Developing Rubrics for AI Scoring of NGSS Learning Progression-Based Scientific Models

Leonora Kaldaras[a,b], Tingting Li[c], Kevin Haudek[c], Joe Krajcik[c]
a. University of Colorado Boulder
b. Stanford Graduate School of Education
c. CREATE for STEM Institute, Michigan State University

Abstract

The Framework for K-12 Science Education (the Framework) recognizes modeling as an essential practice for building a deep science understanding. Modeling assessments should measure the ability to integrate Disciplinary Core Ideas and Crosscutting Concepts, which reflects three-dimensional (3D) understanding. The Framework also promotes using learning progressions (LPs) as guides for organizing the learning process. Artificial intelligence (AI) such as machine learning (ML) holds the potential to streamline the evaluation of student LP-aligned models and improve student learning outcomes in this key scientific practice. However, simply evaluating presence or absence of certain elements in student models is not sufficiently meaningful LP-aligned evaluation. Rather, it is important to ensure that ML algorithms evaluate the same model attributes as a trained human scorer would and according to the criteria described by the LP. Analytic rubrics have been shown to be easier to evaluate the validity of ML-based scores. A possible drawback of using analytic rubrics is the potential for oversimplification of integrated ideas. We demonstrate the deconstruction of a 3D holistic rubric for modeling assessments aligned to NGSS-based 3D LP for Physical Science. We describe deconstructing this rubric into analytic categories for ML training that preserve its 3D nature, the necessary attributes of the modeling practice and the alignment to the NGSS 3D LP. In this context, the 3D LP is used as a guide to develop the rubrics capable of guiding ML algorithms to meaningfully evaluate 3D LP-aligned scientific models. This approach ensures validity of the resulting scores with respect to the 3D LP and the practice of modeling.

Introduction

The Framework for K-12 Science Education (the Framework, NRC 2012) and the Next Generation Science Standards (NGSS, 2013) specify three dimensions of science knowledge including Disciplinary Core Ideas (DCIs), Scientific and Engineering Practices (SEPs) and Crosscutting Concepts (CCCs). Science education researchers refer to the use of these three dimensions as three-dimensional (3D) learning because it reflects in students' ability to integrate these dimensions to make sense of phenomena and solve problems. 3D learning indicates deep science understanding because it reflects student ability to apply their knowledge (NRC, 2012; NGSS, 2013) to solve novel problems and make sense of complex phenomena..

The SEP of *Developing and Using Models* is an essential practice to help students build deep science understanding. Supporting students in developing modeling skills in the context of 3D learning poses several challenges. First, such assessments need to measure student ability to integrate relevant DCIs and CCCs to develop causal models of phenomena (Krajcik, 2021). Models are extremely time-consuming to evaluate and provide feedback for, making supporting students in developing modeling skills challenging (Zhai, Yin, Pellegrino, Haudek, Shi, 2020a). Second, the Framework emphasizes the developmental nature of student understanding which

states that complex science ideas take time, and appropriate scaffolding to develop (NRC, 2012, Smith et al., 2006). The developmental nature is reflected in the idea of a learning progression (LP), which is central to the Framework vision because it represents a roadmap to guide educators in helping students attain higher levels of understanding. Supporting students in developing modeling skills over time requires a validated LP describing increasingly sophisticated ways of integrating the SEP of modeling with relevant DCIs and CCCs. A validated LP helps to meaningfully use assessment results by providing guidance on what supports students need. Very few validated NGSS-aligned 3D LPs exist, which would facilitate the implementation of the NGSS vision.

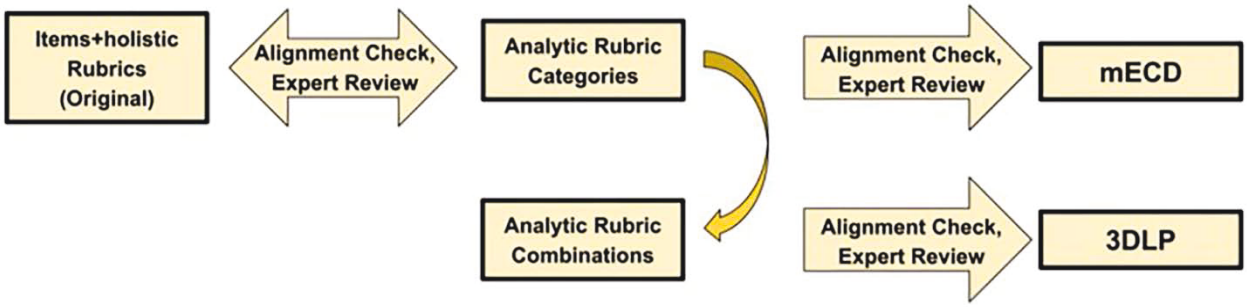
The current work reported here builds on a previously validated NGSS-aligned LP for electrical interactions (Kaldaras, Akaeze & Krajcik, 2021). This project aims to tackle the above challenges to help students build 3D proficiency by developing an automated, supervised machine learning (ML)-based system designed for open-ended formative assessments. Building this system involves several steps including developing high-quality ML-based scoring models and a system to deliver immediate LP-aligned feedback. Here, we report on the design process for using 3D LP to guide the development and alignment of rubrics for student models. When employed with collected student responses, these rubrics provide a well-annotated training set for ML-based models for automatic scoring that measures student ability to model electrostatic phenomena.

