Automated scoring of science explanations for multiple NGSS dimensions and knowledge integration

Brian Riordan[1], Korah Wiley[2], Jennifer King Chen[2], Allison Bradford[2], Sarah Bichler[2],
Matthew Mulholland[1], Libby F. Gerard[2]
[1]ETS
[2]University of California-Berkeley

1. Introduction

The Next Generation Science Standards (NGSS) call for the integration of three dimensions of science learning: disciplinary core ideas (DCIs), cross-cutting concepts (CCCs), and science and engineering practices (SEPs). Science teachers can promote knowledge integration of these dimensions using constructed response (CR) formative assessments to help their students build on productive ideas, fill in knowledge gaps, and reconcile conflicting ideas. However, the time burden associated with reading and scoring student responses to CR assessment items often leads to delays in evaluating student ideas. Such delays potentially make subsequent instructional interventions less impactful on student learning. Effective automated methods to score student responses to NGSS-aligned CR assessment items hold the potential to allow teachers to provide instruction that addresses students' developing understandings in a more efficient and timely manner and can increase the amount of time teachers have to focus on classroom instruction and provide targeted student support.

In this study, we develop a set of constructed response formative assessment items that (1) call for students to express and integrate ideas across multiple dimensions of the NGSS, and (2) can be efficiently and accurately scored with automated methods. Specifically, we address the following research questions:

- How can scoring rubrics be designed to capture NGSS dimensions from knowledge integration-based science explanation CR items?
- Can automated content scoring models be trained for NGSS dimensions from such items?

We describe the design of constructed response items which formatively assess student understanding of multiple NGSS dimensions, namely, using SEPs while demonstrating integrated understanding of DCIs and CCCs.[1]

2. Methods

2.1. Background

We focus on constructed response (CR) items for formative assessments during science units for middle school students accessed via an online classroom system (Gerard & Linn, 2016; Linn et al., 2014)[2]. In past research, items that assessed NGSS performance expectations (PEs) were scored with a single

[1] See resources at https://www.nextgenscience.org/resources/ngss-appendices.
[2] https://wise.berkeley.edu/

knowledge integration (KI) rubric (Liu et al., 2016). KI involves a process of building on and strengthening science understanding by incorporating new ideas and sorting out alternative perspectives using evidence. The KI rubric used to score student short essays rewards students for linking evidence to claims and for adding multiple evidence-claim links to their explanations (Linn & Eylon, 2011). In this study, we developed items that solicit student reasoning about two or more NGSS dimensions of DCIs, CCCs, and SEPs. We scored each item for KI and NGSS "subscores" relating to the DCIs, CCCs, and practices.

2.2. Item and rubric design

We designed formative assessment items and associated rubrics for three units currently used in the online classroom system: Musical Instruments, Photosynthesis and Cellular Respiration, Solar Ovens, and Thermodynamics Challenge.

Photosynthesis and Cellular Respiration (PS). The Photosynthesis and Cellular Respiration unit engages students in exploring these processes by interacting with dynamic models at the molecular level. We designed a CR item that aligns with NGSS performance expectation MS-LS1-6[3] by asking students to express an integrated explanation of how photosynthesis supports the survival of both plants and animals. This item explicitly solicits students' ideas related to the CCC of matter cycling (i.e. change) and energy flow (i.e movement) and reads, "Write an energy story below to explain your ideas about how animals get and use energy from the sun to survive. Be sure to explain how energy and matter move AND how energy and matter change." Successful responses demonstrate proficiency in the SEP of constructing a scientific argument and reflect the synthesis of the DCIs and CCCs.

Solar Ovens (SO). The Solar Ovens unit, asks students to collect evidence to agree or disagree with a claim made by a fictional peer about the functioning of a solar oven. Students work with an interactive model where they explore how different variables such as the size and capacity of a solar oven affect the transformation of energy from the sun. We designed a CR item that addresses NGSS PE MS-PS3-3[4] and assesses students along both the CCC of energy transfer and transformation and the SEP of analyzing and interpreting data. After working with the interactive model, students respond to the CR item with the prompt: "Explain why David's claim is correct or incorrect using the evidence you collected from the model. Be sure to discuss how the movement of energy causes one solar oven to heat up faster than the other."

