Using Machine Learning to Assess Next Generation Science Learning

Xiaoming Zhai

Assistant Professor in Science Education & (Affiliated) Artificial Intelligence

Director of AI4STEM Education Center

Email: Xiaoming.zhai@uga.edu

Website: http://ai4stem.uga.edu

UNIVERSITY OF GEORGIA

Dep. of Mathematics, Science, and Social Studies Education

AI for STEM Education Center

Institute for Artificial Intelligence

This study was partially funded by National Science Foundation(NSF) (Award # 2101104, 2100964, 2101166, 2101112, 1561159). Any opinions, findings, conclusions, or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the NSF.

Team

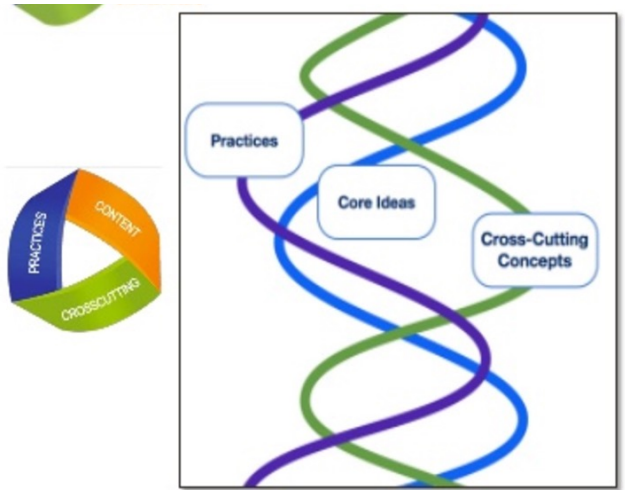
A Roadmap


DEVELOPING ASSESSMENTS FOR THE NEXT GENERATION SCIENCE STANDARDS
NEXT GENERATION SCIENCE STANDARDS
AI4STEM
Expanded Edition
How People Learn
Brain, Mind, Experience, and School
Knowing what Students Know
A FRAMEWORK FOR K-12 SCIENCE EDUCATION
Practices, Crosscutting Concepts, and Core Ideas

Next Generation Science Standards

New Standards for Science Learning

Framework for K-12 Science Education (NRC, 2012)
Next Generation Science Standards (NGSS Lead States, 2013)



- *The NGSS are written as Performance Expectations*
- *Each Standard represents a combination of all three dimensions.*
- *NGSS will require contextual application of the three dimensions by students.*
- *NGSS promotes Competency in Science.*

Three-Dimensional Learning Expectations

Students who demonstrate understanding can:

MS-PS1-4. **Develop a model that predicts and describes changes in particle motion, temperature, and state of a pure substance when thermal energy is added or removed.** [Clarification Statement: Emphasis is on qualitative molecular-level models of solids, liquids, and gases to show that adding or removing thermal energy increases or decreases kinetic energy of the particles until a change of state occurs. Examples of models could include drawing and diagrams. Examples of particles could include molecules or inert atoms. Examples of pure substances could include water, carbon dioxide, and helium.]

The performance expectation above was developed using the following elements from the NRC document *A Framework for K-12 Science Education*:

Science and Engineering Practices

Developing and Using Models
Modeling in 6–8 builds on K–5 and progresses to developing, using and revising models to describe, test, and predict more abstract phenomena and design systems.

- Develop a model to predict and/or describe phenomena.

Disciplinary Core Ideas

PS1.A: Structure and Properties of Matter

- Gases and liquids are made of molecules or inert atoms that are moving about relative to each other.
- In a liquid, the molecules are constantly in contact with others; in a gas, they are widely spaced except when they happen to collide. In a solid, atoms are closely spaced and may vibrate in position but do not change relative locations.
- The changes of state that occur with variations in temperature or pressure can be described and predicted using these models of matter.

PS3.A: Definitions of Energy

- The term "heat" as used in everyday language refers both to thermal energy (the motion of atoms or molecules within a substance) and the transfer of that thermal energy from one object to another. In science, heat is used only for this second meaning; it refers to the energy transferred due to the temperature difference between two objects. *(secondary)*
- The temperature of a system is proportional to the average internal kinetic energy and potential energy per atom or molecule (whichever is the appropriate building block for the system's material). The details of that relationship depend on the type of atom or molecule and the interactions among the atoms in the material. Temperature is not a direct measure of a system's total thermal energy. The total thermal energy (sometimes called the total internal energy) of a system depends jointly on the temperature, the total number of atoms in the system, and the state of the material. *(secondary)*

Crosscutting Concepts

Cause and Effect

- Cause and effect relationships may be used to predict phenomena in natural or designed systems.

Next Generation Science Learning

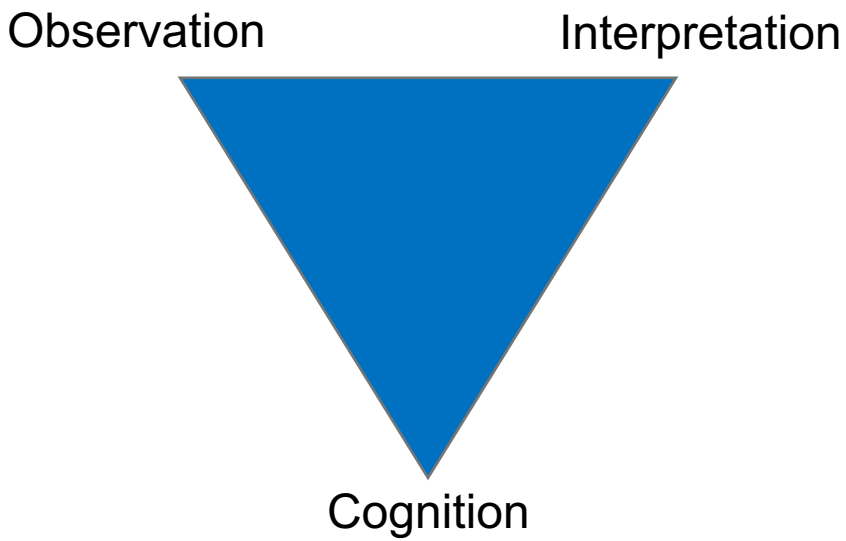
Challenges: Assess students' performance on three-dimensional learning

- Multiple-Choice questions are not able to assess three-dimensional learning
- Constructed responses are needed
- Performance-based assessments are needed
- Scoring students are time-consuming and timely feedback is not available

