# Toward Large-Scale Automated Scoring of Scientific Visual Models

**Chee Wee Leong**
Educational Testing Service
Princeton, NJ, USA
cleong@ets.org

**Lei Liu**
Educational Testing Service
Princeton, NJ, USA
lliu001@ets.org

**Rutuja Ubale**
Educational Testing Service
San Francisco, CA, USA
rubale@ets.org

**Lei Chen**
Liulishuo.com
Shanghai, China
chen.lei.05@gmail.com

## ABSTRACT
Visual models of scientific concepts drawn by students afford expanded opportunities for showing their understanding beyond textual descriptions, but also introduce other elements characterized by artistic creativity and complexity. In this paper, we describe a standardized framework for evaluation of scientific visual models by human raters. This framework attempts to disentangle the interaction between the scientific modeling skills and artistic skills of representing real objects of students, and potentially provides a fair and valid way to assess understanding of scientific concepts e.g. structure and properties of Matter. Additionally, we report ongoing efforts to build automated assessment models based on the evaluation framework. Preliminary findings suggest the promise of such an automated approach.

### ACM Classification Keywords
I.2.6 ARTIFICIAL INTELLIGENCE: Miscellaneous; I.4.10 IMAGE PROCESSING AND COMPUTER VISION: Image Representation

### Author Keywords
Large-Scale Educational Assessment; Automated Scoring; Visual Modeling

## INTRODUCTION
Assessment experts have noted that new reforms in science education require innovative assessments to probe multiple dimensions of science knowledge such as core concepts and science practices [3]. Policy experts and science education researchers have also called for the use of learning progressions (LPs) to guide assessment development in order to develop useful diagnostic tools of knowledge acquired by students and the ability to inform instruction [1]. The Next Generation Science Standards [2], developed for the U.S. K-12 educational system, explicitly identify modeling as one central and valued practice, and modeling is also identified as an important practice in mathematics. The visual models constructed by students can serve as rich vehicles of information for educators interested in supporting and assessing what students know and can do in science. In our previous research, we sought to develop a computer-based science assessment aligned with the NGSS and a learning progression (LP). We selected a disciplinary core concept – Matter, and a central practice i.e. developing and using models, as the target constructs for our assessment prototype that addressed the multidimensional features of science learning. There has been some research employing handwritten drawn models by students as a rich source of evidence to explore what they know about the structure and behavior of Matter [5, 7], and these have been used to construct an LP for Matter [7]. However, there are challenges remaining with regards to a large scale assessment of drawings by students. One obvious hurdle is the expensive labor costs associated with human scoring of such drawings at scale. In this proposed study, we explore ways to automate the scoring process to assess object-based drawings generated by students. Specifically, we posit that a fair and valid assessment of drawings must disentangle the interaction between the scientific modeling skills and the artistic skills of representing real objects by students. The findings will result in updated knowledge of cutting-edge automated scoring methods that can be applied to score student-generated models, and also inform the process of designing modeling prototypes to measure integrated science competency.



*L@S 2018,* June 26–28, 2018, London, United Kingdom



ISBN 978-1-4503-5886-6/18/06…$15.00

DOI: https://doi.org/10.1145/3231644.3231681

## LEARNING PROGRESSION (LP): A STANDARDIZED EVALUATION FRAMEWORK
Human scoring of visual models is based on a carefully developed scoring rubric that is aligned with the LP for Matter. The rubric includes four dimensions that address the understanding of the student in modeling the scale (S), material identity (MI), behavior (B), and distribution (D) of particles. For a given visual model, the scale dimension measures understanding

of composition of Matter beginning with the smallest units, e.g., nanoscopic particles. The material identity dimension examines the anticipated number/identity of particles present. The behavior dimension examines if/how particle movement is represented. Finally, the distribution dimension examines positions of individual particles and space between them in a given Matter state. Each dimension has its own sub-progression levels, starting with the most basic at level 1. Overall, in order to reach a given level of the LP, the student must demonstrate a minimum level of thinking in each of the four dimensions, and that minimum level may vary with the dimension. For instance, the progression from LP-3 to LP-4 requires a mastery of level 4 in scale dimension and a minimum of level 3 in behavior dimension.

