USING AUTOMATED ANALYSIS TO ASSESS MIDDLE SCHOOL STUDENTS' COMPETENCE WITH SCIENTIFIC ARGUMENTATION

Paper presented at the Annual Meeting of the National Council on Measurement in Education (NCME).
Toronto, ON; April 2019

Kevin C. Haudek
Marisol Mercado Santiago
Michigan State University

Christopher D. Wilson
Molly A. M. Stuhlsatz
Brian Donovan
Zoë Buck Bracey
April Gardner
BSCS Science Learning

Jonathan Osborne
Tina Cheuk
Stanford University

INTRODUCTION

Argumentation is fundamental to both science and science education, to the extent that the history of science has been described as "the history of vision and argument" (Crombie, 1994, p3). This perspective is reflected in the Framework for K-12 Science Education (National Research Council, 2012), and the resultant Next Generation Science Standards (NGSS; Lead States, 2013) where argumentation is presented as one of eight fundamental science and engineering practices through which students learn the disciplinary core ideas and crosscutting concepts of science. However, it is widely acknowledged that these new standards will only have a meaningful impact if they are accompanied by high quality assessments that are closely aligned with this three-dimensional vision for teaching and learning science (NRC, 2012, Pellegrino et al., 2013). Such assessments demand a move away from reliance on the efficiency and affordability of multiple choice items, and towards the use of more authentic tasks aligned to NGSS performance expectations. In the case of argumentation in particular, the performance tasks will commonly require significant student written work, which is expensive and resource intensive to score. However, efficiency and affordability remain critical components of new assessment systems, whether for research and evaluation purposes, or for broad scale state and federal measures. We therefore need new, inexpensive approaches to scoring assessments that measure three-dimensional science learning. Achieving this goal is important because "assessments operationalize constructs" (William, 2010) and if there are no assessments of argumentation that assess the performance expectations of the NGSS, it is doubtful that it will be enacted as a practice in the classroom.

Meanwhile, as educators face the emerging challenges associated with measuring new constructs aligned with the NGSS, assessments at all levels are increasingly moving online. For example, PARCC and Smarter Balanced, two federally funded testing consortia, both use computer-based assessment systems. There are also a number of research groups exploring how simulations and digital learning environments can be used to measure three-dimensional learning (NRC, 2014). However, such assessments are still limited by their inability to score student written work efficiently. Further, if the writing of students during online learning experiences could be scored in real time, both formative information could be supplied to the students, and an adaptive experience provided.

This project explores whether we can use automated lexical analysis and machine learning techniques to develop valid and reliable constructed response measures of student scientific argumentation that can be administered and scored at scale. The goal is to develop accurate and reliable scoring models that are able to score written responses at levels equal to human expert scorers, and to accurately place students on a learning progression for argumentation. By developing computer scoring models that accurately replace the time-consuming process of expert human scoring of students' writing on this task, the resulting instruments can be made available online, and can provide rapid argumentation scores for research and evaluation purposes, as well as formative feedback for teachers.

THEORETICAL FRAMEWORK

In this work we draw on an extensive program of research conducted in the field of automated lexical analysis and the body of work conducted on argumentation in science education.

Constructed Response Assessments and Automated Analysis

Constructed response (CR) assessments, in which students use their own language to demonstrate knowledge, are widely viewed as providing greater insight into cognition than closed form (e.g., multiple choice) assessments. In the past, financial and time constraints have made constructed response assessments significantly more challenging to execute in large-enrollment courses than multiple-choice assessment. But today, advances in both technology and measurement research now make it feasible to apply these techniques in instructional settings with the potential to have substantial educational impact (Ha, Nehm, Urban-Lurain, & Merrill, 2011; Haudek, Prevost, Moscarella, Merrill, & Urban-Lurain, 2012; Moharreri, Ha and Nehm, 2014). These studies have shown that 1) it is possible to create computerized scoring models that predict human scoring with inter-rater reliability (IRR) measures approaching that of well-trained expert raters 2) reveal the heterogeneity of student thinking that cannot be revealed by traditional multiple-choice items; and 3) capture, represent, and analyze this multidimensional information in a variety of ways that provide instructors richer insights into student thinking.

Recent work in automated analysis of scientific argumentation and explanation has shown promising results. Liu et al. (2014) examined the scoring accuracy of c-rater when scoring CR concept-based items with multiple levels. They implemented four conceptual CR science items at the middle school level, which were scored using a 5-point holistic rubric. As part of the study, they transformed the holistic levels into analytical rubrics to implement c-rater (Liu et al., 2014). Overall, computer model performance showed between moderate and good agreement with human scores. Linguistic diversity of middle school students, and pronoun resolution (multiple ways to use the same pronoun in the same sentence), and their small sample-size were some of the sources of error that they identified (Liu et al., 2014). Finally, they also commented on the design and implementation challenges of capturing holistic ratings via a series of analytical rubrics.

