Evaluation of Automated Scoring Methods for Students' Claim, Evidence, Reasoning Responses in Science

[1] Haiying Li, [2] Amy Adair, [2] Grace Li, [3] Rachel Dickler, [2] Janice Gobert

[1] haiyli@upenn.edu, University of Pennsylvania

[2] {amy.adair; grace.li; janice.gobert}@gse.rutgers.edu, Rutgers University

[3] rachel.dickler@colorado.edu, University of Colorado Boulder

The Next Generation Science Standards (NGSS Lead States, 2013; NRC, 2012) emphasize constructing explanations as a key science inquiry practice; however, it can be challenging and time-consuming for teachers to develop scoring rubrics and grade students' written science inquiry explanations. The Inquiry Intelligent Tutoring System (Inq-ITS; Gobert et al., 2013) addresses this challenge by using natural language processing (NLP) techniques to automatically assess students' science explanations written in a Claim, Evidence, Reasoning (CER) format (Li et al., 2017a,b; McNeill et al., 2006). In this study, we expand on prior work evaluating the automated scoring of science explanations (Li et al, 2017a,b) by evaluating two different automated scoring methods. Results show that both methods perform moderately well when scoring students' CER responses according to a pre-defined rubric. Our findings hold implications for the future implementation of the automated scoring method across activities and leveraging such algorithms to trigger real-time feedback and scaffolding to support students' CER writing within Inq-ITS.

Methods

We first obtained all sets of students' CER responses completed in Inq-ITS in the 2021-2022 school year for four activity topics (Table 1). We then selected a random sample of 100 response sets from each topic to manually hand-score. One of the authors completed the hand-

scoring using rubrics (Table 2) adapted from previously developed rubrics for CER science explanations (Li et al., 2017a,b). With respect to the algorithms used to automatically score students' responses, there were two different methods for implementing the scoring rules: a RegEx method and a WordDistance method.

The RegEx method uses regular expressions (Thompson, 1968) to match patterns of words associated with the scoring component (e.g., `r'\broughness\b'` for the Claim_IV component in Forces & Motion since the *roughness* of the ramp in the virtual lab simulation is the target independent variable). This method has been described in more detail and used in prior Inq-ITS work (Li et al., 2017a,b). The WordDistance method is an alternative method in which the terms associated with the key concepts in an activity are linked to an identifier in a dictionary, with all similar terms (e.g., "affects", "impacts") using the same identifier. These key terms can then be combined with others to form larger concepts, such as the IV relationship (e.g., "the mass of the green ball increases"). Components are then scored based on the presence of those terms, or concepts, occurring within a specified distance from each other. For both methods, the developer used both the rubric and previously hand-scored student data as a reference when developing the scoring patterns.

Results and Discussion

The average human-computer inter-rater agreement, as measured by Cohen's quadratic weighted kappa (Cohen, 1968), was 88.77% for written claims, 85.98% for written evidence, 85.83% for written reasoning between RegEx method scores and human scores, and 85.87% for written claims, 89.10% for written evidence, 88.48% for written reasoning between WordDistance method scores and human scores (Tables 3-5). These results indicate that the algorithms are performing moderately well in scoring students' CER responses. However,

further analyses revealed that, for three activities, there was low agreement between the hand-scoring and auto-scoring for the “Theory” sub-component of the Reasoning portion for both algorithms. Future work will explore how to further operationalize students’ written scientific theories across the different science domains to improve automated scoring in this area.

References

Cohen, J. (1968). Weighted kappa: Nominal scale agreement provision for scaled disagreement or partial credit. *Psychological Bulletin*, *70*(4), 213.

Gobert, J. D., Sao Pedro, M., Raziuddin, J., & Baker, R. S. (2013). From log files to assessment metrics: Measuring students' science inquiry skills using educational data mining. *Journal of the Learning Sciences*, *22*(4), 521-563.

Li, H., Gobert, J., & Dickler, R. (2017a). Dusting off the messy middle: Assessing students' inquiry skills through doing and writing. In E. André, R. Baker, X. Hu, M. Rodrigo, & B. du Boulay (Eds.), *Artificial Intelligence in Education AIED 2017 Lecture Notes in Computer Science* (Vol. 10331, pp. 175-187). Springer.

Li, H., Gobert, J., & Dickler, R. (2017b). Automated assessment for scientific explanations in on-line science inquiry. In A. Hershkovitz & L. Paquette (Eds.), *Proceedings of the 10th International Conference on Educational Data Mining* (pp. 214-219). EDM Society.

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.

National Research Council (2012). *A framework for K-12 science education: Practices, crosscutting concepts, and core ideas*. National Academies Press.