Assessment Practices

- Identifying learning goals
- Eliciting performance
- Interpretation observations
- Decision making and action-taking

http://nextgenscienceassessment.org



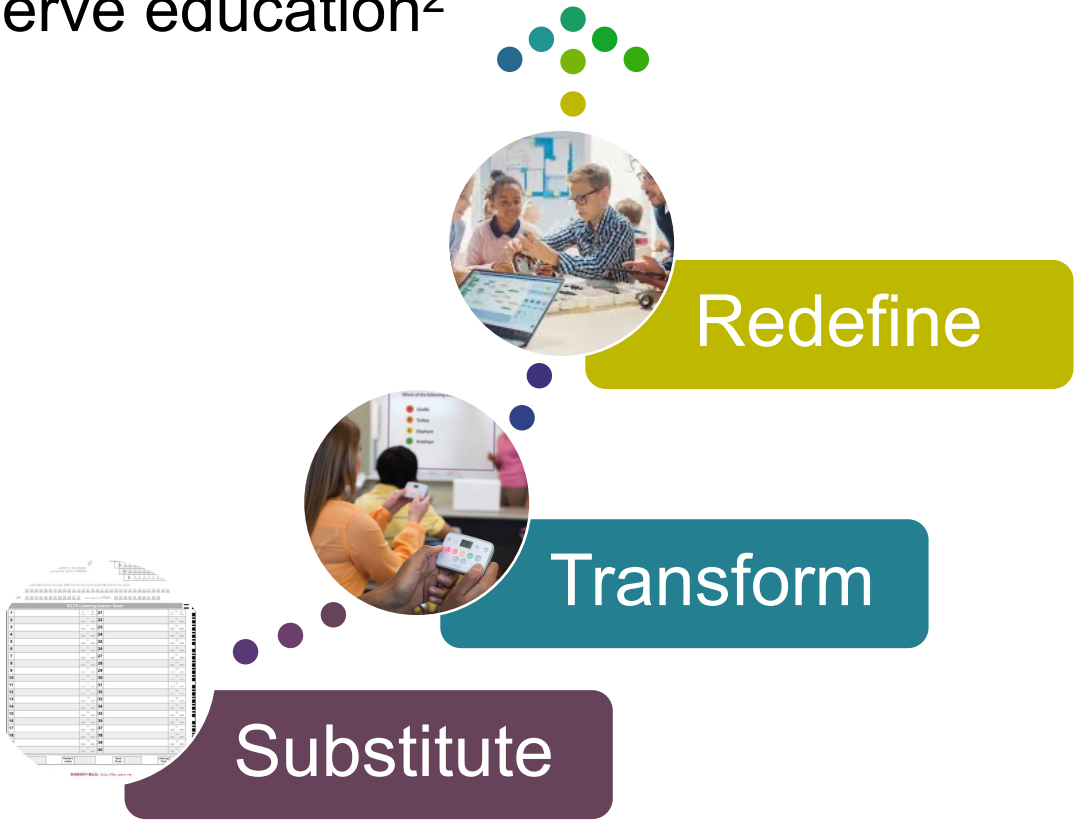
Innovative Assessment

Technologies extended the nature of the problems that can be presented in assessments, as well as the approaches of eliciting and interpreting evidence, and thus enhancing the assessment practices.

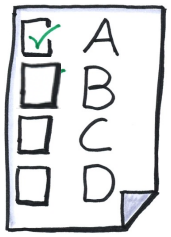
- **Construct**: latent trait of examinees (e.g., knowledge-in-use)
- **Functionality**: evidentiary reasoning processes
- **Automaticity**: human efforts

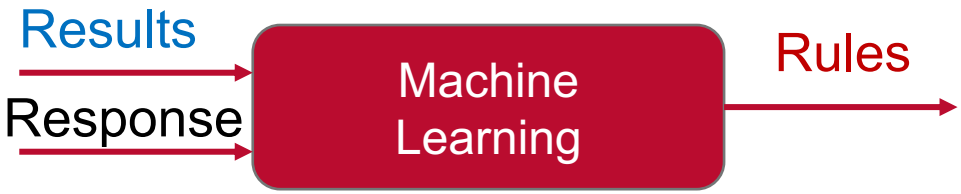
Levels of Innovative Assessments

- Degree to which we could approach the assessment goals that serve education[2]

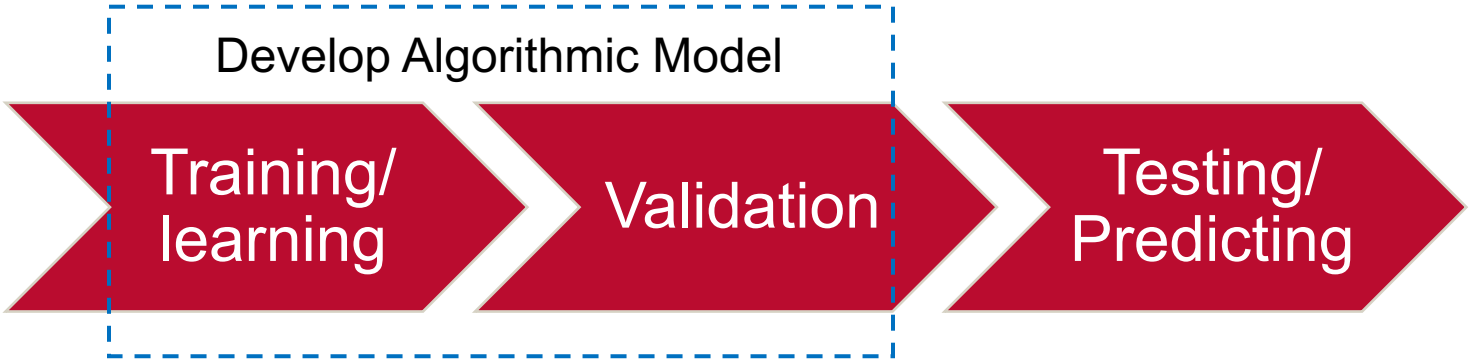


Why Machine Learning?



