International Meeting of the National Association for Research in Science Teaching (NARST). Baltimore, MD Mar 31-Apr 3.

Sieke, S. A., McIntosh, B. B., Steele, M. M., & Knight, J. K. (2019). Characterizing students' ideas about the effects of a mutation in a noncoding region of DNA. *CBE-Life Sciences Education, 18*(2), ar18. https://doi.org/10.1187/cbe.18-09-0173.

Smith, C. L., Wiser, M., Anderson, C. W., & Krajcik, J. S. (2006). Implications of research on children's learning for standards and assessment: a proposed learning progression for matter and atomic-molecular theory. *MEASUREMENT: Interdisciplinary Research and Perspectives, 4*(1–2), 1–98. https://doi.org/10.1080/15366367.2006.9678570.

Sripathi, K. N., Moscarella, R. A., Yoho, R., You, H. S., Urban-Lurain, M., Merril, J., Haudek, K.(2019). Mixed student ideas about mechanisms of human weight loss. CBE Life Sciences Education, 18(3), ar37. https://doi.org/10.1187/cbe.18-11-0227.

Thomas, J., Holste, E., Draney, K., Bathia, S., Anderson, C. W., & Stroupe, D. (2019). Developing automated scoring for large-scale assessments of three-dimensional learning. Paper presented at the Annual International Meeting of the National Association for Research in Science Teaching (NARST). Baltimore, MD. Mar 31-Apr 3.

Tomas, C., Whitt, E., Lavelle-Hill, R., & Severn, K. (2019). Modeling holistic marks with analytic rubrics. *Frontiers in Education, 4*(89). https://doi.org/10.3389/feduc.2019.00089.

Weston, M., Haudek, K. C., Prevost, L., Urban-Lurain, M., & Merrill, J. (2015). Examining the impact of question surface features on students' answers to constructed-response questions on photosynthesis. CBE Life Sciences Education, 14(2), ar19. https://doi.org/10.1187/cbe.14-07-0110.

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

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

Wilson, M. (2009). Measuring progressions: assessment structures underlying a learning progression. *J Res Sci Teach, 46*(6), 716–730. https://doi.org/10.1002/tea.2031.

Yune, S. J., Lee, S. Y., Im, S. J., Kam, B. S., Baek, S. Y. (2018). Holistic rubric vs analytic rubric for measuring clinical performance levels in medical students. *BMC Medical Education, 18*(124). https://doi.org/10.1186/s12909-018-1228-9

Zhai, X. (2019, June) Applying machine learning in science assessment: opportunity and challenges. For *Journal of Science Education and Technology*. https://doi.org/10.13140/RG.2.2.10914.07365.

Zhai, X., Haudek, K. C., Shi, L., Nehm, R. H., & Urban-Lurain, M. (n.d.). From substitution to redefinition: A framework of machine learning-based science assessment. Journal of Research in Science Teaching, 1-30. https://doi.org/10.1002/tea.21658.

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