Mao et al. (2018) utilized c-rater-ML, which uses support vector regression algorithm, to score and provide feedback on students' short responses to scientific argumentation items. Their formative assessment evaluates four components of scientific argumentation based on Lee et al. (2014) and the framework of Toulmin (1958). After predicting scores, the automated system generates feedback the response is correct or provides advice to revise it (Mao et al., 2018). Some of the challenges identified by this work included degradation of model performance on several items and the tendency of the computer model to assign lower scores to shorter responses. In addition, there were few responses at higher levels of argumentation which impacted the ability to build a robust scoring model. On the other hand, their analyses suggested that the feedback provided by the system did lead to improvements in student written arguments (Mao et al., 2018).

Argumentation and Learning Progressions

Over the past several years, there has been a great deal of research around teaching and assessing scientific argumentation, the majority of which relies on Toulmin's (1958) argument structure for analysis of student discourse (e.g., Lee et al., 2013; Cavagnetto, 2010; Osborne, 2010). Toulmin (1958) posited that although there are field-specific elements to every argument, there are structural elements that can be found universally across disciplines. These six field-invariant structural elements are: claim, data, warrant, backing, qualifier, and of rebuttal. We posit that in addition to the construction of an argument composed of Toulmin's elements, proficiency in the practice of argumentation requires students to engage in formulating rebuttals and engage in the act of critique. To perform a critique, students must be able to construct a rebuttal that would explain why the reasoning in a given argument is flawed, by comparing and contrasting the relative merits of two arguments or by constructing an argument for why some evidence has higher epistemic validity than other evidence.

Argumentation is fundamental to both science and science education, and is now a feature of the Common Core State Standard for Language Arts and the framework for Assessment of Mathematics in the OECD PISA tests in 2018. In the U.S. Next Generation Science Standards (NGSS; Lead States, 2013), argumentation is presented as one of eight science and engineering practices through which students learn the core ideas and crosscutting concepts of science. Scientific argumentation – including both construction and critique – is a competency that draws on diverse knowledge and practices (OECD, 2012), including specific domain content (Osborne, 2010), rhetorical knowledge about the conventions of argumentation (Kelly & Takao, 2002), and epistemic commitment to evidence as the basis of belief (Sandoval, 2003). These features make the development of assessments for argumentation particularly complex. Learning progressions are tools that can be used to guide the development of such assessments capturing the complexity of the domain by outlining possible cognitive trajectories that students might follow as they develop a more sophisticated understanding of a core concept. The National Research Council (2007) report *Taking Science to School*, makes the case for such empirically based maps of "successively more sophisticated ways of learning about a topic that can follow and build on one another as children learn and investigate a topic over a broad span of time" (p. 230).

In their research to date, Osborne and his team have developed a learning progression for argumentation in the context of the structure of matter (Osborne, Henderson, MacPherson, & Szu, 2016). The construct map in Table 1 provides a summary of the learning progression. The map for argumentation is innovative in that it includes critique, which is essential for scientific argumentation as the construction of knowledge is a dialectic between construction and critique (Ford, 2008). In other words, being able to explain why an idea is flawed is as important as being able to explain why it is right. Empirical work to date has shown the construct to be psychometrically uni-dimensional, and that it supports the distinction between the two columns (Constructing Arguments vs. Critiquing Arguments) in that the rows shown in Table 1 have been shown to have different difficulties (Yao, 2013). Like all construct maps, it defines a continuum of understandings, providing a "coherent and substantive definition for the content of the construct" (Wilson, 2005).

Level	Constructing	Critiquing	Description
0a	Constructing a claim		Student states a relevant claim.
0b		Identifying a claim	Student identifies another person's claim.
0c	Providing evidence		Student supports a claim with a piece of evidence.
0d		Identifying evidence	
1a	Constructing a warrant		Student constructs an explicit warrant that links their claim to evidence.
1b		Identifying a warrant	Student identifies the warrant provided by another person.

1c	Constructing a complete argument		Student constructs a synthesis between the claim and the warrant.
1d	Providing an alternative counter argument		Student offers a counterargument as a way of rebutting another person's claim.
2a	Providing a counter-critique		Student critiques another's argument.
2b	Constructing a one-sided comparative argument		Student makes an evaluative judgment about the merits of two competing arguments
2c	Providing a two-sided comparative argument		Student provides an evaluative judgement about two competing arguments
2d	Constructing a counter claim with justification		Student explicitly compares and contrasts two competing arguments, and an argument as to why it is superior to each of the previous arguments.

Table 1. Scientific Argumentation Construct Map, (Osborne et al., 2016)

Bringing together the domains of argumentation, learning progression, and automated analysis, our primary **research question** is therefore:

> How can automated lexical analysis and machine learning techniques be applied to developing an efficient, valid, and reliable measure of students' placement on a learning progression for argumentation?

With the following secondary research questions:

- Can we develop automated computer scoring models of students' explanation and argumentation responses that closely correlate with expert human coding?
- What feedback can we provide from the automated computer scoring that will facilitate both quantitative research and evaluation, and formative feedback to instructors and students?