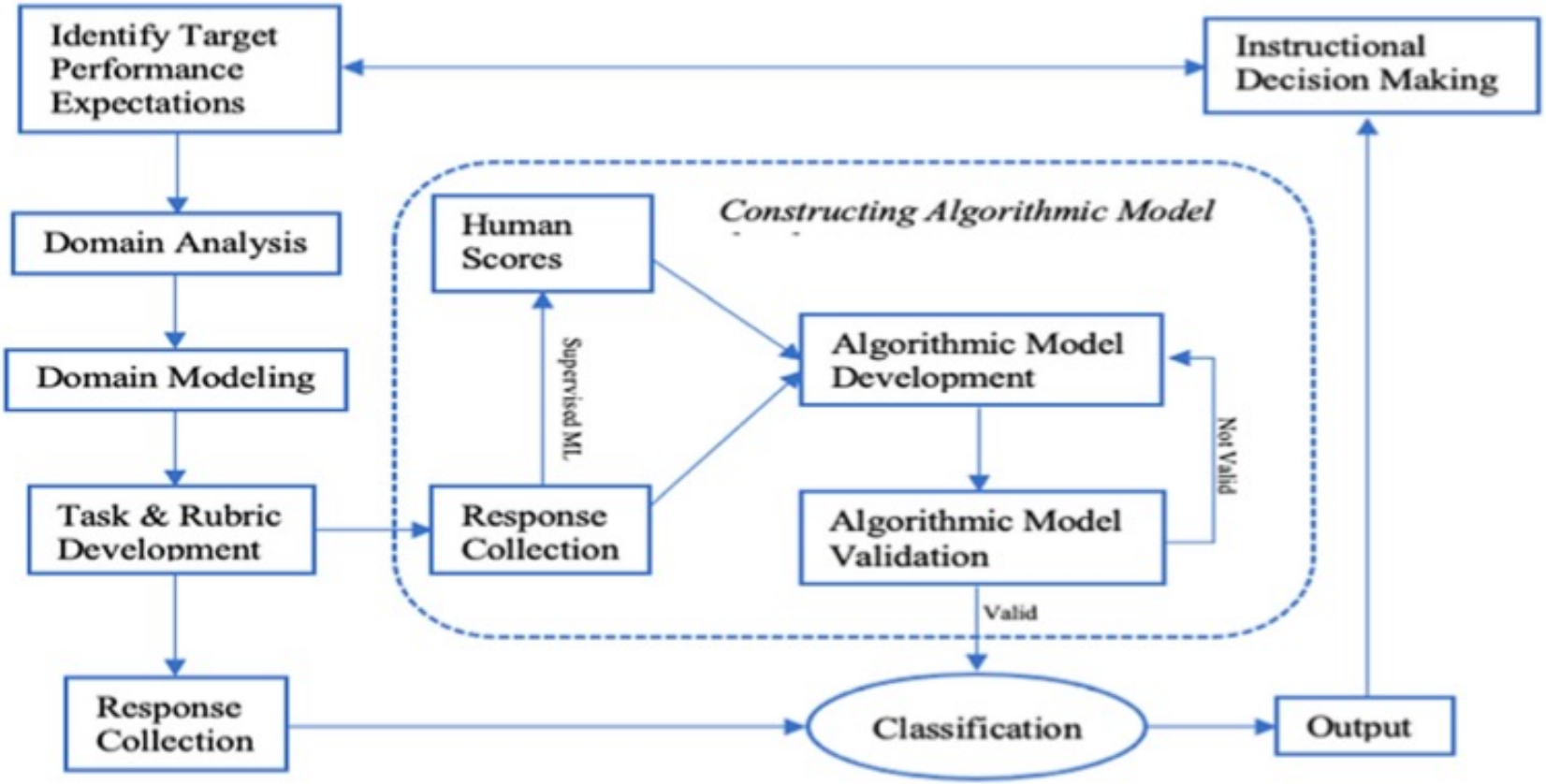


What is Machine Learning?

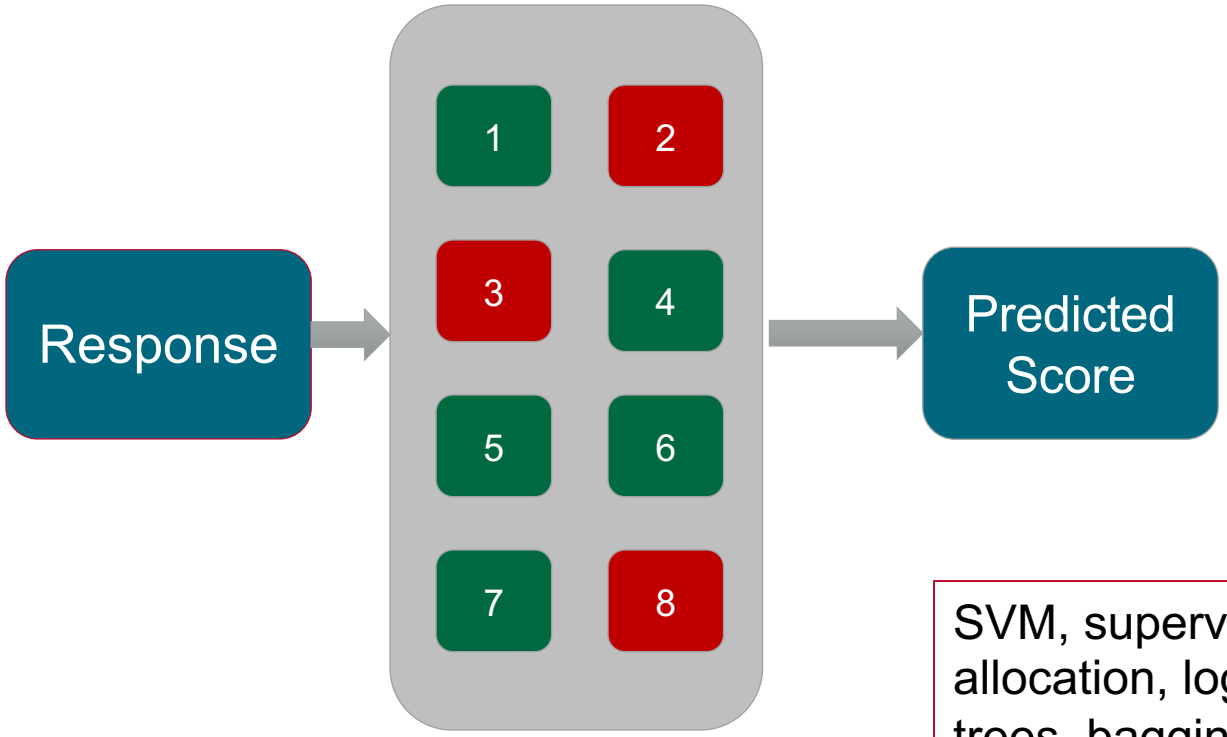


Supervised ML	• Learn from labeled data
Unsupervised ML	• Learn from raw data
Semi-supervised ML	• Learn from both labeled and raw data

Framework for ML-based Next Generation Science Assessment

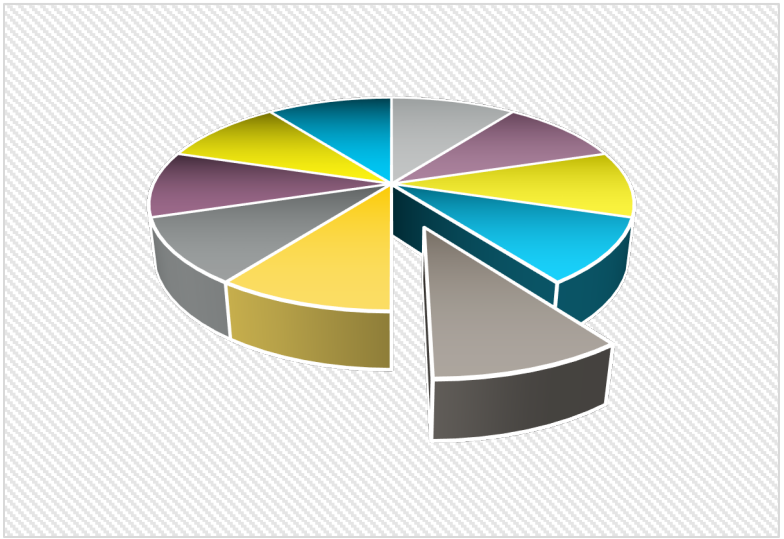


Machine Learning Ensemble Classifiers



SVM, supervised latent dirichlet allocation, logitboost, classification trees, bagging classification trees, random forests, penalized generalized linear models, and maximum entropy models