A central challenge in developing high-quality ML-based automatic scoring approaches for evaluating 3D LP-aligned scientific models lies in building the rubrics that accurately capture the complex 3D nature of student understanding according to the LP levels and yield high agreement between human and machine produced scores (Kaldaras, Yoshida, Haudek, 2022). Specifically, it is important to ensure that the machine scores reflect the modeling skills as opposed to student ability to develop representations with multiple artistic elements (Leong et al., 2018). Moreover, it is important that the machine scores reflect modeling skills integrated with disciplinary content and crosscutting concepts according to the levels of sophistication described by the LP. These properties will ensure that the scores used to develop supervised ML models reflect the integrated nature of 3D understanding according to the cognitive levels described in the 3D LP. The validated LP used in this study describes the elements of the modeling practice relevant to understanding electrical interactions.

Holistic rubrics that assess the overall quality of student performance have been used for 3D LP validation (Haudek et al., 2012; Kaldaras et al., 2021). In contrast, automatic scoring applications rely on analytic rubrics for scoring student responses (Liu et al., 2014). Analytic rubrics are a series of binary statements that identify the presence or absence of a construct. Scores generated by both holistic and analytic approaches have been used to develop functioning, predictive ML models for short, text-based constructed response (CR) items. However, analytic scoring provides an easier way for evaluating the validity of ML-based scores (Kaldaras & Haudek, 2022). Therefore, an analytic rubric can potentially be more useful in designing ML-based models for scoring. We developed a method for deconstruction of LP-aligned holistic rubric on the SEP of Constructing Explanations shown in Figure 1 (Kaldaras, Yoshida & Haudek, 2022). However, no research is available on how to deconstruct an LP-aligned holistic rubric into an analytic rubric based on the SEP of Developing and Using Models.

In this study, we demonstrate a method for deconstructing a 3D holistic rubric for modeling that probes the levels of a validated NGSS-aligned LP for electrical interactions (Kaldaras et al., 2021). We explain how to design an analytic rubric and validate it via human scoring. We also describe using the final consensus human scores based on the resulting analytic rubric to develop an ML model to automatically score responses to the modeling assessment item. This paper addresses the following question (RQ): *How can a validated 3D LP guide the deconstruction of 3D holistic rubrics into 3D analytic rubric categories to aid in developing ML-based scoring models for LP-aligned 3D assessments?*

Figure 1. The Process of Deconstructing a Holistic Rubric into Analytic Rubric (adapted from Kaldaras, Yoshida & Haudek, 2022).



Theoretical Framework

NGSS-Aligned 3D Learning Progression for Electrical Interactions

This study builds on a previously validated NGSS-aligned LP, describing student ability to integrate DCIs, SEPs, and CCCs when modeling and explaining phenomena involving electrical interactions (Kaldaras et al., 2021). The LP describes applying Coulomb's law and charge transfer to develop models that provide causal explanations about electrical interactions in various phenomena. Table 1 shows a description of the levels. The LP focuses on SEPs of Developing and Using Models and Constructing Explanations; the CCC of Cause and Effect and DCIs of Relationship between Energy and Forces and Types of Interactions.

Table 1. 3D LP for Electrical Interactions (adapted from Kaldaras, Akaeze & Krajcik, 2021)