METHOD

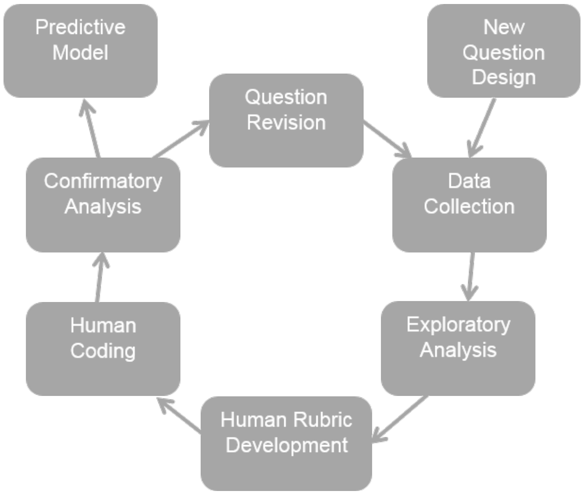


Figure 1 Question Development Cycle (QDC)

Our approach to developing and validating assessments is captured by the Question Development Cycle (QDC) shown in Figure 1 (Urban-Lurain et al., 2013). In the first stage of the QDC, we *Design New Questions* to measure thinking about important constructs. *Data Collection* is typically done by administering the questions online where respondents can enter their answers. *Exploratory Analysis* is performed using a mix of traditional and computer-enabled qualitative analysis, such as lexical analysis software to extract key scientific concept usage context or text mining to identify patterns in student writing. These terms, concepts and patterns may aid in *Rubric Development*. For constructs that already have well defined coding rubrics and can use machine learning algorithms, the *Exploratory* and *Rubric Development* stages are largely bypassed. We use both analytic and holistic rubrics for *Human Coding* of responses. During *Confirmatory Analysis* the *Lexical Resources* are used as dependent variables in statistical and machine classification techniques to predict expert human coding. The final product of the QDC is a *Predictive Model* that can be used to automate the scoring of new responses.

Our work on measuring argumentation is based on the work of Osborne et al., 2016, who have developed a range of assessments for argumentation in the context of matter, complete with rubrics, validated an argumentation learning progression, and collected a body of expert-scored student responses. In Phase 1 of the project we used machine learning techniques to explore how computers could be trained to predict the expert scores of existing items and rubrics developed as part of the learning progression, and to learn how different item types and rubrics lend themselves to this approach. In Phase 2 of the project we engaged in item revision, new data collection, human scoring, and machine classification, to gain greater resolution and sensitivity of the measure, and to use the scores to reliably place students along the learning progression.

Below in Appendix A, we present an exemplar question developed for testing – sugar in water and its associated rubric. To date, three sets of questions have been developed with their scoring guides. Each has been the product of an extensive process of development consisting of cognitive think alouds, testing on MTurk and refinement of the questions and the scoring rubrics. Two approaches have been taken to scoring. The first is an analytic approach which seeks to define the elements of an appropriate answer for a particular level of the learning progression. Each question is examined for its linguistic and cognitive demands by examining what receptive processing is necessary and what the student is required to produce. Exemplar model answers help frame the discourse around the rubric which was then tested with small sample of MTurk respondents. These analytic elements then are combined into a single holistic score, which represents the student ability at the targeted level.

Sample: The data we present here were drawn from three samples of students used in two phases of development and analysis. The initial data was an initial set of 246 8th grade students drawn from a mid-sized urban school district in Northern California that informed the iterative design and revision work of item and rubric development. The second phase of data was drawn from two sources, a private independent school (grades 5-8) with approximately 100 student responses and a set of five middle school (grades 6-8) science classrooms with approximately 900 student responses from a public school district in the California Bay Area, totaling about 1000 responses that were then used in the machine learning modeling and analysis.

Human Coding: For phase II, two human coders were trained on a set of items and associated rubrics. The two coders went through multiple rounds of training using a random subset of 150 student responses to each item. Training rounds were iterated until interrater reliability (Cohen's kappa) between the coders was at least 0.6 on each rubric component. Then, the remaining data set was split into two subsets, including an overlapping set of 150 responses, and each coder scored one subset independently. The final Cohen's kappa value was calculated using the overlapping set of 150 responses coded independently; any scoring disagreements in this subset of responses was resolved by a third scorer and reported as the consensus score.

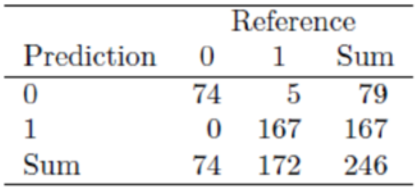
Machine Learning: We employed a supervised machine learning text classification approach to assign student written responses a score (see Aggarwal & Zhai, 2012). During our machine learning process, each individual student response is treated as a document and the bins in the scoring rubric are treated as classes. The computerized scoring system then generates predictions on whether each given document is a member of each class. To generate these predictions we use an ensemble of eight individual machine learning algorithms (Jurka et al. 2012). The algorithms in our ensemble are: support vector machines, supervised latent dirichlet allocation, logitboost, classification trees, bagging classification trees, random forests, penalized generalized linear models, and maximum entropy models. Each algorithm votes independently on the categorization of a particular document, but these individual votes are combined to make a final categorization. These individual votes are combined using a stacking approach, where individual votes are weighted in the final categorization prediction according to individual model performance. Each individual algorithm is trained using a corpus of human-scored student responses. The computer model is generated using a 10-fold cross-validation approach. A computer model is generated using consensus scores, when possible, on responses for each analytic rubric, resulting in the production of a set of predictive scoring models. For this study, we assigned holistic scores to student responses using a combination of analytic scores. Once a computer model performs at an acceptable level of performance, it can be used to assign scores to a testing set of student responses. We have built a set of web-based applications that perform the necessary text parsing and supervised machine learning algorithms employed in this study. The R code is available at https://github.com/BeyondMultipleChoice/AACRAutoReport.