Thermodynamics Challenge (TC). The Thermodynamics Challenge unit asks students to determine the best material for insulating a cold beverage using an online experimentation model. We designed a CR item that aligns with the NGSS PE MS-PS3-3 and assesses student performance proficiency with the targeted DCIs in the PE, understanding of the SEP of planning and carrying out an investigation, and the integration of both of these to construct a coherent and valid explanation.The CR item prompts students to explain the rationales behind their experiment plans with the model, using both key conceptual ideas as

[3] https://www.nextgenscience.org/pe/ms-ls1-6-molecules-organisms-structures-and-processes
[4] https://www.nextgenscience.org/pe/ms-ps3-3-energy

well as their understanding of experimentation as a scientific practice: “Explain WHY the experiments you [plan to test] are the most important ones for giving you evidence to write your report. Be sure to use your knowledge of insulators, conductors and heat energy transfer to discuss the tests you chose as well as the ones you didn't choose.”

Musical Instruments and the Physics of sound waves (MI). The Musical Instruments and Physics of Sound Waves unit engages students in testing and refining their ideas about properties of sound waves (wavelength, frequency, amplitude, and pitch) and guides them applying what they learned to design and build their own instrument, a water xylophone. The CR item we designed aligns with the NGSS PE MS-PS4-2 PE and assesses students' understanding of the relationship of pitch and frequency (DCI) and the characteristics of a sound wave when transmitted through different materials (CCC). Students are prompted to distinguish how the pitch of the sound made by tapping a full glass of water compares to the pitch made by tapping an empty glass. In their answer, they are asked to explain why they think the pitches of the sound waves generated by striking the two glasses will be the same or different.

We designed three scoring rubrics for each item corresponding to two “subscores” representing the degree to which the written responses expressed PE-specific ideas, concepts, and practices and one KI score that represents how the responses integrated these elements.

NGSS subscore rubrics. To evaluate the written responses for the presence of the DCIs, CCCs, and SEPs, we designed subscore rubrics for two of the three dimensions (Table 1). Specifically, we synthesized the ideas, concepts, and practices described in the Evidence Statement documents of each targeted performance expectation to develop the evaluation criteria. We assigned each response a score on a scale of 1 to 3, corresponding to the absence, partial presence, or complete presence of the ideas, concepts, or practices.

KI score rubrics. Table 2 provides an overview of the scoring rubrics for knowledge integration. Target ideas aligned with subsets of the ideas described in the Evidence Statements. For example, the KI scoring rubrics for the Photosynthesis item evaluated written responses for the presence and linkage of five science ideas related to energy and matter transformation during photosynthesis.

2.3. Data collection

Participants were middle school students from 11 schools. Students engaged in the science units and contributed written responses to the CR items as part of pre- and post-tests. Across schools, 44% of students received free or reduced price lunch and 77% percent were non-white.

All items were coded by two researchers using the item-specific subscore and KI and rubrics described above. To ensure coding consistency, both researchers coded at least 10% of the items individually and resolved any disagreements through discussion. After the inter-rater reliability reached greater than 0.90, all of the remaining items were coded by one researcher.

2.4. Automated content scoring models

Content scoring models were built for each item and score type (knowledge integration and two NGSS dimensions). Models for each score type were trained independently on data for each item. In this way, the three models for an item formed different "perspectives" on the content of each response. Human-scored training data for the NGSS dimension models comprised either a subset of or overlapped with the training data for the KI models.

We employed an operational content scoring system that is in ongoing use in both formative and large-scale summative contexts. The system is trained on features extracted from each response. This type of "instance-based" model (cf. Horbach & Zesch (2019)) is effective when exemplar responses are not available and scores responses of any length without additional modeling complexity. The system does not consider grammatical or usage errors that do not relate to the content of each response. The feature set consists of word *n*-grams with *n* in {1, 2}, character *n*-grams with *n* in {2-5}, and syntactic dependency parse trees split into subtrees of size 3, in which a head word links two dependent words.

Nonlinear support vector regression models were trained on the extracted features and human-assigned score for each response. The models predict a class for each response representing the ordinal score.

The models were trained and evaluated by ten-fold cross-validation[5]. Within each training fold, the model hyperparameters were optimized with 5-fold cross-validation.

