Automated Assessment for Scientific Explanations in On-line Science Inquiry

Haiying Li✉
Rutgers University
10 Seminary Place
New Brunswick, NJ 08901
1(848)932-0868
haiying.li@gse.rutgers.edu

Janice Gobert
Rutgers University
10 Seminary Place
New Brunswick, NJ 08901
1(848)932-0867
janice.gobert@gse.rutgers.edu

Rachel Dickler
Rutgers University
10 Seminary Place
New Brunswick, NJ 08901
1(848)932-0869
rachel.dickler@gse.rutgers.edu

ABSTRACT
Scientific explanations, which include a claim, evidence, and reasoning (CER), are frequently used to measure students' deep conceptual understandings of science. In this study, we developed an automated scoring approach for the CER that students constructed as a part of virtual inquiry (e.g., formulating questions, analyzing data, and warranting claims) in an intelligent tutoring system (ITS), called Inq-ITS. Results showed that the automated scoring of CER was strongly correlated with human scores when validated using independent sets of data from both the same inquiry task/question, as well as when using data from a different inquiry task/question. These findings imply that automated CER is a very promising approach to reliably and efficiently score scientific explanations in open response format for both small- and large-scale assessments. It also provides Inq-ITS with the capability to assess the full complement of inquiry practices described by NGSS.

Keywords

automated assessment, scientific explanation, claim, evidence, reasoning

1. INTRODUCTION
The implementation of the Next Generation Science Standards (NGSS) has led to a need for assessments that are able to capture students' competencies at science inquiry practices [21]. Open-response tasks have been used in assessments for science inquiry because they can elicit students' communication skills, conceptual understandings, and ability to reason from evidence due to the measurement constraints of traditional multiple-choice items [10]. Rubrics for scoring students' explanations have been developed according to frameworks, such as Toulmin's [27] model of argumentation [8,16]. A modified version of Toulmin's model consists of three components: claim (an assertion about an investigated question), evidence (data or observations that support the assertion, i.e., the claim), and reasoning (articulating how the evidence supports the claim and how scientific principles explain the relationship between the data and claim).

Previous studies have developed rubrics to assess the accuracy of claim, evidence, and reasoning (CER) in students' scientific explanations. Gotwals and Songer [8] applied a rubric following the CER framework in order to score middle school students' explanations in an ecological science assessment. The rubric scoring for each component of CER was on a scale from 0 to 2 according to the accuracy and depth students' responses. McNeill et al. [16] scored students' responses to explanation prompts for middle school chemistry with a rubric that also followed the CER format using a 0 to 2 scale. These general rubrics for open response items provide some insight regarding the argumentation skill level of students, which can be valuable for guiding teachers' instruction and feedback. Open response items, however, can be time consuming and costly to score [28]; they can be inaccurately scored due to human factors such as rater fatigue [19], and rubrics can be interpreted and used differently by different raters [1]. One way to resolve these issues is through the use of automated scoring techniques [30].

Automated scoring techniques also permit automated feedback to students *as* they write scientific explanations or immediately following their writing tasks, when students have the opportunity to revise their writing. Automated, real-time feedback has been found to: significantly reduce the time between response submission and feedback relative to human scorers [15] and be as, if not more, effective than feedback presented by teachers [3]. While automated scoring presents an efficient and accurate means for promoting student learning gains, no studies, to date, have developed techniques for online, automated scoring of scientific explanations according to CER.

The current paper presents a new automated scoring approach to CER using the techniques of both natural language processing and machine learning. The approach addresses accuracy as well as important structural components of explanations as identified in the CER framework. The approach was validated using correlations between human scores and automated scores for scientific explanations produced in the Inq-ITS learning environment. Automated scoring of CER will: dramatically reduce time and expense, improve the efficiency and accuracy of CER scores, allow for instantaneous feedback, and make individualized instruction from teachers and/or automated scaffolding possible. Furthermore, scoring these data is critical because our data show that many students who have acquired a deep understanding of science content and inquiry practices, cannot articulate in words what they have learned. Conversely, some students are able to simply parrot what they have heard/read when doing written CER tasks, but do not actually understand the science content or practices [4, 11].

1.1 Automated Open Response
Automated scoring techniques have been developed to assess students' open responses in computer-assisted assessments and learning environments for science. Techniques include natural language-processing (NLP), such as regular expressions [12], to determine whether students' scientific explanations include key conceptual phrases [3, 13, 14]. The specific techniques and rubrics

used for automated scoring of science open response items vary across programs as described below.

The Summarization Integrated Development Environment (SIDE) uses a combination of NLP techniques and machine learning algorithms to score scientific explanations for the inclusion of biology concept knowledge [9, 20]. This system yielded correlations between human-scored and computer-scored responses ranging from 0.79 to 0.87 depending on the sample of participants. Disagreement was attributed to differences in linguistic tendencies across samples [9]. A later study by Nehm et al. [20] on the same system found that agreement between human and computer scoring was strong (i.e. $k > .81$). The SIDE program may be a valuable tool in scoring student scientific explanations [20], but is limited to identifying the presence of concepts within responses, and as such is not useful at scoring students' competencies at generating claims, evidence, and reasoning, which are critical to NGSS inquiry practices.

Another program that has been used to autoscore scientific explanations is the SPSS Text Analysis (SPSSTA) program [29], which uses language-processing procedures to identify terms and note patterns within texts [25]. A study by Weston et al. [29] applied SPSSTA to score undergraduate responses to biology explanation prompts. The agreement between human-coded responses and the SPSSTA for different levels of an analytic rubric ranged from a kappa of 0.67 to 0.88. The SPSSTA program relates to SIDE in terms of its potential to identify important concepts, but is unable to automatically produce machine learning algorithms from a trained data set [9], and this is limited in utility.

