Automated Analyses of Students' Difficulties with Explanations in Science Inquiry

Introduction

The NGSS (2013) requires students to conduct inquiry and communicate their scientific findings. However, students often struggle with constructing explanations and engaging in argumentation in science (Chinn & Brewer, 1993; Klahr & Dunbar, 1988). Further, teachers must be able to assess students and provide instruction on these argumentation practices. Although some coding schemes have been developed to support teachers' assessment and instruction of students' argumentation competencies (e.g., Osborn et al., 2016), grading students' writing is arduous and subject to grading biases by teachers (Myford & Wolfe, 2009). Thus, automated scoring algorithms, such as those developed by our team (Authors, 2017a), can be beneficial for teachers as they can provide actionable data about their students' difficulties, which can inform their instruction (Ruiz-Primo & Furtak, 2006).

In the present study, we leverage a Claim-Evidence-Reasoning (C-E-R) framework (McNeill et al., 2006) to elicit students' competencies with constructing scientific explanations and engaging in argumentation. Additionally, we use natural language processing algorithms (Authors, 2017a) to assess students' C-E-R statements and their respective fine-grained sub-components (described later). Ours differs from other automated assessments of scientific writing, which often assess students' responses holistically (e.g., Liu et al., 2016). By developing and applying multiple algorithms to score students on the *fine-grained* sub-components of their three separate C-E-R responses, this approach helps to better capture students' difficulties, thus informing future scaffolds for students and instructional supports for teachers.

Methods

Participants and Materials

Participants include 76 eighth-grade students taught by one teacher in the Northeastern United States. The students completed two virtual labs in the [ITS] environment, conducting scientific inquiry about a sled moving down a ramp. In Lab 1 (i.e., Task 1), students “determine how the ramp roughness affects the time to the end of the ramp”; in Lab 2 (i.e., Task 2), students “determine how the sled size affects the distance traveled from the end of the ramp.” In both labs, students form a question and collect and analyze data to determine whether it supports or refutes their hypothesis. At the end of the lab, students communicate their findings in a Claim-Evidence-Reasoning format (McNeill et al., 2006).

Measures

Students’ scores for each of the three components (i.e., Claim, Evidence, Reasoning) were calculated as a sum of the scores for sub-components underlying each component, as defined by previously developed rubrics (Figure 1; Authors et al., 2017a). Students’ C-E-R responses were automatically scored at the sub-component level based on previously developed automated scoring algorithms; see prior work (Authors et al., 2017a) for more information about the rubrics and algorithms.

Figure 1

Claim-Evidence-Reasoning Sub-Components, Descriptions, and Possible Points

CER	Sub-Component	Description	Possible Point Values
Claim	Claim: IV	Did the student state the target independent variable (IV)?	No Credit: 0 Max Credit: 1
	Claim: IVR	Did the student say how they changed the independent variable (i.e., the independent variable relationship; IVR)?	No Credit: 0 Partial Credit: 0.5, 0.8 Max Credit: 1
	Claim: DV	Did the student state the target dependent variable (DV)?	No Credit: 0 Max Credit: 1
	Claim: DVR	Did the student say how the dependent variable changed in the experiment (i.e., the dependent variable relationship; DVR)?	No Credit: 0 Partial Credit: 0.5, 0.8 Max Credit: 1
Evidence	Evidence: Sufficient	Did the student state data for at least two trials (i.e., a sufficient amount of data)?	No Credit: 0 Partial Credit: 0.5, 1 Max Credit: 2
	Evidence: Appropriate IVR	Did the student state the appropriate data for the independent variable?	No Credit: 0 Partial Credit: 0.5, 0.8 Max Credit: 1
	Evidence: Appropriate DVR	Did the student state the appropriate data for the dependent variable?	No Credit: 0 Partial Credit: 0.5, 0.8 Max Credit: 1
Reasoning	Reasoning: Theory	Did the student explain the scientific principle behind the phenomena?	No Credit: 0 Max Credit: 1
	Reasoning: Connection	Did the student say how the claim relates to the evidence?	No Credit: 0 Partial Credit: 0.5, 0.8 Max Credit: 1
	Reasoning: IV/IVR	Did the student state the independent variable and/or say how they changed the independent variable?	No Credit: 0 Partial Credit: 0.5 Max Credit: 1
	Reasoning: DV/DVR	Did the student state the dependent variable and/or say how they changed the dependent variable?	No Credit: 0 Partial Credit: 0.5 Max Credit: 1