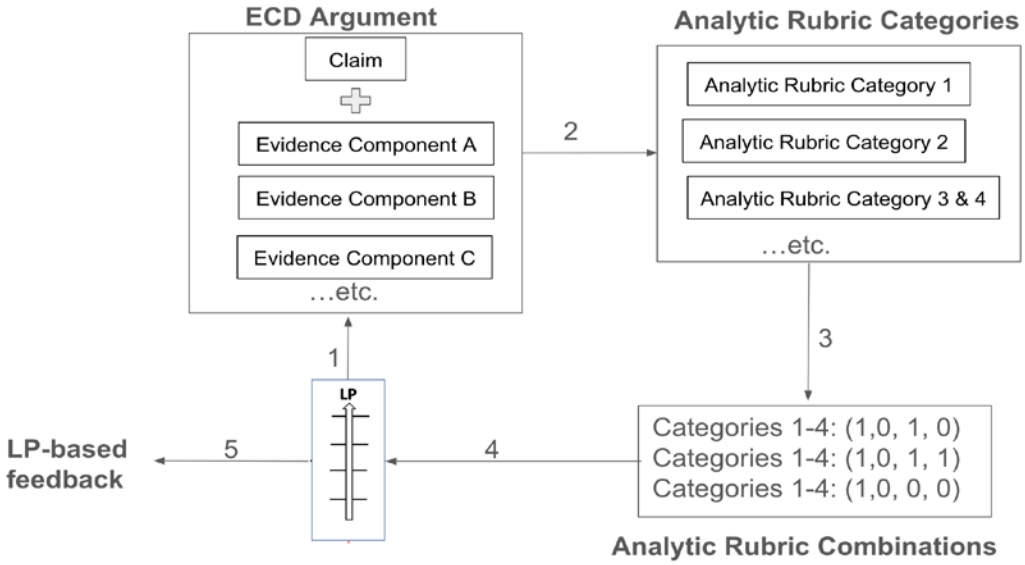
Level 3: Models and explanations represent causal relationships that use ideas of Coulombic interactions and electron transfer at the atomic-molecular level to explain phenomena.
Level 2: Models and explanations represent causal relationships that use the ideas of Coulombic interactions and charge transfer at the macro or atomic-molecular level to explain phenomena; may contain some inaccuracies related to atomic-level descriptions.
Level 1: Models and explanations represent partially causal relationships that use ideas of Coulombic interactions and electron transfer with inaccurate/incomplete ideas to explain phenomena.
Level 0: Models and explanations don't represent causal relationships and use Coulomb Law and/or electron transfer with significantly inaccurate/incomplete ideas.

The LP was validated by developing 3D modeling and explanation assessment tasks and aligned holistic rubrics (Kaldaras et al., 2021). The current work begins with a previously developed item and holistic rubric aligned to the 3D LP (Table 1). We demonstrate a LP-guided process of deconstruction of a holistic rubric into a series of analytic rubric categories.

A critical step in the LP-guided process of developing an analytic rubric is designing (or using existing) evidence centered design (ECD) argument to carefully specify the evidence that needs to be present in student responses to meet the requirements of the claim (step 1 in Figure 2). The claim in this context describes what students should be able to do in accordance to proficiency levels specified by the 3D LP. This evidence is a collection of statements that reflect what students should be able to do with respect to integrating the three relevant NGSS dimensions (DCIs, SEPs and CCCs) to meet the requirements of the claim ("ECD argument" in Figure 2, Evidence component A, B, C etc.). These evidence components then become the basis of the analytic rubric categories that will be developed and used to score student models (step 2 in Figure 2). The analytic rubric categories are described in a way that reflect presence or absence of ideas specified by a given component or components of the evidence statement ("Analytic Rubric Categories" in Figure 2). Lastly, analytic rubric categories are combined in specific ways guided by the 3D LP to yield specific analytic rubric combinations that reflect 3D LP-aligned classifications of student ability to develop scientific models using relevant DCIs and CCCs (step 3, "Analytic Rubric Combinations" in Figure 2). Each analytic rubric combination is then aligned to specific LP level (step 4 in Figure 2), and LP-based feedback specific feedback is designed for each scientific model evaluated using this process. A detailed diagram of this process is shown in Figure 2.

Note that the process rubric development shown in Figure 2 can be used for developing an LP-based analytic rubric from scratch without the need for the holistic rubric. However, since in the current study we had a holistic rubric available from prior work as well, we used a combination of approaches shown in Figures 1 and 2 to design our final analytic rubric. Specifically, we ensured that the resulting analytic rubric categories meaningfully align to the ECD argument and the previously developed holistic rubric as well. We will further describe this process in more detail.

Figure 2. The process of analytic rubric development and alignment to the 3D LP via ECD argument.