RESULTS

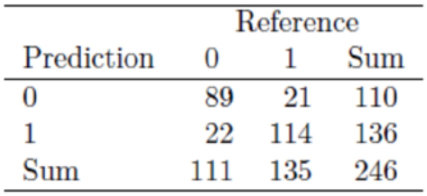
Phase 1 Findings

In our phase 1 analyses, we had mixed success in obtaining human-computer agreement. The two confusion matrices below provide an example of a high level of agreement obtained for some items, and a lower level of agreement found for others.

Kappa
= 0.953



	Reference		
Prediction	0	1	Sum
0	74	5	79
1	0	167	167
Sum	74	172	246



	Reference		
Prediction	0	1	Sum
0	89	21	110
1	22	114	136
Sum	111	135	246

Kappa
= 0.647

Table 2. High and low levels of agreement obtained from analysis of the Phase 1 data.

Our team has explored several possible explanations for this variation in agreement, including a) the item’s scientific context, b) the level of the item in the learning progression, c) the level of human-human agreement achieved, d) the amount of branching in the survey logic of the item, and e) characteristics of the sample.

Phase 2 Findings

In light of the phase 1 findings, items and rubrics were revised and new data collection was conducted Fall 2018 in the California Bay Area middle schools.

As an example of our work, we will report on findings from two items from a single context: sugar dissolving in water. We have developed computer scoring models using 775 student responses for an item which targets level 1c (*Constructing a complete argument*) of the learning progression (see Appendix A for item, task audit and coding rubric). Scoring for this item requires identification of all three elements of scientific argumentation: claim, warrant and evidence.

Our coding rubric for this item contains a single component for the claim; a single component for evidence, which could be stated in two different ways and three possible components for different warrants or reasoning. In earlier forms of the rubric contained more individual components for evidence and different reasoning. However, during and after human coding, we decided to combine some coding categories due to low occurrence in the student responses and/or highly overlapping language used when expressing similar ideas. Here is an example student response which contains all three elements of a complete argument:

> *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.*

Using our coding rubric, this response would have been scored as having a valid claim (sugar gets dissolved), evidence (you can no longer see it) and reasoning (particles break up) all present. We have developed a scoring model for each of the analytic components found in the final coding scheme.

Component	Cohen's kappa (Human-Human)	Cohen's kappa (Human-Computer)	95% CI Accuracy Human-Computer
Claim	0.869	0.719	0.900-0.939
Evidence	0.834	0.809	0.904-0.943
Reasoning – C1	0.644	0.741	0.953-0.979
Reasoning – C2	0.782	0.775	0.906-0.944
Reasoning – C3	0.921	0.911	0.950-0.977

Table 3. Interrater reliabilities between human coders and computer model performance for seven rubric components for Sugar – Level 1c item.

Overall, we have very good model performance across all analytic components. Overall, each component had accuracy ranges at or above 0.9 and Cohen's kappa all above 0.7, including two components above 0.8 which is considered as near perfect agreement. Results that show that models for the *Claim* component and one of the *Reasoning* components are the lowest performers, although the models still have inter-reliability measures of greater than 0.7. In the case of *Reasoning- C1,* the computer model showed better inter-rater reliability on the C1 warrant than the two human coders. Conversely, the performance of the model for identifying the *Claim* component was below the ability of human coders to detect the same construct.

We have combined these analytic component scores into a single, four-level, holistic score for the targeted learning progression level, then used these holistic scores to generate a predictive model. The holistic score represents whether a student has a complete and valid argument, a partially complete argument or no argument at all. A score of 3 would represent a complete and valid argument; a score of 0 would represent off-task or no attempt at argumentation. Here is an example student response that would have a holistic score of 2; it is a partially complete argument because it is missing a relevant piece of evidence:

The sugar has been dissolved when the sugar was mixing in with the water.

The results of the predictive model for these holistic scores are given below.

		Human generated holistic score				
		0	1	2	3	Sum
Computer	0	**4**	3	0	0	7
predicted	1	23	**361**	75	13	475
holistic	2	0	17	**190**	52	259
score	3	0	7	1	**26**	34
	Sum	27	388	269	91	775

Table 4. Confusion matrix between human and computer holistic scores for Sugar-1c item. Bolded numbers on the diagonal represent scores in agreement between human and computer.