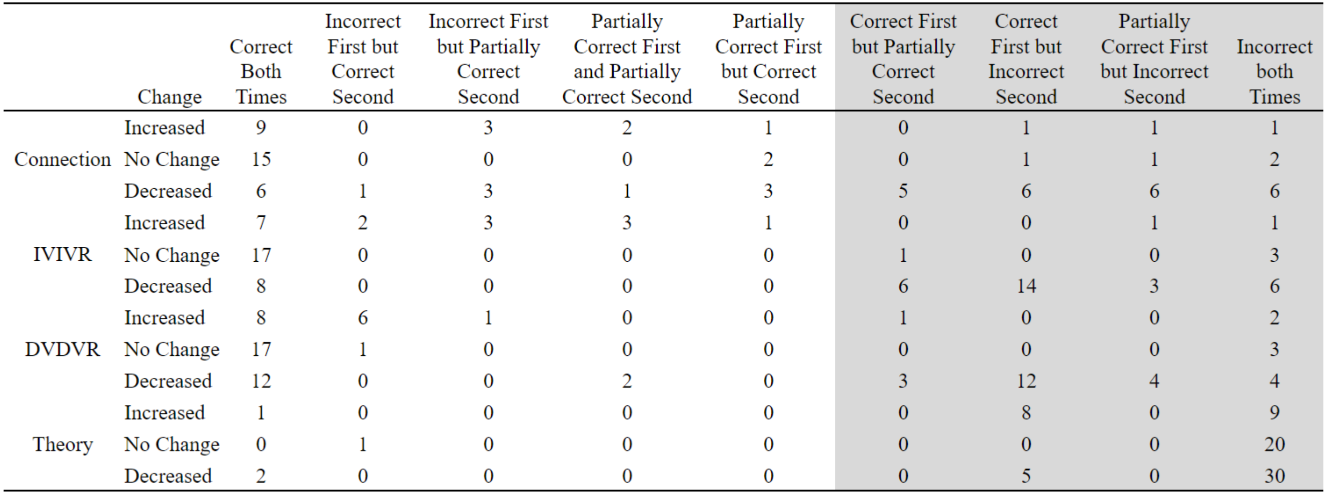
Analysis & Results

Three paired samples t-tests were conducted to compare scores from Task 1 to Task 2 in each of the C-E-R scores. There was not a significant change in Claim scores from Task 1 (*M* = 2.80, *SD* = 1.19) to Task 2 (*M* = 2.78, *SD* = 1.42); *t*(75) = 0.13, *p* = .899 or in the Evidence scores from Task 1 (*M* = 2.86, *SD* = 1.38) to Task 2 (*M* = 2.93, *SD* = 1.40); *t*(75) = 0.23, *p* = .654. However, there was a significant difference, namely, a decrease in the average Reasoning scores from Task 1 (*M* = 3.49, *SD* = 1.37) to Task 2 (*M* = 2.96, *SD* = 1.75); *t*(75) = 2.68, *p* = .009. These results suggest that students need support for all three components (i.e., Claim, Evidence, and Reasoning), because without support, students may not improve at this critical NGSS (2013) practice. To illustrate *why* students are not improving, our presentation will unpack students' difficulties with each of the C-E-R sub-components (Figure 1); however, due to space limitations, we are currently focusing on their Reasoning responses.

There were four sub-components for Reasoning: *Theory* (explaining scientific principles behind phenomena), *Connection* (explaining how claims relate to evidence), *IV/IVR* (listing the independent variable and/or how they changed it), and *DV/DVR* (stating the dependent variable and/or how it changed) (Figure 1). Students were grouped based on the change in their total score of these 4 sub-components for Reasoning (i.e., *increased*, *decreased*, *no change*). Their scores on each of the four Reasoning sub-components were assessed for correctness (i.e., correct, partially correct, or incorrect), to determine their outcomes on each task, and frequency (i.e., first task, second task, or both tasks), to determine how their scores changed between the two tasks at the sub-component level (Figure 2). The results show that, of the four Reasoning sub-components, *Theory*, or explaining relevant scientific phenomena, was the most difficult, with 77.6% (59/76) getting it incorrect (i.e., 0 points) both times, regardless of whether they increased, decreased, or stayed the same on their overall Reasoning score.