Cross-validation



Training set: ☺ Human labeled

Computer Algorithm

Validation set: ☺ Human labeled

Machine labeled

Human-machine Cohen's Kappa

HA.0 Modules

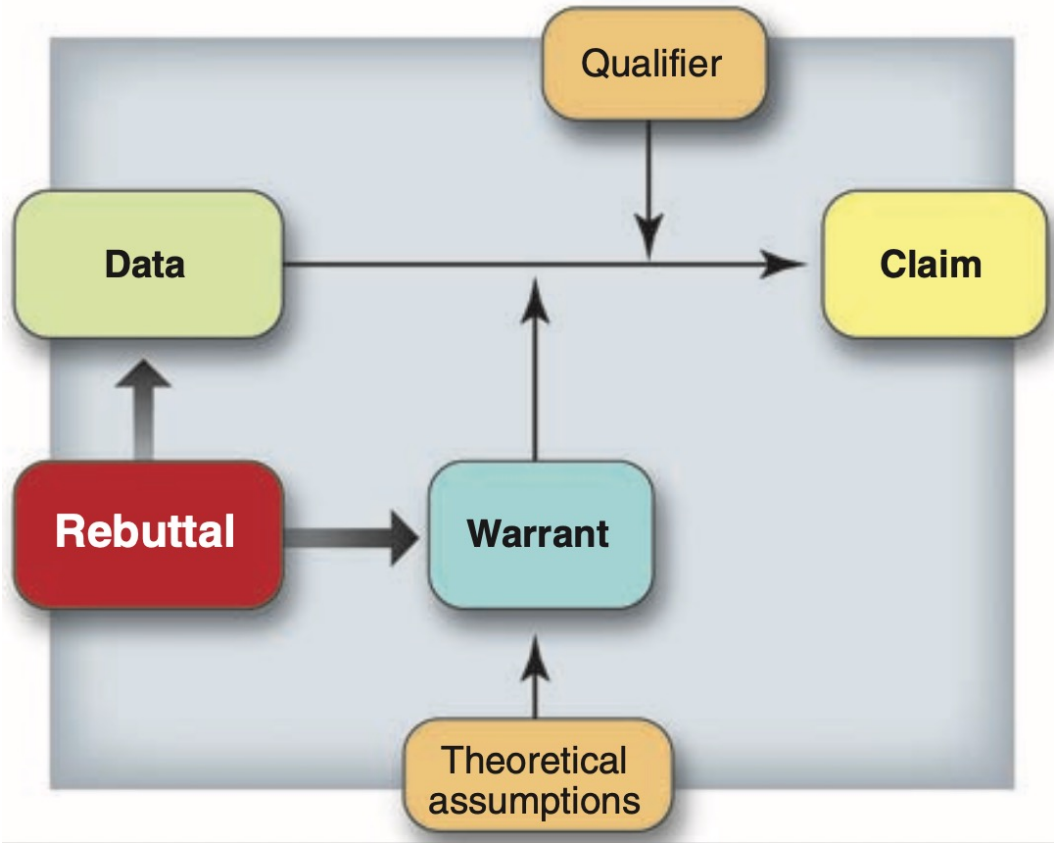
HĀ

Learning Progression

Application 1: Automatically Assess Argumentation Learning Progression

A Framework for Argumentation in Science

Osborne et al. (2010) based on Toulmin's structure for practical arguments.



Methods

- Created 19 items to cover nearly the entire construct map for scientific argumentation. Set in three different science contexts.
- 931 responses were collected from 2 sources in California
- Analytic coding rubrics were designed to identify the key components of argumentation for each item
- Two coders trained for each item set until IRR > 0.7; then coded ½ remaining data set independently, with an overlapping subset for final IRR calculations

Methods-Sample Item and Rubrics

Laura and Mary do an experiment and pour grains of sugar into a glass of water. After stirring the glass with a spoon for a few minutes, they cannot see the grains of sugar.

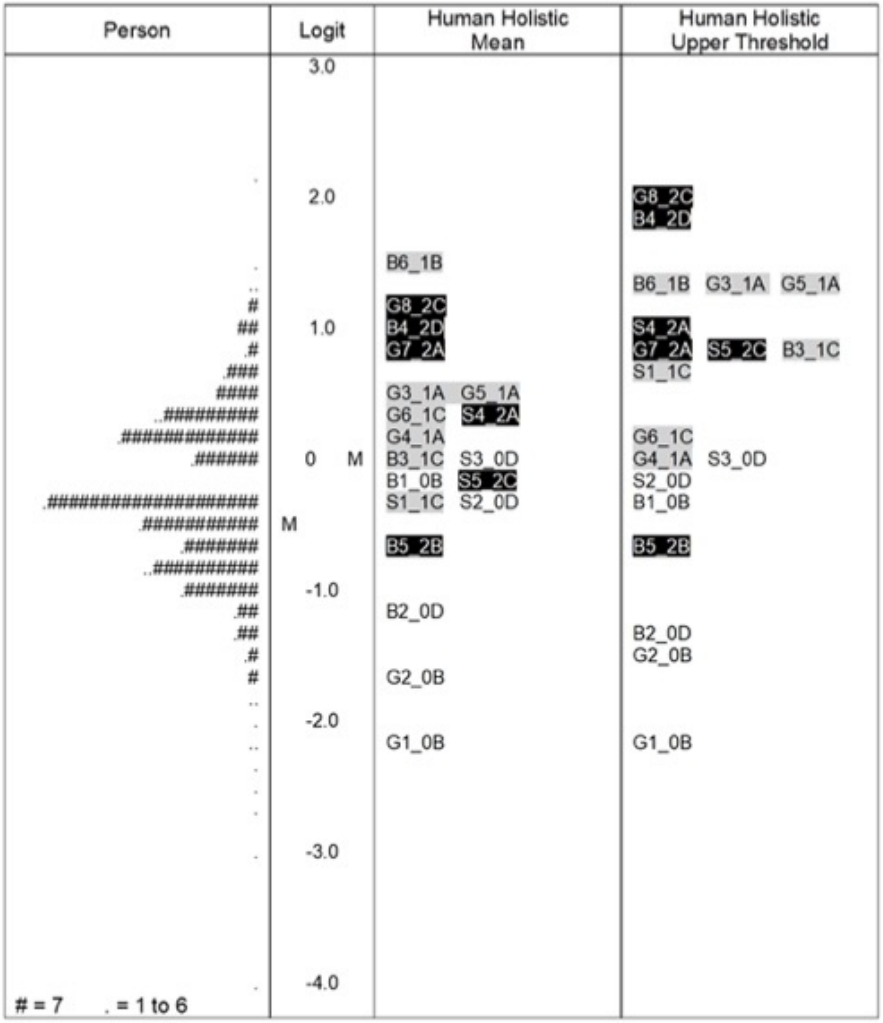
1. Make a scientific argument about what happened to the sugar using the information above.