Other transitionary requirements for progression of LP levels are shown in Table 1. A full treatment of the learning progression of Matter can be found in [6].

| | S | | | | MI | | | B | | | | D | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | **1** | **2** | **3** | **4** | **1** | **2** | **3** | **1** | **2** | **3** | **4** | **1** | **2** | **3** |
| **LP-1** | X | X | | | X | X | | X | | | | X | | |
| **LP-2** | | X | X | | | X | | X | X | | | X | | |
| **LP-3** | | | X | X | | X | X | | X | X | | | X | |
| **LP 4** | | | | X | | X | X | | | X | X | | X | X |
| **LP-5** | | | | X | | | X | | | | X | | | X |

**Table 1. A mapping between Learning Progression (LP) levels and levels in each rubric dimension. For a given LP level, indication of a 'X' means the minimum level that must be mastered in that associated dimension.**

## VISUAL MODELING

Visual models were collected through a pilot study that explored the implementation of a *formative assessment* prototype in two science classroom settings. In both classroom settings, teachers used the prototype assessment task to help students learn about the core concept of Matter. The formative assessment task was delivered online and students worked in pairs to input responses due to lack of access to technology in both classroom settings. Both teachers implemented the assessment task during a relevant unit of science instruction. Prior to the task, both teachers received a one-day professional development training on strategies for using formative assessment, the underlying science competency model, and the Matter LPs (see details of the competency model in [6]), the formative assessment task and other relevant supporting materials.

Specifically, in this assessment task, students were asked to draw and refine visual models of *pure water* and *ocean water*. Modeling involves the use of a computer-based drawing tool in which students used a virtual pen or selected from a pool of predefined objects – micro-objects (circle, square, triangle, diamond), macro-objects (fish, water drops, water steam, algae, salt, etc.) and labels – to express their idea of the structure of Matter. The drawing tool also allows students to change the size, color, and position of objects, add arrows to represent particle motion, and label objects. In total, we collected 148 student-pair visual models of Matter, of which 53 are generated for pure water model, 95 for ocean water model. Two educational experts doubly scored each drawing using the evaluation framework, with an agreement of 97.1% on the four dimensions and 100% on the LP level. Where there is disagreement, the two experts discussed to reach an consensus. Scoring is based uniquely on the content of the drawings. Overall, 85, 25, 35 and 3 drawings received a score of LP-1, LP-2, LP-3 and LP-4 respectively. No drawing was given a score of LP-5. Figure 1 shows examples of visual models generated by students.

## AUTOMATED SCORING

In this exploratory work, we implemented a standard approach to automated scoring in educational assessment [9]. First, we extract construct-relevant features from the data that are potentially correlated to human ratings on the scoring dimensions. Second, we attempt to build statistical models using these features to automate the scoring process. While textual description of models such as *"my model shows rain drops and blue squares because the square represents the ocean and the drops represent the rain."* are also available for analysis using natural language processing techniques, we focus on a unimodal approach here given the visual nature of the dataset.

### Data Format

Each visual model is a response by the student(s) elicited in the form of a drawing in the assessment task. The computer-based drawing toolkit allows extraction of visual attributes of each drawing using the data-interchange JavaScript Object Notation (JSON) file format that is self-descriptive. For each object drawn, JSON encodes its type, color (RGB with an alpha channel for specific opacity), text, X-Y coordinates, height, width and rotation in degrees. Each object drawn can be one of macro-objects, micro-objects, label, or arrow. A textual description is always provided by the student in the case of label.

### Feature Engineering

Importantly, the multidimensional scoring rubric targets different constructs which compositely estimates the learning progression level of a student. We hypothesize two categories of features in Table 2, each principally aligned with one or more of the constructs to ensure coverage in the scoring process.

#### *Counting-based features*

Crucially, a basic understanding of each scientific concept during visual modeling rests on knowledge of the number of type (identity) of particles present. This Expected Identity Count (EIC) is the number of distinctive types of nanoscopic particle present in a given model i.e. 1 (water particle) for the pure water model and 2 (salt and water particles) for the ocean water model. A unique particle identity can be specified by its color, type or a combination of both using micro-objects. Deviation from these expected counts indicates a gap in material identity awareness. Likewise, macro-objects such as fish and water drops when overused relative to micro-objects signal shallow understanding in the scale dimension. Behavioral-wise, arrows indicate direction of movement of particles and their lengths are used to gauge velocity of such movements.

#### *Spatial-based features*

Spatially, we are concerned with two aspects that specifically targets the distributive property of particles. To estimate spatial tightness and looseness, we adopt the $k$-Nearest Neighbor ($k$-NN) [4] algorithm to compute inter-particle distances. We