2.5. Evaluation metrics

To evaluate the agreement of human scores and machine scores, we report Pearson's correlation, quadratic weighted kappa (QWK), root mean squared error (RMSE), and standardized mean difference (SMD) between human and machine scores. QWK is a measure of agreement that ranges between 0 and 1 and is motivated by accounting for chance agreement (Fleiss & Cohen, 1973). SMD is measured in standard deviation units and measures the extent to which automated scores are centered on a value similar to the human scores (Williamson, Xi, & Breyer., 2012)[6].

3. Results

3.1. Score distributions

Figure 1 displays the score distributions for the NGSS subscores for each item. By examining the shape of the distributions of scores across items, we can see that students' expression of different aspects of NGSS performance expectations differed. For the Musical Instruments and Photosynthesis item, students expressed the disciplinary core ideas less than the cross-cutting concepts. For both the Solar Ovens and Thermodynamics Challenge items, students often did not explicitly articulate science concepts. The

[5] See Witten et al. (2016) chapter 5.3.
[6] See Williamson et al. (2012) for further discussion of QWK and SMD.

Thermodynamics Challenge item was particularly challenging, as many students did not express the targeted science or experimentation concepts.

Figure 2 displays the score distributions for the KI scores for each item. First, the highest score of 5 has relatively fewer responses than other score levels. Second, the score distribution for the Thermodynamics Challenge item is skewed toward score level 2. Overall, fewer students attained the higher score levels of 3 and above, indicating that this item was relatively more difficult.

3.2. Human-machine agreement

For NGSS subscore models (Table 3), those with robust score distributions (cf. Figure 1) showed good human-machine agreement, while the models trained on the most skewed data distributions showed lower levels of human-machine agreement. Specifically, Solar Ovens Science and the Thermodynamics Challenge subscore models were trained on data where about 80% of responses had the lowest score. Each of these models' agreement with the human-scored data was relatively low and significantly below the recommended 0.7 QWK threshold (Williamson et al., 2012).

The models for the KI scores showed mostly good agreement with human scores (Table 4). QWK was substantially higher than 0.7 for the Photosynthesis and Solar Ovens items. All models met the standard criterion of SMD <= 0.15 (Williamson et al., 2012).

4. Discussion

We described a set of CR items for middle-school science curricula that simultaneously assess students on expression of NGSS DCIs, CCCs, and SEPs, and the integrative linkages between each, as part of engaging in scientific explanations and argumentation. We demonstrated that human and automated scoring of such CRs for the NGSS dimensions (via independent subscores) and the integration of knowledge (via KI scores) is feasible. We demonstrated that automated scoring can be developed with promising accuracy.

Results showed that students often scored at the lowest levels of all three rubrics, which increased skewness in the datasets and likely contributed to reduced model accuracy. This finding potentially reflects a need to revise the CR item prompts and associated unit supports to reliably elicit the targeted concepts from more students. Automated scoring research will explore more robust methods for learning scoring models from less data in formative settings, especially from highly skewed score distributions, while continuing to provide accurate scoring.

Our findings demonstrate the ability to both develop and automatically score NGSS-aligned CR assessment items. With further refinement, we can provide teachers with both the instructional and technological assistance they need to effectively and efficiently support their students to demonstrate the multidimensional science learning called for by the NGSS.

Acknowledgements

This material is based upon work supported by the National Science Foundation under Grant No. 1812660. 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 National Science Foundation.

References

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

Gerard, L. F., & Linn, M. C. (2016). Using Automated Scores of Student Essays to Support Instructor Guidance in Classroom Inquiry. *Journal of Science Instructor Education*, *27*(1), 111–129.

Horbach, A., & Zesch, T. (2019). The influence of variance in learner answers on automatic content scoring. *Frontiers in Education*, *4*, 28.

Linn, M. C., & Eylon, B.-S. (2011). *Science Learning and Instruction: Taking Advantage of Technology to Promote Knowledge Integration*. New York: Routledge.

Linn, M. C., Gerard, L., Ryoo, K., McElhaney, K., Liu, O. L., & Rafferty, A. N. (2014). Computer-guided inquiry to improve science learning. *Science*, *344*(6180), 155–156.