Overall, the holistic model had an accuracy of 0.75 (0.718-0.78 accuracy 95% confidence interval) and a Cohen's kappa of 0.562. Overall, the holistic model tends to under-predict students' argumentation ability. Specifically, the model has the most difficulty predicting a complete argument (i.e. level 3); this is likely due to the lower model performance on the *Claim* analytic component (see Table 3), as having a valid claim is essential for a complete argument. This may also be the cause of the underprediction of level 2 responses. Additional responses at level 3 would be helpful in developing the computer model.

We have done similar work for another item in the Sugar context, but targeted at the 0d level (*Providing evidence*) of the learning progression. For this item, students need to specifically identify the relevant evidence without conflating evidence with claim (see Appendix B for item, task audit and coding rubric). Therefore, there were only two relevant components during coding. We coded and used a total of 763 student responses to train a computer model.

Component	Cohen's kappa (Human-Human)	Cohen's kappa (Human-Computer)	95% CI Accuracy Human-Computer
Evidence	0.940	0.953	0.981-0.996
Claim	0.727	0.772	0.861-0.908

Table 5. Interrater reliabilities between human coders and computer model performance for two rubric components for Sugar – Level 0d item.

In order to generate a holistic code, we combined these two components to identify responses that included evidence and did not include a claim. Therefore, the holistic score was dichotomous (0 or 1) to reflect the ability of a student to specifically identify a relevant piece of evidence. We created a computer model to predict this holistic score, and the results are given below.

		Human generated holistic score		
		0	1	Sum
Computer predicted holistic score	0	**376**	43	419
	1	27	**317**	344
	Sum	403	360	763

Table 6. Confusion matrix between human and computer holistic scores for Sugar-0d item.

The holistic model performance had an accuracy of 0.908 (0.886-0.928 accuracy 95% confidence interval) and a Cohen's kappa of 0.815. This model had a slight tendency to make false negative predictions. Overall, this holistic model had better performance metrics than the model created for the Sugar-1c item, likely because there were fewer components required and the construct (Identifying evidence) is more straightforward to identify. We have found similar results for items in another disciplinary context at the same learning progression levels. We are continuing this work to develop predictive models for items which target higher levels of the learning progression, which constitute more complex argumentation skills (e.g., providing a comparative argument).

DISCUSSION AND CONCLUSIONS

Educational reforms demand assessments move away from reliance on the efficiency and affordability of multiple-choice items, and towards the use of more authentic tasks aligned to broader skills and performance expectations. We therefore need new, inexpensive approaches to scoring assessments. Achieving this goal is important because "assessments operationalize constructs" (William, 2010) and if no assessments of important constructs exist, it is doubtful that they will be valued or enacted as a practice in the classroom. The work described here addresses these issues by applying machine learning techniques to efficiently measure students' ability to engage in scientific argumentation. By developing computer scoring models that accurately replace the time-consuming process of expert human scoring, the resulting instruments can provide rapid scores for formative feedback and research purposes.

We have attempted to design a set of items, scoring rubrics and associated computer scoring models aligned with an empirically validated learning progression for scientific argumentation. So far our results show that we have been able to map student written arguments in several contexts to a learning progression, which is useful to describe how students develop in this area. This is important for returning interpretable and summary class-wide statistics to teachers. We have also been able to generate holistic scores of student ability using a series of analytic rubrics. There are some challenges to identifying the necessary and relevant "pieces" to a holistic score (for example, see Liu et al, 2014). However we have employed task audits (see Part 2 in Appendices A & B) during item development, which help align rubrics, expected responses and identify components necessary for a holistic score. By adopting such an approach we have been able to produce a rubric which can lead to reliable scoring between human coders computer and addresses some of the challenges of identifying and combining the relevant components for a holistic score. This helps overcome an issue we encountered in Phase I of this project with low human-human agreement on holistic scores of student responses, especially on some of the complex argumentation practices. Finally, computer scoring models for the analytic components have shown good performance with little model tuning or pre-processing efforts on text parsing. The ensemble approach employed in this study has demonstrated good to very good performance over several item contexts and rubric components and may overcome some of the limitations of employing only a single classification algorithm. Although further testing of model performance will be conducted to determine model degradation measures.

ACKNOWLEDGEMENT

This material is based upon work supported by the National Science Foundation under Grant No. DUE 1561159. 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

Aggarwal C.C., Zhai C. (2012) A Survey of Text Classification Algorithms. In: Aggarwal C., Zhai C. (eds) Mining Text Data. Springer, Boston, MA

Cavagnetto, A., Hand, B. M., & Norton-Meier, L. (2010). The Nature of Elementary Student Science Discourse in the Context of the Science Writing Heuristic Approach. *International Journal of Science Education, 32*(4), 427 - 449.

Crombie, A. C., (1994). *Styles of scientific thinking in the European tradition: The history of argument and explanation especially in the mathematical and biomedical sciences and arts (Vol. 3).* London: Duckworth.

Ford, M. J. (2008). Disciplinary authority and accountability in scientific practice and learning. *Science Education, 92*(3), 404-423.

Ha, M., Nehm, R. H., Urban-Lurain, M., & Merrill, J. E. (2011). Applying computerized-scoring models of written biological explanations across courses and colleges: prospects and limitations. *CBE—Life Sciences Education, 10*(4), 379-393.