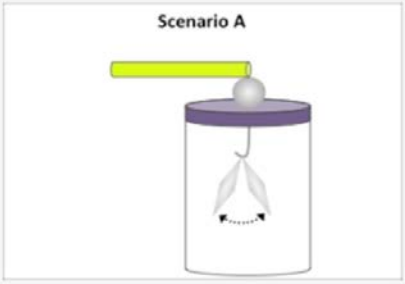
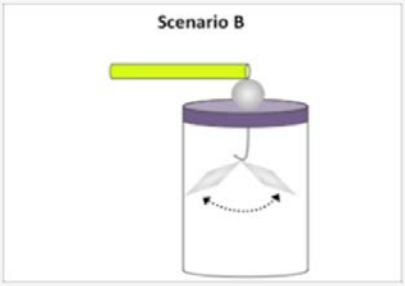
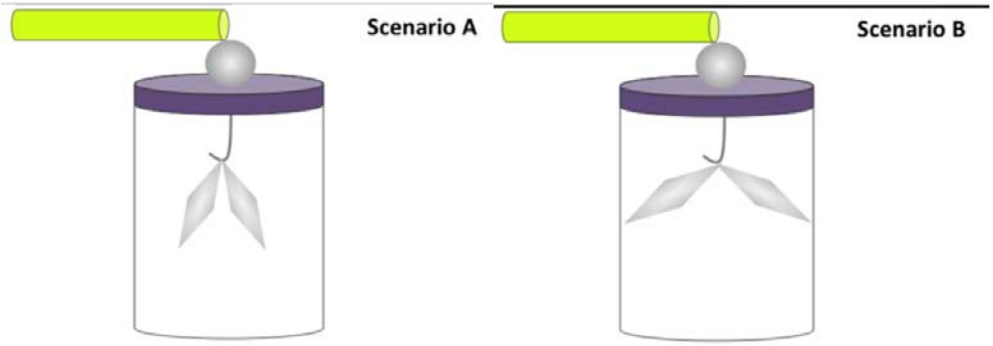
Background of existing modeling item and holistic rubric

Table 2 shows the information used in the original assessment development process. The item focused on probing modeling skills when explaining the interaction between the electroscope and the charged rod (Table 2). Table 2 provides information about each step of a modified evidence centered design process (mECD, Harris et al., 2019) for item development and corresponding holistic rubric for responses. A critical aspect of mECD is developing a claim describing what students should be able to do with respect to the three dimensions of NGSS targeted by the LP. Each level of the rubric aligns to the level of the LP shown in Table 1 and reflects the ideas that should be present in student responses.

The modeling item has several important features. First, this item is part of a group of items with the same storyline. Initially, students watch a video where they observe that when a rod is brought close to the electroscope, the foil leaves move apart. They are asked to write an explanation for what causes the leaves to move apart. Next, they are asked the modeling tasks shown in Table 2. Second, the item provides several important components of the model as a scaffold: the electroscope and the rod.

Table 2. Modified Evidence-Centered Design for the Electroscope item.

NGSS PE	HS-PS2-4. Use mathematical representations of Newton's Law of Gravitation and Coulomb's Law to describe and predict the gravitational and electrostatic forces between objects. Note: The gray text indicated the parts of the PE not evaluated by the item. Specifically, the item focused on evaluating student ability to use only Coulomb's law to develop *qualitative* models of electrostatic phenomena.
Claim	Students will construct a model to represent what causes neutral objects to become charged when put in contact with the charged objects and how the magnitude of the charge on the charged object affects the observations.
Evidence	1. Students show macroscopic causal mechanisms to model how neutral objects become charged (they will not use electron transfer). a. Charged objects are modeled as containing either positive or negative charges. b. Charge is modeled as point charge (either positive or negative). c. Charge transfers from the charged object to the neutral object when a charged object is touching the neutral object, which causes neutral objects to become charged. d. If a neutral object A is in contact with a neutral object B that was touched by a charged object, the charge will transfer from the charged object to the neutral object B and subsequently to neutral object A because object A is in contact with the now charged object B. This process causes the neutral objects that are not in direct contact with the charged objects to become charged. e. Models can show either positive or negative charges being transferred. 2. Students' models will show a causal relationship between the type and amount of charge and the magnitude and direction of associated electric

	force. Their models will include these ideas as appropriate: a. Opposite charges attract and like charges repel. b. Electric forces occur between two charged objects. The forces can be represented in terms of direction and strength. When two charged objects are attracted to each other, the force is directed toward the other object; when two objects repel each other, the force is directed away from each other. c. The amount of charge that two objects have affects the magnitude of interaction between them; the greater the charge, the stronger the interaction (Coulomb, amount).
Item	Scenario A below shows a diagram of what occurred in the video when a charged rod touched the ball.  In Scenario B, a rod touches the ball and makes the leaves move much further apart.  Draw a model to show what the differences are in the rod and foil leaves in the two scenarios. 
Holistic Rubric	Level 0: no model/no justification, model/justification is inaccurate, model/justification does not use charges and charge transfer, only observable components (for example, foil leaves are more open in scenario B). Level 1: model/justification only uses ideas related to the amount of charge and charge transfer (charge is shown on at least the rod and the leaves). Level 2: justification related amount of charge to generated repulsive force and shows/discusses charge transfer (charge is shown on at least the rod and the leaves).

Methods

Analytic Rubric Development

The Electroscope modeling item measures student ability to develop macroscopic level models of electrostatic phenomena, which is consistent with levels 0-2 of the LP. Students are not expected to use electron transfer or show atomic-level components at this point in the curriculum, nor was the item designed to do so. Notice the item as no prompts or scaffolds that would indicate to the learn to discuss electron transfer or atomic-level components. Rather, for this item it is sufficient to show charge transfer from the charged rod to all the components of the electroscope. Additionally, students need to show the magnitude of the repulsive force to be larger for scenario B compared to A.