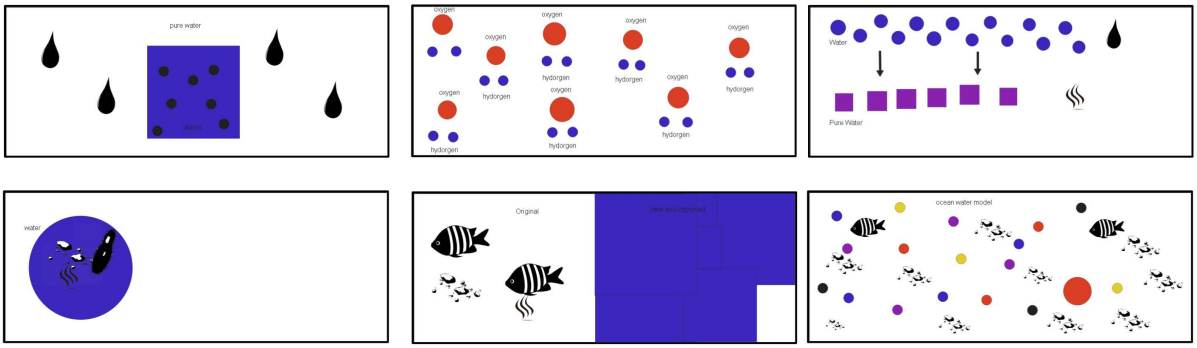


**Figure 1. Examples of drawings by students to illustrate scientific models of *pure water* (first row) and *ocean water* (second row). Specifically, Directional arrows can also be used to illustrate behavioral patterns of Matter particles i.e. water molecules.**

| Features | Brief Description | S | MI | B | D | LP |
|---|---|---|---|---|---|---|
| **Counting-based** | | | | | | |
| mico-object types | # types of micro-objects | 0.397 | 0.454 | 0.025 | 0.457 | 0.424 |
| macro-object types | # types of macro-objects | -0.467 | -0.449 | -0.340 | -0.414 | -0.469 |
| mico-object color types | # types of micro-object colors | 0.252 | 0.291 | 0.014 | 0.315 | 0.265 |
| EIC deviation | \| EIC - # types of micro-object color-shape \| | -0.376 | -0.394 | -0.132 | -0.342 | -0.373 |
| arrows | # arrow instances | 0.319 | 0.253 | **0.772** | 0.257 | 0.443 |
| arrow strengths | # mean length of arrow instances | 0.153 | 0.123 | 0.360 | 0.132 | 0.211 |
| arrow randomness | variance of direction of arrow instances | 0.182 | 0.158 | 0.597 | 0.173 | 0.324 |
| **Spatial-based** | | | | | | |
| *k*-NN-3 | mean distance to 3 NN micro-objects | 0.517 | 0.546 | 0.128 | 0.568 | 0.541 |
| *k*-NN-10 | mean distance to 10 NN micro-objects | 0.512 | 0.550 | 0.128 | 0.573 | 0.544 |
| dispersion | mean normalized spread between micro-objects | **-0.548** | **-0.589** | -0.128 | **-0.614** | **-0.578** |

**Table 2. Marginal correlations of individual feature against human scoring dimensions. S = Scale, MI = Material Identity, B = Behavior, D = Distribution, LP = Learning Progression. Except for macro-object types, all other features are based on micro-objects. Magnitude-wise, the largest correlation in each feature category per dimension is underlined, while the largest correlation overall within a dimension is in bold.**

used $k$=3 for a local approximation of proximity, and $k$=10 for a more global approximation. For a given visual model, the dispersion feature computes the number of particles per unit area per particle type, and averaging over all particle types. Consequently, a larger dispersion value is indicative of the same number of particles drawn over a larger canvas area in the visual model.

### Scoring Model

Given the numeric labels assigned to each LP level, it is possible to formulate the score prediction process as a supervised task of regression (with rounding) or classification using learners with matured statistical properties and explainable outputs, which are recommended for educational assessment tasks. We use scikit-learn toolkit [8] via SKLL[1] for building and evaluating the learners that are potentially deployable as our automated scoring system.

### Experiments

For experimentation, the 148 visual models represented by JSONs are shuffled randomly for a 10-fold cross-validation using each learner. Counting-based and spatial-based features are extracted for a total of 10 features per visual model. Specifically, log transformations are applied to spatial features for data smoothing. Marginal correlations (Pearson) based on a randomly chosen training partition for each feature are computed against each of the scoring dimensions, as well as LP, as shown in Table 2. Given that LP level prediction is the overall indicator for learning assessment, we used scikit-learn/SKLL to experiment with typically used regressors and classifiers using all available features and report their performance in Table 3. The learners are tuned on quadratic weighted kappa (QWK) which measures how well machine predicted scores agree with human scores.

[1]http://github.com/ EducationalTestingService/skll

## DISCUSSION

Referring to Table 2, a number of interesting observations regarding correlations can be made. First, the consistent negative correlations of the macro-object types feature across the different dimensions indicate that understanding levels are less sophisticated when students focus on drawing more macroscopic objects rather than explaining the nanoscopic aspects of Matter. This is particularly evident in the modeling the scale dimension. Secondly, the number of arrows, their direction and randomness almost exclusively account for showing understanding of the behavior of particles in Matter by students when compared to other features. Though, an indiscriminate use of arrows by students with no proper understanding is possible for gamification of the scoring model, and a realistic usefulness of this feature requires further examination. Third, we observe that all spatial-based features bear promising correlations ($|r| \geq 0.56$) in modeling the distribution dimension. Specifically, the dispersion feature stands out among all features in its consistent modeling all except for the behavior dimension. The Expected Identity Count (EIC) deviation is the only one that is engineered to target the concept-specific