Haudek, K. C., Prevost, L. B., Moscarella, R. A., Merrill, J., & Urban-Lurain, M. (2012). What are they thinking? Automated analysis of student writing about acid–base chemistry in introductory biology. *CBE—Life Sciences Education, 11*(3), 283-293.

Jurka,TP., Collingwood, L., Boydstun, AE., Grossman, E., Van Atteveldt, W. (2012). RTextTools: Automatic Text Classification via Supervised Learning. R package version 1.3.9. http://CRAN.R-project.org/package=RTextTools

Kelly, G., & Takao, A. (2002). Epistemic Levels in Argument: An Analysis of University Oceanography Students' Use of Evidence in Writing. *Science Education, 86*, 314-342.

Lee, O., Quinn, H., & Valdés, G. (2013). Science and Language for English Language Learners in Relation to Next Generation Science Standards and with Implications for Common Core State Standards for English Language Arts and Mathematics. *Educational Researcher, 42*(4), 223-233

Lee, H.-S., Liu, O. L., Pallant, A., Roohr, K. C., Pryputniewicz, S., & Buck, Z. E. (2014). Assessment of uncertainty-infused scientific argumentation. *Journal of Research in Science Teaching*, *51*(5), 581–605. https://doi.org/10.1002/tea.21147.

Liu, O. L., Brew, C., Blackmore, J., & Gerard, L. (2014). Automated scoring of constructed-response science items: Prospects and obstacles. *Educational Measurement-Issues and Practices*, *33*(2), 19–28. https://doi.org/10.1111/emip.12028

Mao, L., Liu, O. L., Roohr, K., Belur, V., Mulholland, M., Lee, H.-S., & Pallant, A. (2018). Validation of automated scoring for a formative assessment that employs scientific argumentation. *Educational Assessment*, *23*(2), 121–138. https://doi.org/10.1080/10627197.2018.1427570

Moharreri, K., Ha, M., & Nehm, R. H. (2014). EvoGrader: an online formative assessment tool for automatically evaluating written evolutionary explanations. *Evolution: Education and Outreach, 7*(1), 1-14. doi:10.1186/s12052-014-0015-2

National Research Council. (2007). *Taking Science to School: Learning and Teaching in Grades K*-8. Washington DC: National Research Council.

National Research Council. (2012). *A Framework for K-12 Science Education: Practices, Crosscutting Concepts, and Core Ideas*. Washington, D.C.: The National Academies Press.

National Research Council (2014). Developing assessments for the Next Generation Science Standards. Washington, DC: National Academies Press.

NGSS Lead States. (2013). *Next Generation Science Standards: for States, By States*. Washington, DC: The National Academies Press.

Osborne, J. F., Henderson, B., MacPherson, A., Szu, E., Wild, A., & Shi-Ying, Y. (2016). The Development and Validation of a Learning Progression for Argumentation in Science. *Journal of Research in Science Teaching, 53*(6), 821-846.

Osborne, J. F. (2010). Arguing to Learn in Science: The Role of Collaborative, Critical Discourse. *Science*, *328*, 463-466.

Pellegrino, J. W., Wilson, M. R., Koenig, J. A., & Beatty, A. S. (Eds.). (2013). *Developing Assessments for the Next Generation Science Standards*. Washington DC: National Academies Press.

Sandoval, W. A. (2003). Conceptual and epistemic aspects of students' scientific explanations. *Journal of the Learning Sciences*, *12*(1), 5-51.

Toulmin, S. (1958). *The Uses of Argument*. Cambridge: Cambridge University Press.

Urban-Lurain, M., Prevost, L., Haudek, K. C., Henry, E. N., Berry, M., & Merrill, J. E. (2013). *Using computerized lexical analysis of student writing to support Just-in-Time teaching in large enrollment STEM courses.* Paper presented at the 2013 IEEE Frontiers in Education Conference (FIE).

Yao, X. M. (2013, March). Automated Essay Scoring: A Comparative Study. In Applied Mechanics and Materials (Vol. 274, pp. 650-653).