Liu, O. L., Rios, J. A., Heilman, M., Gerard, L., & Linn, M. C. (2016). Validation of Automated Scoring of Science Assessments. *Journal of Research in Science Teaching*, *53*(2), 215–233.

Williamson, D. M., Xi, X., & Breyer, F. J. (2012). A framework for evaluation and use of automated scoring. *Educational Measurement: Issues and Practice*, *31*(1), 2–13.

Witten, I. H., Frank, E., Hall, M. A., & Pal, C. J. (2016). *Data Mining: Practical machine learning tools and techniques*. Morgan Kaufmann.

Table 1: NGSS performance expectations (PE) and targeted components: disciplinary core idea (DCI), cross-cutting concept (CCC), and science and engineering practices (SEP) targeted by each item.

Item	PE	DCI	CCC	SEP
Photosynthesis	MS-LS1-6	x	x	
Solar Ovens	MS-PS3-3		x	x
Thermodynamics Challenge	MS-PS3-3	x		x
Musical Instruments	MS-PS4-2	x	x	

Table 2: Knowledge integration (KI) scoring rubrics and example responses.

Score	Description	Item			
		Photosynthesis	Solar Ovens	Thermodynamics Challenge	Musical Instruments
1	Off-task	she can do it herself	David's claim is ... because ...idk	I dont know.	it is just how it works
2	On-task but lacks normative ideas	Energy comes from the sun. Energy moves trough the air. Energy goes to any living thing thats need it. Energy is realesed by the suns light shines down and thats how its relased	he is correct because when you look on how fast it heats up mostly all the heat energy was there.	I chose these items because I feel those items may be used in my water bottle that keep my water cold.	It will always stay the same because the spoon is the same
3	Partial link - normative ideas without any valid links between normative ideas	I think that the rabbit gets energy from its food that it ate before moving. Energy got to the food by sun rays that come from the sun if the rabbit ate a plant of some sort.The energy goes in to the plants cells.	David's claim is wrong because the wide short was a bigger target for the sunrays to hit so more heat got into the box.	The best materials to keep water cold are the conductors(alumi num, copper and steel)because they are better conductors than insulators.	The pitch is lowered by the water in the glass. This means that the glass full of water will have a lower pitch than the glass that is empty.
4	Full link - one valid link between normative ideas	The energy comes from the glucose in the plant and the rabbit ate the plant so the rabbit gets the energy from the plant. The energy moves through the chloroplast in	David's claim is was completely wrong because the skinny long box opening was too small not allowing sun light to go inside. That why its better to use the wide box	The tests I chose are the most helpful for determining the best material to keep a cold beverage cold because I think that both of the materials I chose	They will be different because when you had a more dense medium like water into a cup instead of less dense air the sound gets caught more between the particles resulting in

		which it makes glucose. The energy is stored in the plants chloroplast, so then it can make the energy for the whole plant, so that it can grow tall and big. The energy changes from light energy to chemical energy.	because it has more of a bigger window for the sun light to in.	are the best insulators out of the options given.	a lower pitch.
5	Complex link - multiple valid links between normative ideas	The rabbit eats the leaf in the classroom. The leaf already has stored energy. The leaf's energy came from this: First, light energy, water, and carbon dioxide are entered into the chloroplast to convert into oxygen and glucose. The glucose has energy in it, but it's not usable yet. To make it usable, the glucose and oxygen go into the Mitochondria where water, carbon dioxide, and USABLE chemical energy come out of the plant to be stored everywhere inside the plant. Then,	David's claim is incorrect because based on the information I collected form the computer model, the short and wide increased its temperature. The movement of energy causes one solar oven to heat up faster than the other because the wide opening gap lets the infrared radiation, in the inside, becomes heat.	I chose these because from the previous tests I learned that any type of metal is most likely a conductor, not an insulator. So I picked materials other than metal. I also chose in a hot room for all of them because that would put them all to the test to see if they keep the liquid cold even in a hot room.	I think that the glass full of water would have a lower pitch because the cup would have more mass which would make the cup harder to vibrate which makes the sound so it would have a lower pitch. The sound waves would also have a longer wavelength and would have a lower frequency.

when the rabbit
eats the leaf, it
takes some of that
stored, usable,
chemical energy
which gives the
rabbit energy.

Figure 1: Score distributions for NGSS subscore items (percentages).