| Learner | QWK | S.D. |
|---|---|---|
| Linear Regression | 0.624 | 0.12 |
| Decision Tree Regression | 0.517 | 0.29 |
| Random Forest Regression | 0.623 | 0.17 |
| Support Vector Regression (LibSVM) | 0.551 | 0.17 |
| Logistic Regression Classifier | 0.581 | 0.10 |
| Decision Tree Classifier | 0.559 | 0.07 |
| Random Forest Classifier | 0.538 | 0.06 |
| Support Vector Classifier (LibSVM) | 0.443 | 0.29 |

**Table 3. Quadratic Weighted Kappa and its standard deviation (overall for both pure and ocean models) reported by popular learners.**


**Figure 2. Learning curves for the best regressor/classifier**

visual model, where its value is dependent on whether the model is pure water and ocean water. Expectedly, this feature has a correlation of -0.394 for the material identity dimension, which indicates students would be penalized for deviating away from the expected number of identities anticipated.

After controlling for all other variables, analysis suggests that micro-object types, macro-object types, arrows and dispersion features are the most correlated with LP with partial correlations of 0.20, -0.31, 0.23 and -0.31 respectively. To demonstrate an in-depth understanding, students should target an all-around visual modeling approach that focuses on nanoscopic aspects of Matter, its behavior and taking advantage of the entire canvas while doing so.

Though LP levels can be classified numerically, prediction correlation with human scorers is better overall using regression-based learners of which the maximum QWK is achieved at 0.624. In the Linear Regression model, we also report an adjusted $R^2$ of 0.53, which suggests that approximately half of the variance in LP level differences can be accounted for by a linear regression model using the feature set proposed. Looking at Figure 2, we observe consistently that there is a small error gap in learning curves across both the best-performing regressor/classifier. QWK stagnates at approximately 0.6-0.7 for both training and cross-validation curves, indicating a high-bias problem. This corresponds to a case where adding more data for training is *less important* than conceiving the right features. Indeed, feature engineering efforts should be targeted toward constructs with high partial correlations to LP, namely the behavior and distribution dimension, as shown in Figure 3.

## CONCLUSION

In this paper, we explored potential automated scoring techniques for computer-based visual modeling of scientific concepts e.g. Matter. Scaling up assessment of such visual models require standardization of a feature framework to disentangle

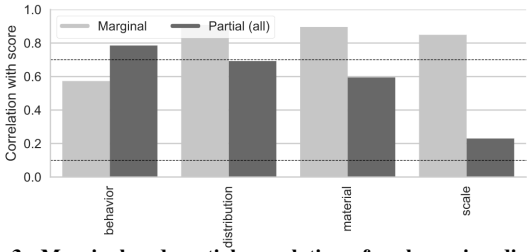


**Figure 3. Marginal and partial correlation of each scoring dimension against LP levels.**

artistic elements from modeling skills to ensure fairness in the scoring process. While we show that it is possible to build scoring models that are interpretable, most drawings by students in our dataset reside in the lower end of the LP framework (below LP-3), hence findings here are preliminary though promising. Future work will focus on validating the constructed models on a larger drawing sample encompassing all LP levels, engineering more informed features, and adopting a multimodal-based scoring approach that considers textual description of the visual model by students.

## REFERENCES


1. Thomas B Corcoran, Frederic A Mosher, and Aaron Rogat. 2009. Learning progressions in science: An evidence-based approach to reform. (2009).
2. National Research Council and others. 2013. *Next generation science standards: For states, by states*.
3. National Research Council and others. 2014. *Developing assessments for the next generation science standards*. National Academies Press.
4. Thomas Cover and Peter Hart. 1967. Nearest neighbor pattern classification. *IEEE transactions on information theory* 13, 1 (1967), 21–27.
5. Kenneth Forbus, Jeffrey Usher, Andrew Lovett, Kate Lockwood, and Jon Wetzel. 2011. CogSketch: Sketch understanding for cognitive science research and for education. *Topics in Cognitive Science* 3, 4 (2011), 648–666.
6. Lei Liu, Aaron Rogat, and Maria Bertling. 2013. A CBAL science model of cognition: Developing a competency model and learning progressions to support assessment development. *ETS Research Report Series* 2013, 2 (2013).
7. Joi DeShawn Merritt. 2010. *Tracking students' understanding of the particle nature of matter*. Ph.D. Dissertation. University of Michigan.
8. Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, and others. 2011. Scikit-learn: Machine learning in Python. *Journal of machine learning research* 12, Oct (2011), 2825–2830.
9. Mark D Shermis and Jill C Burstein. 2003. *Automated essay scoring: A cross-disciplinary perspective*. Routledge.