Developing an analytic rubric involved specifying the necessary components of the model, the relationships between the components and the connection to the phenomenon. A detailed description of each component of the model is reflected in categories 1-10 (Table 3). Developing analytic rubric categories involves providing a fine-grained and detailed description of model components and relationships. The research team spent a significant amount of time revising the description of the categories to ensure accuracy and sufficient level of detail. Notice that each analytic rubric category is aligned to specific parts of the mECD argument ensuring that all the components, relationships and connections to phenomenon are captured in the rubric categories.

Table 3. Alignment between mECD argument and the analytic rubric categories.

Category	Description	ECD Evidence
1	Point charge (either + or –) on the rod in scenario A	1a, b,e
2	Point charge on the metal ball. The charge must be the same type as shown in the rod in scenario A. *Alternatively, models can show charge transfer from the rod to the ball with arrows, and not explicitly show point charges on the ball (there should be charges on the rod).*	1c,d,e
3	Point charge on the hook of the electroscope. The charge must be the same type as shown on the rod in scenario A. *Alternatively, models can show charge transfer from the ball to the hook/foil leaves with arrows, and not explicitly show point charges on the hook (there should be charges on the ball).*	1c,d,e
4	Point Charge on the leaves of the electroscope in scenario A. The charge must be the same type as shown in the rod in scenario A.	1c,d,e
5	Clearly indicates repulsive Electric force causes leaves to move, by using arrows or force representations and pointing in opposite directions between the leaves in scenario A	2a,b,c
6	Point charge on the rod in scenario B. The charge must be the same type as shown on the rod in scenario A. There must be more point charges on the rod in scenario B than in scenario A.	1c,d,e

7	Point charge on the sphere of the dome in scenario B. The charge must be the same type as shown on the sphere of the dome in scenario A. There must be more point charges on the sphere in scenario B than in scenario A. *Alternatively, models can show charge transfer from the rod to the ball with arrows, and not explicitly show point charges on the ball.*	1c,d,e
8	Point charge on the hook of the electroscope in scenario B. The charge must be the same type as shown on the hook in scenario A. There must be more point charges on the hook in scenario B than in scenario A. *Alternatively, models can show charge transfer from the ball to the hook with arrows, and not explicitly show point charges on the hook.*	1c,d,e
9	Point Charge on the leaves of the electroscope in scenario B. The charge must be the same type as shown in the leaves in scenario A. A. There must be more point charges on the leaves in scenario B than in scenario A.	1c,d,e
10	Clearly indicates repulsive Electric force causes leaves to move, by using arrows or force representations and pointing in opposite directions between the leaves in scenario B.. The repulsive arrows should be bigger or bolder (or both) for scenario B than for scenario A.	2a,b,c
11	Model shows both types of charges on one or more parts of the electroscope in one or both scenarios. *This can be ignored if positive and negative charges are not accumulated in specific locations.*	**Address inaccuracy**
12	Similar amount of charge on one or more parts of the electroscope in scenario A and B. This category only applies if they show the same type of charge through the entire model.	**Address inaccuracy**
13	Either the rod in scenario A is not charged or the whole electroscope is not charged in scenario A.	**Address inaccuracy**

To ensure that the 3D nature of the modeling item is reflected in the final assigned model score, the final assigned score should reflect student ability to develop a causal model for explaining the difference between scenario A and B by using ideas of Coulomb's law and charge transfer as described by the LP (Table 1). The mECD argument guided this process and served as a necessary link between the LP and the final score. Examples of how the resulting final score on the analytic rubric category combinations aligned to LP level is shown for level 1 and level 2 sample models in Figures 2 and 3 respectively. For example, Figure 3 shows a model that contains evidence for all necessary categories including charge transfer in both scenarios and the difference in the amount of charge and the associated repulsive force for both scenarios. A sample combination of "1" for categories 1-10 is indicative of level 2 of the LP. Similarly, Figure 4 shows a sample model containing evidence for categories 1, 4, 5, 6, 9, 10. This combination of categories is consistent with level 1 because the charge transfer is not fully shown. Notice that the LP and the LP-aligned mECD argument helped guide both the development of the analytic rubric categories and the alignment between the category combinations and the LP levels.

Figure 3. Sample Level 2 Model with rubric categories indicated.