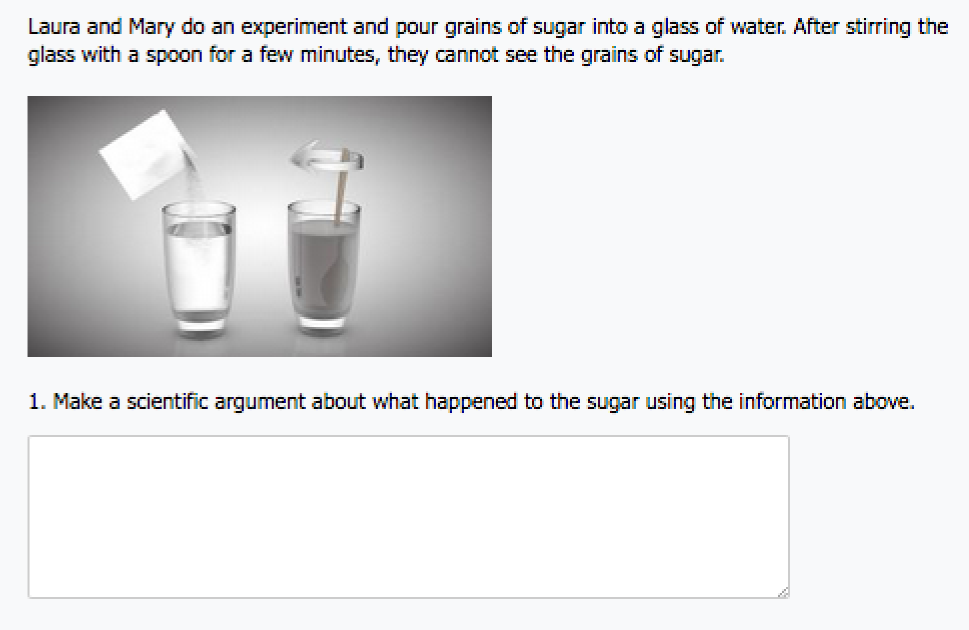
William, D. (2010). What Counts as Evidence of Educational Achievement? The Role of Constructs in the Pursuit of Equity in Assessment. *Review of Research in Education, 34*, 254-284.

Wilson, M. (2005). *Constructing Measures: An item response modeling approach*. Mahwah, NJ: Lawrence Erlbaum Associates.

Appendix A

Name: Rubric_S1_1c_A1B1C3

Part 1: Task



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.

Part 2: Audit of task

LP level: 1c, Constructing a complete argument **Definition:** Student makes a claim, selects evidence that supports that claim, and constructs a synthesis between the claim and the warrant.
1. RECEPTIVE: What science constructs / disciplinary core ideas (DCI) do students need to make sense of and reason with in the task?
1. Students are aware that sugar is added to the water and then it disappears after stirring. This is the phenomenon they have to explain.
2. RECEPTIVE: What elements (or components) of argumentation practice must students negotiate with in the task?
• Students need to know what an “argument” is (and what it represents). Students need to construct an explanation for the disappearing sugar. They need to make a claim about the mechanism that caused the sugar to disappear and this mechanism needs to be consistent with the disappearing act and also involve some kind of reason having to do with the process of dissolving.
3. PRODUCTIVE: What science constructs must students demonstrate knowledge in?
1. Physical change vs. chemical change. This scenario represents a physical change (not chemical change)

2. Substance (sugar as solute) has dissolved in water (solvent), not melted.
4. PRODUCTIVE: What elements (or components) of argumentation practice must students demonstrate competency in? [Reflective of the LP level assessed]
• Students need to know what an "argument" is (and what it represents).

"Ideal response(s)"
The sugar dissolved in the water (claim). The sugar is still there but you cannot see it (evidence), because it broke up into tiny pieces (reasoning). **The sugar disappeared because …. Insert reason …. Insert evidence (cannot see the sugar)**

Part 3: Rubric

Table 1: Rubric components

Component	**0/1**	**Examples (include ID#)**
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 unoticable to the naked eye.
COMPONENT B: Possible Evidence		
B1: Cannot see the sugar. OR The sugar disappeared. Synonyms for "disappeared" and "cannot see" are OK.		21. The sugar was dissolved in the water. Because the grains of sugar are so small, they blend with the water and become unoticable to the naked eye. 29. The sugar dissolved into the water, forming a mixture, so that the individual sugar grains are no longer visible.
COMPONENT C: Possible reasons		
C1: 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.
C2: The sugar (molecules) bonded with water (molecules).		41. The sugar grains dissolved in the liquid, essentially bonding with the liquid molecules