Figure 2

Student Sub-scores within Communication Score Reasoning Grouped by Overall Change (N=76)



	Change	Correct Both Times	Incorrect First but Correct Second	Incorrect First but Partially Correct Second	Partially Correct First and Partially Correct Second	Partially Correct First but Correct Second	Correct First but Partially Correct Second	Correct First but Incorrect Second	Partially Correct First but Incorrect Second	Incorrect both Times
Connection	Increased	9	0	3	2	1	0	1	1	1
	No Change	15	0	0	0	2	0	1	1	2
	Decreased	6	1	3	1	3	5	6	6	6
IVIVR	Increased	7	2	3	3	1	0	0	1	1
	No Change	17	0	0	0	0	1	0	0	3
	Decreased	8	0	0	0	0	6	14	3	6
DVDVR	Increased	8	6	1	0	0	1	0	0	2
	No Change	17	1	0	0	0	0	0	0	3
	Decreased	12	0	0	2	0	3	12	4	4
Theory	Increased	1	0	0	0	0	0	8	0	9
	No Change	0	1	0	0	0	0	0	0	20
	Decreased	2	0	0	0	0	0	5	0	30

Discussion

We found that students struggled the most with the *Theory* sub-component of Reasoning, which involves explaining scientific theory. Students often demonstrate difficulties with incorporating scientific theories and principles into their reasoning across science domains both in [ITS] (Authors, 2023) and elsewhere (McNeill et al., 2006), indicating they require more support in this area. To help students improve with generating C-E-R statements, it is important to operationalize the competencies at a fine-grained level (Kubsch et al., 2022). This type of fine-grained approach can identify students' specific difficulties, as we have here, and, then, help us design fine-grained scaffolds to target students' specific difficulties once they are implemented into [ITS]. Secondly, this approach can provide teachers with detailed, actionable information about where students are struggling in C-E-R tasks, thereby, informing a learning progression for this important and challenging science practice.

Resources

Authors. (2013; 2017a; 2017b; 2019).

Chinn, C. A., & Brewer, W. F. (1993). The role of anomalous data in knowledge acquisition: A theoretical framework and implications for science instruction. *Review of Educational Research*, *63*(1), 1-49.

Klahr, D., & Dunbar, K. (1988). Dual space search during scientific reasoning. *Cognitive science*, *12*(1), 1-48.

Kubsch, M., Czinczel, B., Lossjew, J., Wyrwich, T., Bednorz, D., Bernholt, S., Fiedler, D., Strauß, S., Cress, U., Drachsler, H., Newmann, K., & Rummel, N. (2022). Toward learning progression analytics—Developing learning environments for the automated analysis of learning using evidence centered design. In, *Frontiers in Education* (p. 605). Frontiers.

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.

McNeill, K. L., Lizotte, D. J., Krajcik, J., & Marx, R. W. (2006). Supporting students' construction of scientific explanations by fading scaffolds in instructional materials. *The Journal of the Learning Sciences*, *15*(2), 153-191.

Myford, C. M., & Wolfe, E. W. (2009). Monitoring rater performance over time: A framework for detecting differential accuracy and differential scale category use. *Journal of Educational Measurement*, *46*(4), 371-389.

Next Generation Science Standards Lead States. (2013). *Next Generation Science Standards: For states, by states*. The National Academies Press.

Osborne, J. F., Henderson, J. B., MacPherson, A., Szu, E., Wild, A., & Yao, S. Y. (2016). The development and validation of a learning progression for argumentation in science. *Journal of Research in Science Teaching*, *53*(6), 821-846.

Ruiz-Primo, M. A., & Furtak, E. M. (2006). Informal formative assessment and scientific inquiry: Exploring teachers' practices and student learning. *Educational Assessment, 11*(3-4), 237-263.