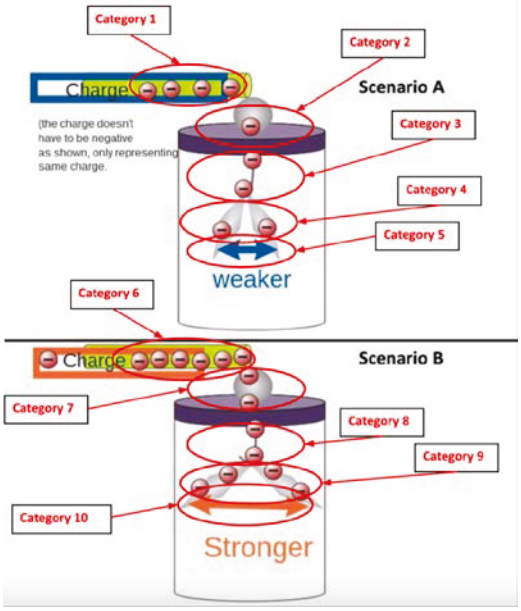


Figure 4. Sample Level 1 model.

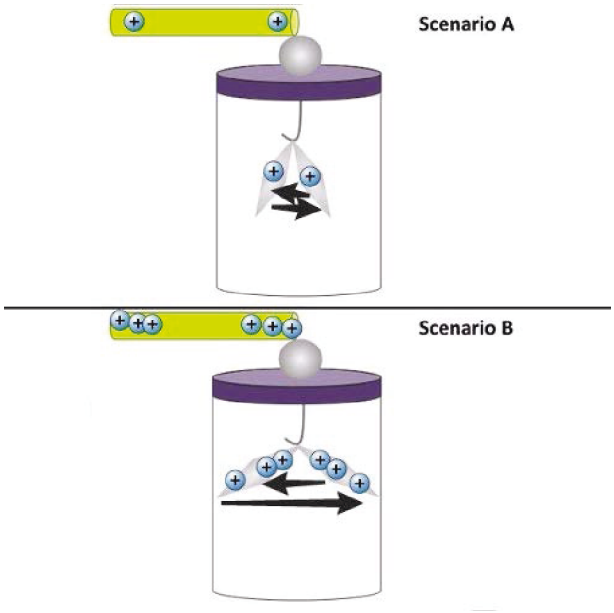
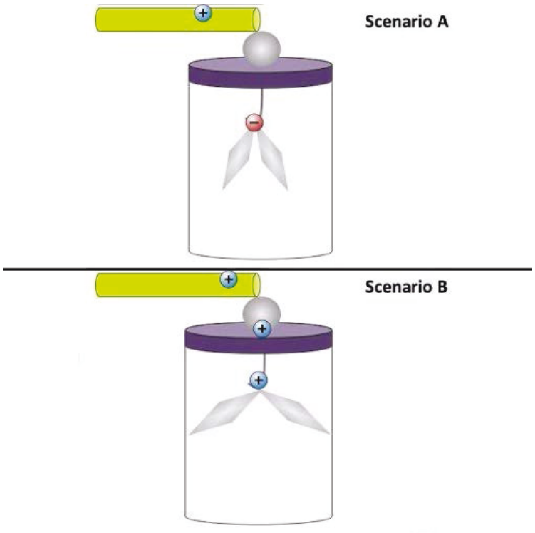


Figure 5. Sample student response for category 11.



The final consideration in developing the analytic rubric is how to deal with student responses that contain various inaccurate and/or incomplete ideas. This consideration is important for ensuring that any subsequent feedback provided to students meaningfully builds on the actual ideas reflected in student responses. To tackle this issue, we looked at a wide range of sample models and identified three broad categories of responses reflecting specific inaccurate and incomplete ideas. A sample student model reflecting Category 11 is shown in Figure 5. For this example, only category 11 is assigned a “1”, and all other categories are assigned a “0”.

Data sources

The Electroscope item was administered to 9th-grade students participating in the NGSS-aligned curriculum study. Unit 1 focused on ideas related to Coulomb's law as related to electrical interactions. The Electroscope item was administered as part of the Unit 1 pre and post-test and student responses from the posttest were used for the analysis reported here.

Analytic Scoring and Human Inter-Rater Reliability

The first researcher coded ~200 randomly selected student models to ensure that the rubric was easy to use and applied to a range of models. The rubric and the coded models were then reviewed by other researchers. Clarifications of rubric criteria and necessary additions were made to ensure the usability of the rubric. Three undergraduates were trained to apply the rubric to student models. Training was done in subsets of several hundred responses and coded independently by coders. Results from independent coding on subsets were checked for IRR (Krippendorff, 2004). We used a threshold of Krippendorff's alpha greater than 0.8 between human coders for each analytic category (Krippendorff, 2004). We then checked for human IRR. Categories that showed <0.8 Krippendorf's alpha between coders were discussed by the coders until agreed upon and the rubric was updated. A total of 1211 responses from students collected in 9th grade Physical Science classroom were scored by trained human scorers. This data set is subsequently used to train the ML model.

Results

Human Coding

As shown in Table 4, Krippendorff's Alpha values for most categories are at least 0.8, indicating strong human–human agreement (Krippendorff, 2004). We also observed very high accuracy (i.e. absolute agreement) measures between coders among nearly all categories. This suggests that the rubric is clear and interpretable by new coders and leads to reliable scoring.