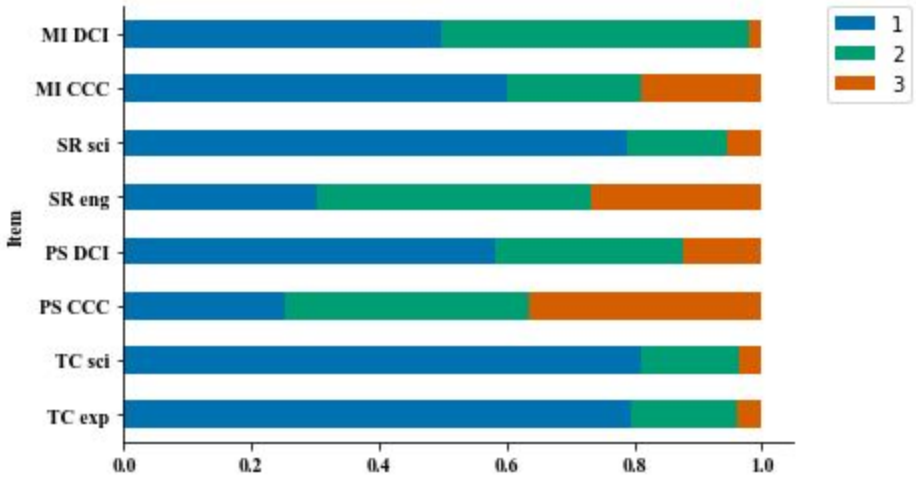

Figure 2: Score distributions for Knowledge Integration (KI) items (percentages).

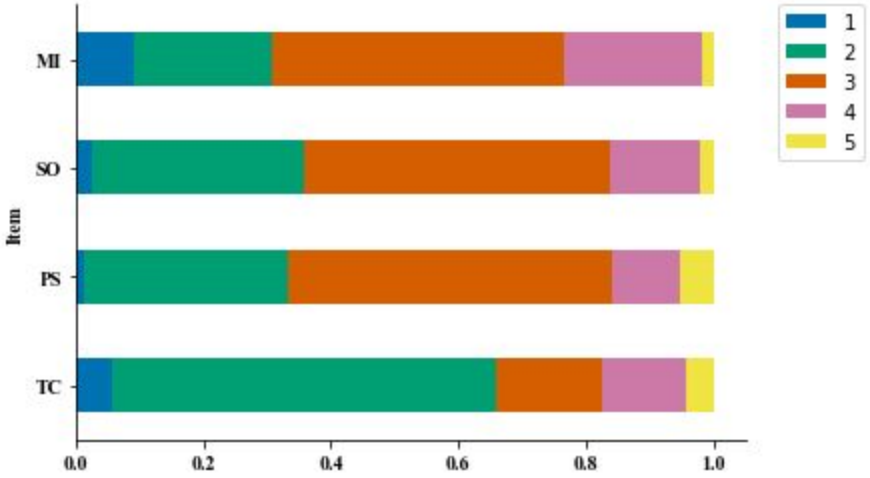

Table 3: Human-machine agreement for NGSS subscore models. QWK = quadratic-weighed kappa, RMSE = root mean squared error, SMD = standardized mean difference.

Item		Correlation	QWK	RMSE	SMD
Musical Instruments	CCC	0.731	0.727	0.555	-0.035
	DCI	0.756	0.756	0.369	0.006
Photosynthesis	CCC	0.726	0.683	0.537	0.007
	DCI	0.751	0.707	0.470	-0.029
Solar Ovens	SEP: engineering	0.735	0.662	0.487	-0.006
	SEP: science	0.619	0.529	0.402	-0.052
Thermodynamics Challenge	SEP: experimentation	0.572	0.478	0.425	-0.004
	DCI: science	0.491	0.463	0.443	-0.045

Table 4: Human-machine agreement for KI score models. QWK = quadratic-weighed kappa, RMSE = root mean squared error, SMD = standardized mean difference.

Item	Correlation	QWK	RMSE	SMD
Musical Instruments	0.761	0.761	0.632	0.017
Photosynthesis	0.829	0.793	0.467	0.027
Solar Ovens	0.772	0.739	0.529	-0.005
Thermodynamics Challenge	0.678	0.646	0.715	-0.013