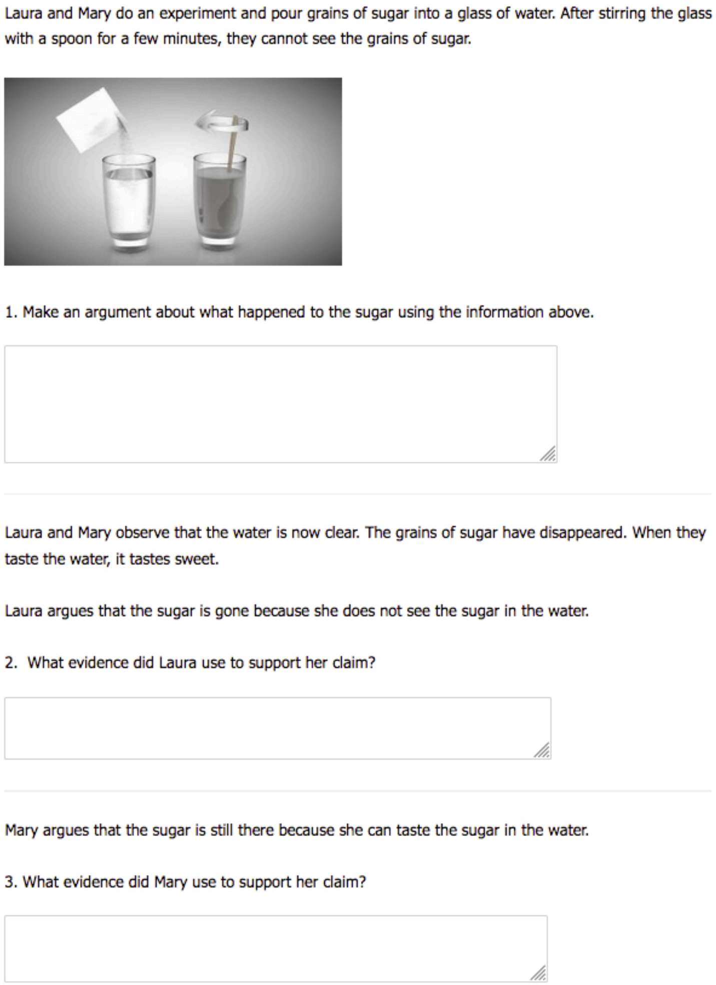
OR Sugar mixed/blended/combined with water.		21. The sugar was dissolved in the water. Because the grains of sugar are so small, they blend with the water and become unoticable to the naked eye.
C3: Physical act of **stirring**		[2505] i think that the sugar in the cup dissolved in the water because after they **stirred it**, the sugar was no longer visible.
Holistic Score (3 POINTS MAX)		
3 – Scientifically accurate claim, evidence, AND reasoning. (All three components must be present)		2114 When the sugar in the water was mixed, the sugar dissolved so it's as if we can't see it. 2121 The sugar dissolves after steering it in the water for a few minutes, That's the reason they cant see the sugar particles anymore. 4517 this happens because the gains of sugar desolve and you cant see them. like for ex. with a cube and put it in hot water and stir it in a cup it will deslove like the sugar.
2 – Scientifically accurate claim AND reasoning OR Scientifically accurate claim AND evidence		5110 The grains of sugar dissolved into the water, because the water molecules have bonded together with the sugar molecules, thus making it a sugary water. 5115 The grains of sugar dissolved after stirring the glass with a spoon. 7205 when they put the sugar in the cup it started to dissolve a little but when the spoon came and they mixed it around to make it dissolve
1 – [*Only* scientifically accurate claim OR *only* reasoning OR *only* evidence] OR Evidence AND reasoning (no claim)		6417 The sugar disappeared and the sugar never turned back to it's original shape. 6616 after they put the sugar in the water for a few minutes it was goon . 1514 when the put in the glass it goes away

0 – Out of context or not scientifically accurate		97 I think the sugar became darker due to irritating and spinning it in circles. 5509 What had happened to the grains of sugar? 1307 vision came and ate it 6619 The sugar melted in the water.

Appendix B

NEW name: Rubric_S3_0d_A1B1

Part 1: Task



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 an argument about what happened to the sugar using the information above.

Laura and Mary observe that the water is now clear. The grains of sugar have disappeared. When they taste the water, it tastes sweet.

Laura argues that the sugar is gone because she does not see the sugar in the water.

2. What evidence did Laura use to support her claim?

Mary argues that the sugar is still there because she can taste the sugar in the water.

3. What evidence did Mary use to support her claim?

Part 2: Audit of task

LP level: Identifying evidence
Definition: Student identifies another person’s piece of evidence.

1. RECEPTIVE: What science constructs / disciplinary core ideas (DCI) do students need to make sense of and reason with in the task?

Sugar is a solute that can dissolve in a liquid solvent such as water. A physical changes has taken place from solid to liquid. It is NOT a chemical reaction.

2. RECEPTIVE: What elements (or components) of argumentation practice must students negotiate with in the task?

Familiarity with the terms that are present in the task: argument, claim, evidence.

3. PRODUCTIVE: What science constructs must students demonstrate knowledge in?
Student is taking the perspective of Mary and understanding that Mary used her sense of taste to make her claim.
4. PRODUCTIVE: What elements (or components) of argumentation practice must students demonstrate competency in? [Reflective of the LP level assessed]
Students need to know what “evidence” means and Mary’s “claim” (terminology/meaning/application in context).

“Ideal response(s)”
Mary can taste the sugar in the water. OR The sugar taste is present in the water.

Part 3: Rubric

Table 1: Rubric components

Component		**Examples**
COMPONENT A: Evidence **A1**: It/sugar can be tasted. OR She/her/Mary can taste the sugar. OR The water tastes sweet	1/0	2. Mary used taste evidence (she could taste the sweetness of the sugar in the water). 5. The sweetened taste of the water after the sugar dissolved.
COMPONENT B: Claim **B1**: Student repeats Mary’s claim that the *sugar is still in the water/cup* (or any variation of this claim).	1/0	16. The sugar had nowhere to go. 22. Mary believes that the sugar is still present because she can still taste it.
Holistic score		
Response contains Evidence and NO Claim (A1=1 AND B1=0)	**1**	7. She can taste the typical sweetness of sugar incorporated in the water. 12. She uses her taste buds and taste of sugar to support her claim.

Response contains claim (with or without evidence)	**0**	16. The sugar had nowhere to go. 22. Mary believes that the sugar is still present because she can still taste it.