Table 4. Human agreement measures for model scoring.

Category	Agreement	Krippendorff's Alpha
1	0.966	0.945
2	1.00	1.00
3	0.989	0.937
4	1.00	1.00
5	0.966	0.934
6	0.977	0.953

7	0.977	0.881
8	0.977	0.867
9	0.976	0.909
10	0.966	0.932
11	0.989	0.953
12	0.914	0.871
13	0.993	0.910

Machine Learning Model Training and Testing

We used supervised ML, specifically convolutional neural network analysis approach with Res-Net18 architecture as feature extraction network (He et al., 2016). The training data set contained 884 responses. Table 5 below shows the final human-machine agreement for each scoring category for the training stage. As shown in Table 5, human-machine agreement for all categories is above 85% accuracy, reflecting very high agreement (Lu et al., 2023). This suggests that the supervised ML approach accurately detected the model components and critical relationships within the model that were outlined in the rubric. We note that some of the categories with the lowest performance metrics are rubric categories associated with inaccuracies.

During training, we use the pretrained ResNet-18 (Residual Network) architecture, modifying its final fully connected layer to deliver binary output for our classification needs. The ResNet-18 architecture, noted for its deep residual learning framework, was employed as our feature extraction network (He et al., 2016). This network, with its depth of layers and residual connections, is particularly adept at learning from small datasets, which often pose challenges for deep learning models due to the risk of overfitting (He et al., 2016). To accommodate the input dimensionality and maintain consistency with the ResNet architecture, we set $d = 512$ (feature dimensionality) and resized all images to $W = H = 224$ (pixels).

Our model was implemented in PyTorch, benefitting from its flexible programming environment and efficient computational graph dynamics (Paszke et al., 2019). Optimization during training was conducted using the Adam optimizer, with a learning rate of $1e - 4$, balancing the advantages of adaptive gradient methods with the need for precision in the weight update process (Kingma & Ba, 2014). An NVIDIA GeForce GTX 1080Ti graphics card expedited the training process, enabling the efficient optimization of the model. Throughout the cross-validation process, we systematically assessed and saved the best-performing models according to validation metrics, opting for F1 score or accuracy based on the dataset's balance.

Table 5. Human-machine agreement for the ML training stage of model scoring.

	Validation			
	Accuracy (% +-SD)	Precision (% +-SD)	Recall (% +-SD)	F1 score (% +-SD)
C1	92.14+-9.66	92.44+-7.49	91.98+-7.99	87.27+-10.49
C2	94.90+-5.15	91.95+-6.89	90.76+-8.39	90.45+-8.30
C3	94.37+-6.31	90.40+-9.10	86.30+-9.18	86.56+-9.61
C4	91.25+-3.52	87.39+-5.17	86.00+-8.43	72.39+-6.92
C5	91.15+-4.98	89.49+-7.40	85.19+-8.14	85.96+-8.35
C6	87.19+-7.97	84.17+-7.52	82.56+-8.86	81.82+-9.90
C7	90.20+-10.65	83.54+-12.65	79.61+-11.42	78.75+-13.29
C8	93.24+-4.82	83.52+-12.04	81.78+-12.05	79.97+-11.01
C9	91.58+-5.84	87.91+-6.71	87.10+-8.02	86.28+-8.07
C10	91.43+-5.42	89.96+-7.11	84.91+-9.88	85.55+-9.80
C11	87.15+-9.40	87.15+-9.40	61.51+-15.13	57.34+-10.21
C12	87.05+-12.93	61.51+-15.57	58.64+-9.76	56.37+-10.67
C13	90.71+-5.32	78.02+-13.79	75.90+-14.09	74.92+-13.17

Discussion

It is challenging to train AI algorithms to recognize complex reasoning such as that reflected in students' scientific models. This is because machine learning algorithms should be trained to go beyond simple features of a given image to recognize specific aspects that are important for the practice of modeling focusing on evaluating causal aspects of scientific models explaining phenomena. The task of training AI to recognize such aspects related to scientific reasoning and skills becomes even more challenging when we are dealing with LP-aligned scientific models as called for by the NGSS and the Framework. The reason is that often scientific models at various LP levels might look very similar (compare level 2 and level 1 models in figures 3 and 4 respectively), but in reality, represent qualitatively different levels of understanding. If machine learning algorithms are not able to accurately compare these important differences, then we will not be able to design accurate and targeted LP-aligned feedback, which in turn defies the purpose of using ML techniques to solve one of the central current problems in education. It is therefore important to design approaches that leverage everything we know about how proficiency in a given construct develops, which is reflected in the LP-based vision, when designing AI-based methods for evaluating student learning in the context of NGSS aligned tasks that require the use of scientific reasoning. In this context, rubrics represent a crucial link between broad proficiency levels outlined in the LP and item-specific descriptions of how these proficiencies are reflected in student performance on a given assessment item. In this work we present a method for designing LP-aligned rubrics that serve as a roadmap for the ML algorithm guiding it as to what features are important to consider when scoring a scientific model. The proposed rubric development process yields a rubric that can be reliably used by coders and effectively guide ML algorithms to accurately evaluate scientific models at different LP levels. Final scores from application of the rubric exhibited high interrater reliability (Table 4) and human-machine agreement (Table 5). Moreover, final scores on the model reflect student 3D understanding with respect to modeling electrostatic phenomena, making this approach highly effective for scoring NGSS LP-aligned scientific models.