NGSS Lead States. (2013). *Next generation science standards: For states, by states*. The National Academies Press.

Thompson, K. (1968). Programming techniques: Regular expression search algorithm. *Communications of the ACM*, *11*(6), 419-422.

Appendix

Table 1. Inq-ITS Virtual Lab Activity Topics Used for CER Auto-Scoring Evaluation

Automated Method	**Science Domain**	**Topic**	**NGSS Performance Expectation**	**# of CER Responses**	**Investigative Goal**
RegEx	Physical Science	Forces & Motion	MS-PS2-2	2880	Determine how the roughness of the ramp impacts the time to end of the ramp.
	Life Science	Animal Cell	MS-LS1-2	1122	The cell cannot break down food. Determine how lysosomes impact the ability to break down food.
Word Distance	Earth Science	Plate Tectonics: Convergent Plates	MS-ESS2-3	2439	Investigate how the duration of plate movement impacts the formation heights at the convergent boundary
	Physical Science	Collisions	MS-PS2-1	1489	Determine how the mass of the green ball affects the final velocity of the green ball.

Table 2. Summary of CER Science Explanation Rubric

Response	**Sub-Component**	**Possible Points**
Claim	IV – States the independent variable (IV)?	0, 1
	IVR – Describes how the IV changed?	0, 0.5, 0.8, 1
	DV – States the dependent variable (DV)?	0, 1
	DVR – Describes how the DV changed?	0, 0.5, 0.8, 1
Evidence	Sufficient – Mentions at least 2 trials?	0, 0.5, 1, 2
	Appropriate IV – Refers to data/trials related to their IV?	0, 0.5, 0.8, 1
	Appropriate DV – Refers to data/trials related to their DV?	0, 0.5, 0.8, 1
Reasoning	Theory – Describes a scientific theory, concept, or principle related to the phenomenon?	0, 0.8, 1
	Connection – States how their data/evidence relates to the claim?	0, 0.5, 0.8, 1
	IV/IVR – States the IV and/or describes how the IV changed?	0, 0.5, 1
	DV/DVR – States the DV and/or describes how the DV changed?	0, 0.5, 1

Table 3. Human-Computer Inter-Rater Agreement (Cohen's Kappa) between Hand-Scoring and Auto-Scoring across Virtual Lab Topics (Claim)

Automated Method	**Topic**	**CLAIM**				
		TOTAL	**IV**	**IVR**	**DV**	**DVR**
RegEx	Forces & Motion	94.39%	90.34%	88.25%	100.00%	83.21%
	Animal Cell	83.15%	88.37%	67.81%	100.00%	64.42%
	Average	**88.77%**	**89.36%**	**78.03%**	**100.00%**	**73.82%**
Word Distance	Plate Tectonics: Convergent Plates	92.88%	90.07%	78.44%	96.93%	77.46%
	Collisions	78.86%	87.90%	66.86%	87.01%	70.30%
	Average	**85.87%**	**88.99%**	**72.65%**	**91.97%**	**73.88%**

Table 4. Human-Computer Inter-Rater Agreement (Cohen's Kappa) between Hand-Scoring and Auto-Scoring across Virtual Lab Topics (Evidence)

Automated Method	**Topic**	**EVIDENCE**			
		TOTAL	**SUFFICIENT**	**APPROPRIATE IV**	**APPROPRIATE DV**
RegEx	Forces & Motion	90.51%	87.94%	90.81%	93.52%
	Animal Cell	81.45%	68.23%	87.19%	81.16%
	Average	**85.98%**	**78.09%**	**89.00%**	**87.34%**
Word Distance	Plate Tectonics: Convergent Plates	89.76%	80.77%	91.39%	94.35%
	Collisions	88.44%	79.67%	98.23%	73.60%
	Average	**89.10%**	**80.22%**	**94.81%**	**83.98%**

Table 5. Human-Computer Inter-Rater Agreement (Cohen's Kappa) between Hand-Scoring and Auto-Scoring across Virtual Lab Topics (Reasoning)

Automated Method	**Topic**	**REASONING**				
		TOTAL	**THEORY**	**CONNECTION**	**IV/IVR**	**DV/DVR**
RegEx	Forces & Motion	85.20%	77.30%	94.69%	84.32%	72.88%
	Animal Cell	86.45%	23.66%	93.79%	95.09%	99.49%
	Average	**85.83%**	**50.48%**	**94.24%**	**89.71%**	**86.19%**
Word Distance	Plate Tectonics: Convergent Plates	86.78%	6.32%	91.70%	90.79%	90.80%
	Collisions	90.17%	35.10%	97.02%	91.79%	93.35%
	Average	**88.48%**	**20.71%**	**94.36%**	**91.29%**	**92.08%**