Level	Constructing	Critiquing
0a	Constructing a claim	
0b		Identifying a claim
0c	Providing evidence	
0d		Identifying evidence
1a	Constructing a warrant	
1b		Identifying a warrant
1c	Constructing a complete argument	
1d	Providing an alternative counter argument	
2a	Providing a counter-critique	
2b	Constructing a one-sided comparative argument	
2c	Providing a two-sided comparative argument	
2d	Constructing a counter claim with justification	

Methods – Sample Item and Rubric

Analytic Component	Examples
COMPONENT A: Possible claims	
A1: The sugar dissolved	21. The sugar was dissolved in the water. Because the grains of sugar are so small, they blend with the water and become unnoticeable to the naked eye.
COMPONENT B: Possible evidence	
B1: The sugar disappeared, or you can no longer see the sugar.	29. The sugar dissolved into the water, forming a mixture, so that the individual sugar grains are no longer visible.
COMPONENT C: Possible Warrant/Reason	
W1: The sugar broke into pieces	9. The sugar gets dissolved, and the particles break up until they are spread out throughout the liquid and you no longer can see them with the naked eye.
W2: The sugar mixed or combined with water.	21. The sugar was dissolved in the water. Because the grains of sugar are so small, they blend with the water and become unnoticeable to the naked eye.
W3: Physical act of stirring	2505. I think that the sugar in the cup dissolved in the water because after they stirred it, the sugar was no longer visible.

Wright Map of Argumentation Items



Machine Scoring Accuracy

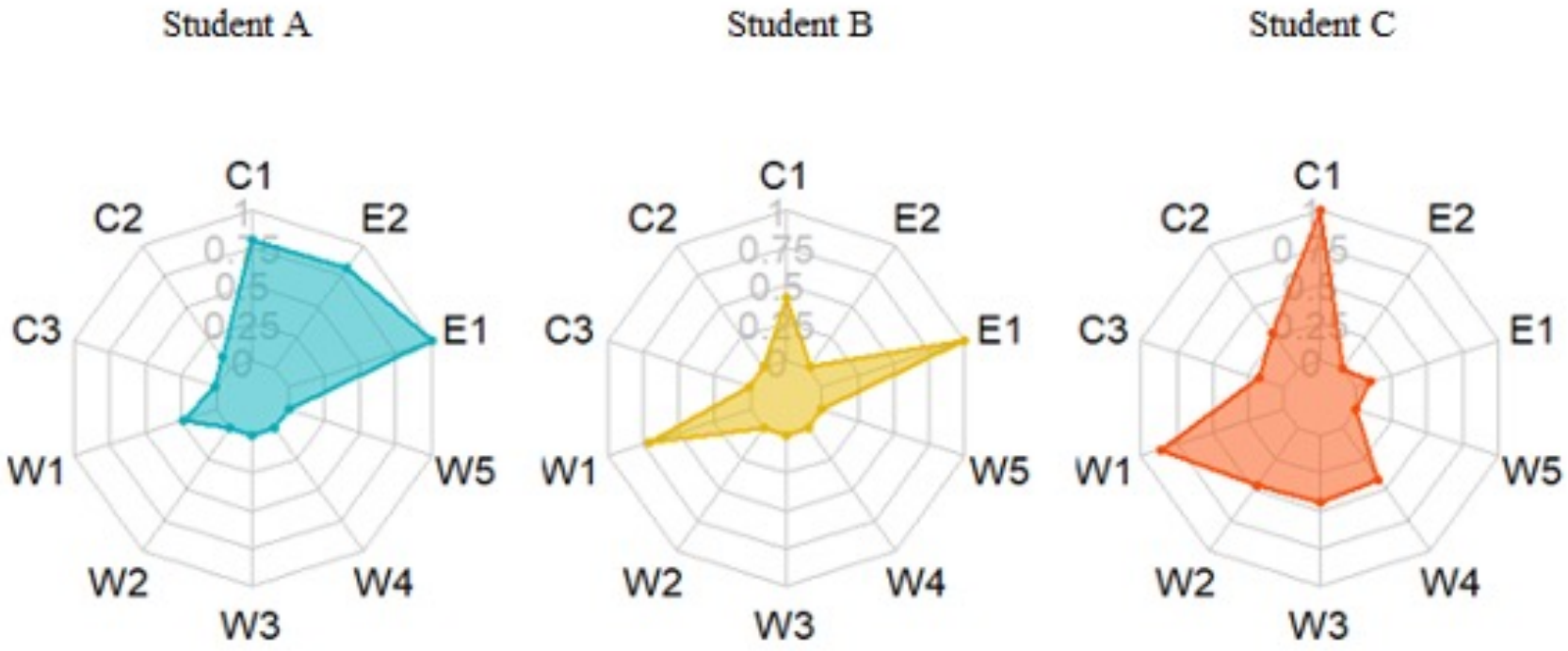
Item	Human-Computer agreement (κ)	# of levels in rubric		Item	Human-Computer agreement (κ)	# of levels in rubric
S1	0.75	4		B6	0.88	3
S2	0.82	2		G1	0.63	2
S3	0.81	2		G2	0.74	2
S4	0.73	4		G3	0.60	4
S5	0.68	4		G4	0.71	3
B1	0.93	3		G5	0.71	3
B2	0.87	3		G6	0.65	3
B3	0.69	4		G7	0.59	3
B4	0.71	4		G8	0.70	4
B5	0.74	4				

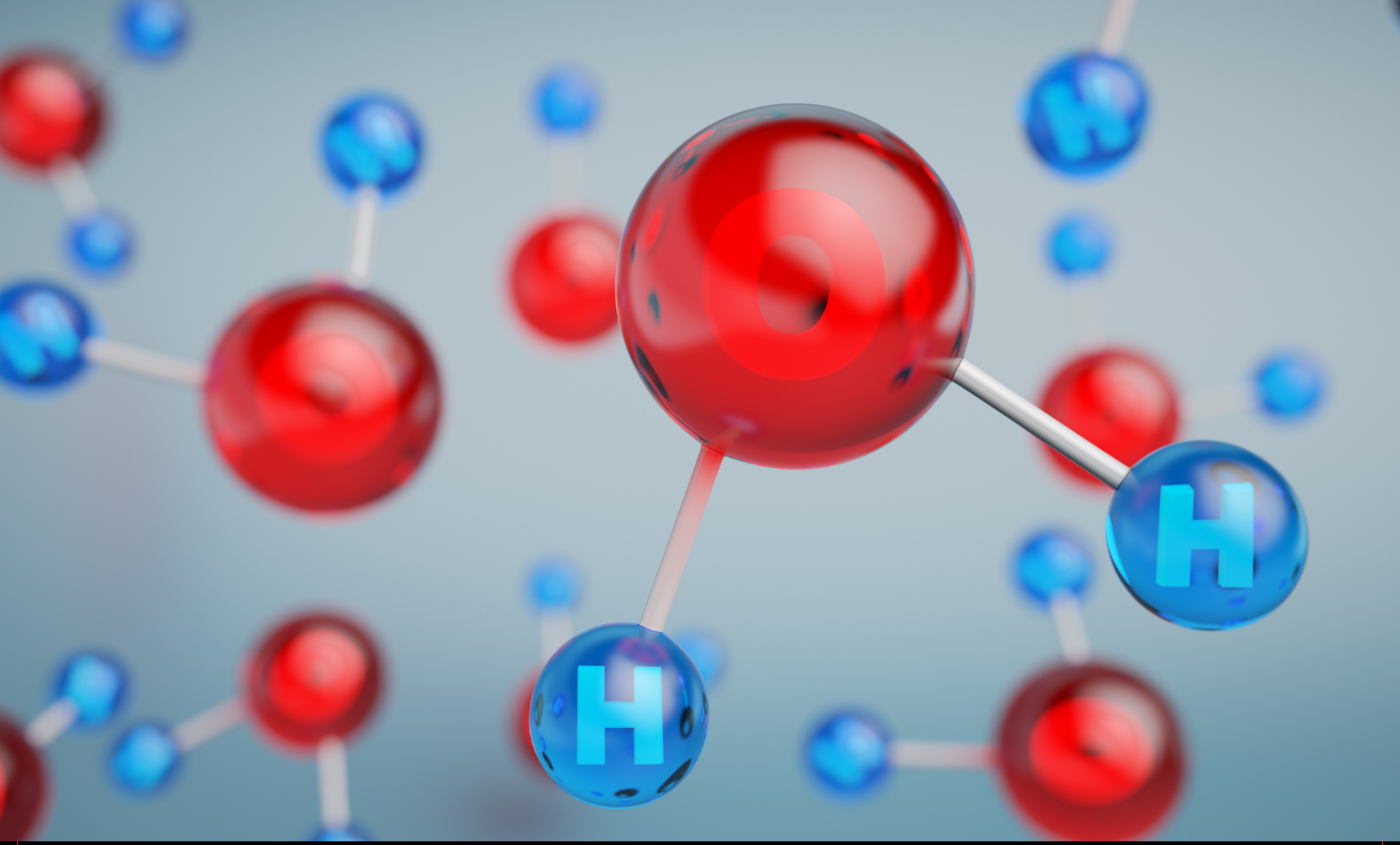
Students' mastery patterns in the seven classes