Study's Significance

This process of rubric development described here (Figure 2) represents a transparent and principle-based approach for designing LP-aligned analytic rubrics for AI scoring of any constructed response assessments (including scientific models, text-based explanations etc.). Defining analytic categories in this manner allows for easy identification of human-machine misscores by providing a straightforward way to pinpoint specific analytic rubric categories that were misscored. This property has the potential to improve overall validity of the associated AI-based scoring system. The process demonstrated here can serve as a guide for helping develop ML-based scoring approaches aimed at supporting students in developing scientific modeling skills using a curriculum that aligns with a validated LP and the associated assessment items. Using such an approach allows the possibility of implementing individualized and meaningful LP-aligned feedback statements as part of an automated assessment system. The mECD argument is an essential link between the LP and the final score. Finally, we extended an approach to analytic rubric development originally designed for Constructing Explanations to the SEP of Developing Models, while preserving the 3D nature of the item, rubric and resulting scores.

Acknowledgements

This material is based upon work supported by the National Science Foundation (Grant No. 2200757). Any opinions, findings and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the supporting agencies.

References

Harris, C. J., Krajcik, J. S., Pellegrino, J. W., and DeBarger, A. H. (2019). Designing knowledge-in-use assessments to promote deeper learning. Educ. Meas. Issues Pract. 38, 53–67. doi: 10.1111/emip.12253

Haudek, K. C., Prevost, L. B., Moscarella, R. A., Merrill, J., and Urban-Lurain, M. (2012). What are they thinking? Automated analysis of student writing about acid–base chemistry in introductory biology. CBE—Life Sci. Educ. 11, 283–293. doi: 10.1187/cbe.11-08-0084

He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 770–778).

Kaldaras, L., & Haudek, K. C. (2022a). Validation of automated scoring for learning progression-aligned Next Generation Science Standards performance assessments. In Frontiers in Education (Vol. 7, p. 968289). Frontiers Media SA.

Kaldaras, L., Akaeze, H., and Krajcik, J. (2021). Developing and validating next generation science standards-aligned learning progression to track three-dimensional learning of electrical interactions in high school physical science. J. Res. Sci. Teach. 58, 589–618. doi: 10.1002/tea.21672

Kaldaras, L., Yoshida, N. R., & Haudek, K. C. (2022b). Rubric development for AI-enabled scoring of three-dimensional constructed-response assessment aligned to NGSS learning progression. In Frontiers in Education (Vol. 7, p. 983055). Frontiers.

Krajcik, J. S. (2021). Commentary—applying machine learning in science assessment: Opportunity and challenges. J. Sci. Educ. Technol. 30, 313–318. doi: 10.1007/s10956-021-09902-7

Krippendorff, K. (2004). Content analysis: An introduction to its methodology. Thousand Oaks, California: Sage.

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

Leong, C. W., Liu, L., Ubale, R., & Chen, L. (2018, June). Toward large-scale automated scoring of scientific visual models. In Proceedings of the Fifth Annual ACM Conference on Learning at Scale (pp. 1-4).

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

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

Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., ... & Chintala, S. (2019). Pytorch: An imperative style, high-performance deep learning library. Advances in neural information processing systems, 32.

Smith, C. L., Wiser, M., Anderson, C. W., and Krajcik, J. (2006). Implications of research on children's learning for standards and assessment: A proposed learning progression for

matter and the atomic-molecular theory. Meas. Interdiscip. Res. Perspect. 4, 1–98. doi: 10.1080/15366367.2006.9678570

Zhai, X., Yin, Y., Pellegrino, J. W., Haudek, K. C., and Shi, L. (2020). Applying machine learning in science assessment: A systematic review. Stud. Sci. Educ. 56, 111–151. doi: 10.1080/03057267.2020.1735757.

Zongxing, L., Baizheng, H., Yingjie, C., Bingxing, C., Ligang, Y., Haibin, H., & Zhoujie, L. (2023). Human-machine interaction technology for simultaneous gesture recognition and force assessment: A Review. IEEE Sensors Journal.