Table 5 Argumentation mastery patterns revealed from CDM analysis

C1	C2	C3	W1	W2	W3	W4	W5	E1	E2	%
x	x	x	x					x		13.7%
x			x							9.3%
x			x	x	x	x				8.6%
x	x							x		7.8%
										7.5%
								x		7.0%
x	x	x								6.5%
x	x		x					x	x	5.3%
x			x					x		5.2%
x			x					x	x	3.7%
x	x		x	x						3.4%
x								x	x	3.1%
x	x		x					x		2.5%
x	x									2.3%
x			x	x	x	x	x			2.1%
x	x	x						x		1.7%
			x					x		1.6%
x	x		x							1.5%
x										1.2%
								x	x	1.2%
x	x	x	x	x	x			x		1.0%

Probability of mastering each skill for three students



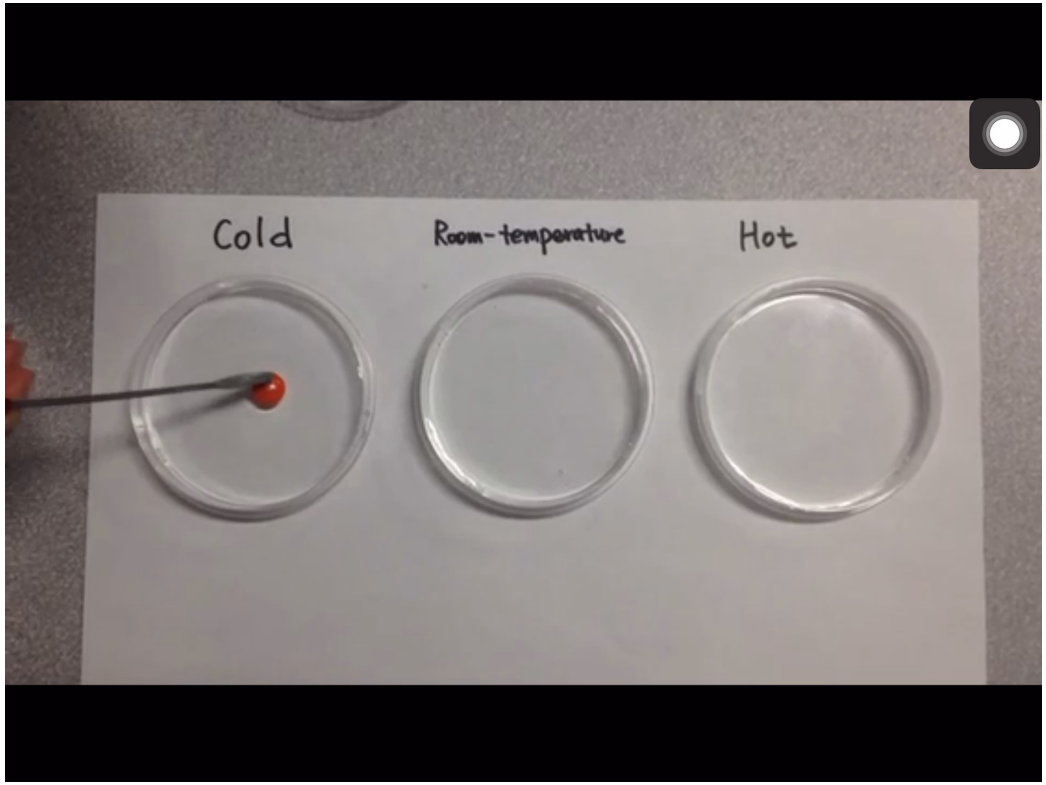


Application 2. Automatically Assess Drawn Models

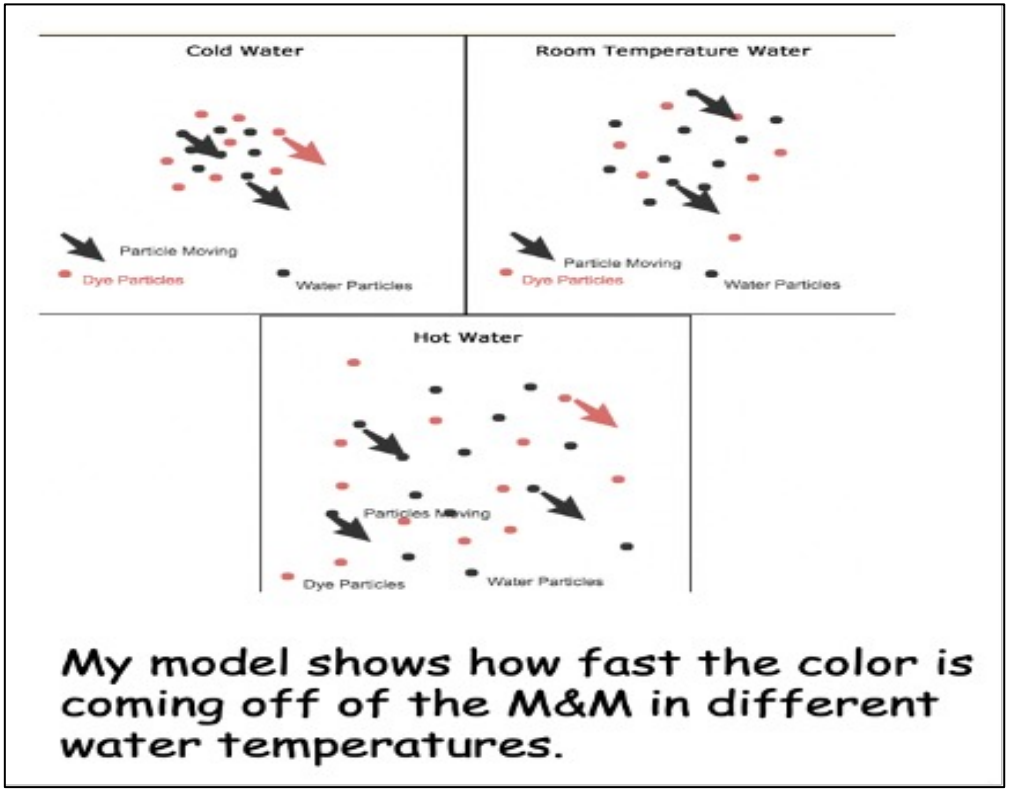
What is Scientific Model?

- Constitutes System, components, relations between components
- Reflect causal relationships or underlying mechanism of phenomena
- Abstraction
- Multi-representative
- Revisable
- Use and reflect consensus knowledge of the science community

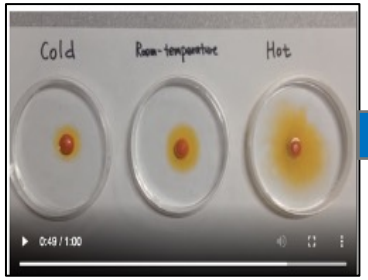
Theoretical Perspective: What is Scientific Modeling?

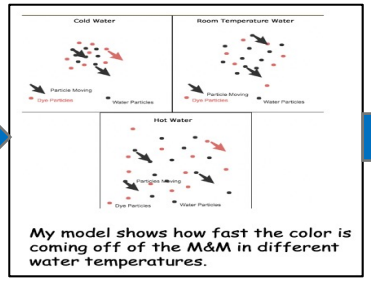


Student Explanation of the Phenomena--Model



What is Scientific Modeling?





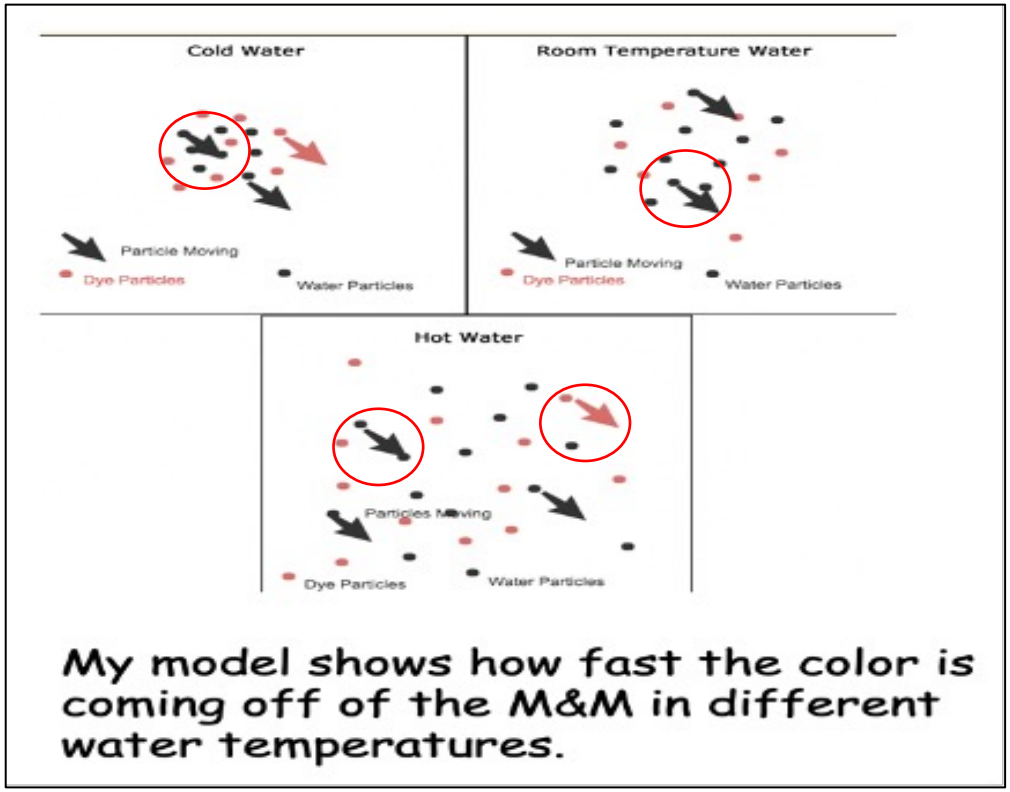
Phenomena

Mental Model

Representation (Model)

Scientific Model

Is this a scientific model?



Convolutional neural network building blocks and residual learning

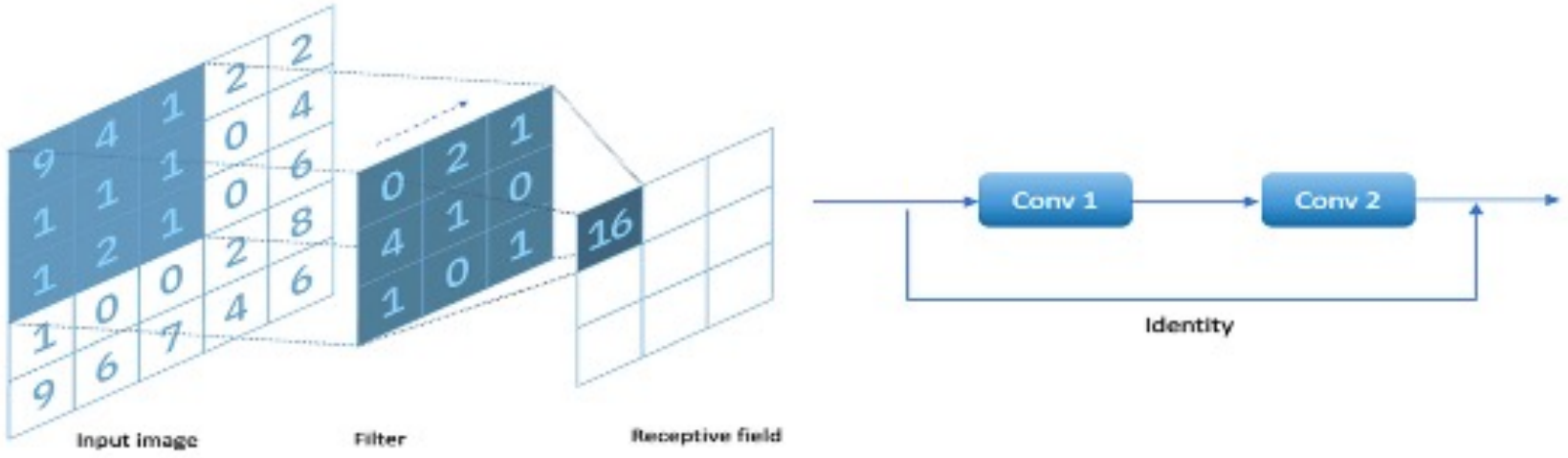


FIGURE 1 Convolutional neural network building blocks (left) and residual learning (right)

Findings—Drawing Representations

Table 2 Results of machine scoring

Task	Training Result			Validation Result		
	Accuracy	95% CI	Cohen's k	Accuracy	95% CI	Cohen's k
R1_1	0.97	(0.96, 0.98)	0.95	0.86	(0.81, 0.91)	0.76
J2_1	0.95	(0.94, 0.96)	0.92	0.79	(0.73, 0.85)	0.64
M3_1	0.96	(0.95, 0.97)	0.95	0.83	(0.77, 0.89)	0.74
H4_1	0.97	(0.96, 0.98)	0.95	0.88	(0.83, 0.93)	0.76
H5_1	0.95	(0.94, 0.96)	0.94	0.82	(0.76, 0.88)	0.71
J6_1	0.97	(0.96, 0.98)	0.96	0.89	(0.84, 0.94)	0.82

Findings: Written Descriptions

TABLE 5 Machine–human scoring agreement for written responses to modeling assessment

Task	Accuracy	95% CI	Cohen's k	S	Prec	Prev
R1-2	0.91	(0.89, 0.93)	0.74	0.97	0.91	0.75
J2-2	0.92	(0.90, 0.94)	0.81	0.96	0.93	0.70
M3-2	0.93	(0.92, 0.95)	0.81	0.97	0.94	0.77
M3-3	0.94	(0.92, 0.95)	0.76	0.98	0.94	0.82
H4-2	0.94	(0.92, 0.95)	0.81	0.98	0.95	0.78
				0.60	0.83	0.10
				0.88	0.92	0.12
H4-3	0.86	(0.83, 0.88)	0.78	0.94	0.89	0.39
				0.93	0.79	0.36
				0.39	0.93	0.25
H5-2	0.94	(0.92, 0.95)	0.87	0.96	0.93	0.56
J6-2	0.93	(0.91, 0.95)	0.85	0.95	0.93	0.58
J6-3	0.89	(0.86, 0.91)	0.62	0.98	0.89	0.78
				0.56	0.85	0.18
				0.00	NA	0.03

Abbreviations: Prec, precision; Prev, prevalence; S, sensitivity.

Correlation on Drawn Models and Written Descriptions

TABLE 3 The correlations between student performance on drawn models and on written responses

Task (N)	Drawn models (*d*)	Written description (*d*)	Pearson coefficient (*r*)	Task	Drawn models (*d*)	Written description (*d*)	Pearson coefficient
R1 (844)	R1-1 (0.18)	R1-2 (0.40)	0.306**	H4 (890)	H4-1 (−0.29)	H4-2 (0.28)	0.442**
J2 (883)	J2-1 (0.30)	J2-2 (0.46)	0.365**		H4-1 (−0.29)	H4-3 (−0.25)	0.115**
M3 (809)	M3-1 (−0.30)	M3-2 (1.13)	0.328**	H5 (834)	H5-1 (0.11)	H5-2 (1.59)	0.450**
	M3-1 (−0.30)	M3-3 (1.66)	0.339**	J6 (743)	J6-1 (−0.62)	J6-2 (−1.03)	0.438**
					J6-1 (−0.62)	J6-3 (1.51)	0.274**

Note: *d* = difficulty value calibrated from Rasch measurement. ** indicates statistically significant at level <0.01. Pearson coefficients were calculated based on raw scores.

Conclusions and Future directions

- Multi-representations as a means of developing inclusive, equitable assessments
- Insights into applying machine learning in responsive assessments
- Accuracy issues
- Validity issues
- Efficiency issues

Factors Impact Accuracy

Assessment Format: ICC = 0.28
Subject Domain: ICC = 0.42
Construct: ICC = 0.21
School Levels = 0.15
Algorithm: ICC = 0.45
Supervision: ICC = 0.17
ICC = Intraclass Correlation Coefficient

Zhai, X., Shi, L. Nehm, R. (2020). A Meta-analysis of machine learning-based science assessments: Factors impacting machine-human score agreements. *Journal of Science Education and Technology.* 30(3), 361-379.

News~~

- Special Issue in *Journal of Science Education and Technology,* Applying Machine Learning in Science Assessment
- AI for Tackling STEM Education Challenges, Frontiers in Education
- Machine Learning Applications in Educational Studies, Frontiers in Education
- Uses of Artificial Intelligence in STEM Education. (Zhai & Krajcik, Eds.), Oxford University Press.

References

- Maestrales, S., **Zhai, X.,** Touitou, I., Baker, Q., Krajcik, J., Schneider, B. (2021). Using machine learning to score multi-dimensional assessments of chemistry and physics. *Journal of Science Education and Technology*. 30(2), 239-254.
- Wilson, C., Haudek, K., Osborne, J., Stuhlsatz, M., Cheuk, T., Donovan, B., Bracey, Z., Mercado, M., & **Zhai, X**. (accepted). Using Automated Analysis to Assess Middle School Students' Competence with Scientific Argumentation. *Journal of Research in Science Teaching*.
- **Zhai, X.** (2021). Practices and theories: How can machine learning assist in innovative assessment practices in science education. *Journal of Science Education and Technology*. 69(1), 255-258.
- **Zhai, X.** (2021). Advancing automatic guidance in virtual science inquiry: From ease of use to personalization. *Educational Technology Research and Development*. 69(1), 255-258.
- **Zhai, X.,** Krajcik, J., Pellegrino, J. (2021). On the validity of machine learning-based Next Generation Science Assessments: A validity inferential network. *Journal of Science Education and Technology*. 30(2), 298-312.
- **Zhai, X.,** Shi, L. Nehm, R. (2020). A Meta-analysis of machine learning-based science assessments: Factors impacting machine-human score agreements. *Journal of Science Education and Technology.* 30(3), 361-379.
- **Zhai, X.,** Haudek, K., Shi, L., Nehm, R., Urban-Lurain, M. (2020). From substitution to redefinition: A framework of machine learning-based science assessment. *Journal of Research in Science Teaching, 57*(9), 1430-1459.
- **Zhai, X.,** Haudek, K., & Ma, W. (2022). Assessing argumentation using machine learning and cognitive diagnostic modeling. Research in Science Education. DOI: 10.1007/s11165-022-10062-w
- **Zhai, X.**, Haudek, K., Stuhlsatz, M., Wilson, C. (2020). Evaluation of construct-irrelevant variance yielded by machine and human scoring of a science teacher PCK constructed response assessment. *Studies in Educational Evaluation, 67,* 1-12.
- ***Zhai, X.*** *& Pellegrino, J. W. (In Press). Large-Scale Assessment in Science Education. In N. G. Lederman, D.L. Zeidler, & J.S. Lederman (Eds.), Handbook of Research on Science Education, Volume III (pp. xxx- xxx). New York, NY: Routledge.*
- **Zhai, X.,** Yin, Y., Pellegrino, J., Haudek, K., Shi., L. (2020). Applying machine learning in science assessment: A systematic review. *Studies in Science Education. 56*(1), *111-151.*
- **Zhai, X.,** He, P., & Krajcik, J. (2022). Applying machine learning to automatically assess scientific models. Journal of Research in Science Teaching. 1-30.