EvoGrader automatically scores constructed explanations using machine-learning algorithms [17]. A study compared EvoGrader scores to human scores based on the identification of nine key evolution concepts and strong agreement was found, as indicated by kappas above 0.85 for all concepts except one ($k = 0.71$) [17]. The EvoGrader automated assessment system was able to produce human-like scoring of key evolutionary concepts, but would need retraining in order to be generalized to other domains.

The c-Rater program scores scientific explanations based on the presence of central concepts using natural language processing [13]. A study by Liu et al. [13] compared human and c-Rater scores for four energy open response questions and found moderate agreement with Pearson correlations ranging from 0.67 to 0.72. While c-Rater was able to capture the presence of concepts, the program did not perform highly enough to be recommended for use as a sole scorer. Liu et al. [14] examined the agreement between human scorers and c-rater-ML, which is an autoscoring program that uses support vector regressions, a machine learning technique. Kappas across eight science explanation items ranged from 0.62 to 0.90, indicating good to very good agreement between human raters and c-rater-ML on a 5-point rubric for connecting key ideas [14]. The high agreement on certain explanation items demonstrated the potential for c-rater-ML to be used as a sole scorer, but, as noted by the authors, sensitivity to variations in phrasing of central concepts needed to be improved.

Automated scoring programs for scientific explanations exemplify the potential for accurate and efficient scoring of open responses in terms of the presence of scientific concepts, but do not provide opportunity for scoring more fine-grained components of explanations. That is, auto-scoring techniques have yet to address argumentative components of explanations that are central to science inquiry, namely students' competencies at generating claims, evidence for claims, and articulating the link between the two using reasoning, which are required by NGSS. Auto-scoring specific sub-components of responses, as we have done in our work, enables automated scaffolds that can, in turn, target specific areas of student difficulty. The rubrics for CER in previous studies broadly categorized responses into incorrect, partially correct, or fully correct, but failed to break down CER into finer-grained sub-skills or sub-components. As a result, previous rubrics have been unable to pinpoint exactly why students are having difficulties constructing explanations. In the present study, we developed a fine-grained rubric modified from McNeill et al. [16].

1.2 Description of Inq-ITS

Inq-ITS is a web-based intelligent tutoring system for Physical, Life, and Earth science that automatically assesses scientific inquiry practices at the middle school level in real time within interactive microworld simulations [5, 24]. Within each microworld, inquiry practices proposed in the NGSS for middle school are assessed including: question asking/hypothesizing, collecting data, analyzing data, warranting claims, and communicating findings using a CER framework.

Automated scoring has been implemented within Inq-ITS with patented algorithms [5] to measure sub-skills of each inquiry practice based on actions recorded in log files [7, 23]. Automated scoring of sub-skills in Inq-ITS required building detectors based on data-mined algorithms that captured variations of complex behaviors, such as designing controlled experiments [7, 24]. In order to build detectors, human raters used text-replay tagging to identify key behavioral features and train models that determined the presence of particular sub-skills [23]. The additional implementation of Bayesian Knowledge Tracing and Knowledge Engineering has enabled real-time, automated feedback that scaffolds students as they engage in inquiry practices in Inq-ITS [7] and has been found to result in significant inquiry learning gains for students [18, 22]. Sao Pedro and his colleagues [22, 24] found that students who had no experience with designing controlled experiments and testing stated hypotheses were able to acquire these skills after receiving scaffolded feedback from Inq-ITS's pedagogical agent, Rex. Moussavi, Gobert, and Sao Pedro [18] found that students who received scaffolds on data interpretation skills in one science topic of Inq-ITS were better able to apply those skills in a new science topic.

While automated scoring and feedback has been successfully applied to student actions in Inq-ITS, automated scoring has yet to be developed for written explanations. The automated scoring approach presented in this paper allows for automatic scoring of students' written scientific explanations in Inq-ITS, as well as lays the groundwork for the development of specific, automated feedback for open response items.

2. METHOD

2.1 Participants and Materials

Participants were 293 middle school students from 18 classes in six public middle schools who completed the Inq-ITS density virtual lab. The Density Virtual Lab contained three activities aimed to foster understanding about density of a liquid when using: different shapes of a container (narrow, square, and wide), different types of liquid (water, oil, and alcohol), and different amounts of liquid (quarter, half, and full). This study validated the automated scoring for the scientific explanations that students constructed in the first two activities: shape-density ($N = 293$) and type-density ($N = 268$) after a series of scientific investigations. The type-density data set was used to train and test the model with the method of 10-fold cross-validation. The shape-density data set was used to further test

the model to examine how well the model performed when it was generalized to an independent data set.

2.2 Rubrics and Inter-Rater Reliabilities

Scientific explanations in Inq-ITS consisted of three components: claim, evidence, and reasoning (CER). As previously stated, other rubrics have been unable to pinpoint exactly why students are having difficulties when constructing explanations. In the present study, we developed a fine-grained rubric modified from McNeill et al. [16], described as follows.

Claim was graded by four sub-skills: independent variable (IV), IV relationship (IVR; the conditions that students changed in the controlled target IV), dependent variable (DV), and DV relationship (DVR; the effect of IV on DV). For example, a good claim that a student wrote in the type-density activity was: I found out when you change the *type* (IV) of the liquid from *water* to *oil* (IVR), the *density* (DV) will *decrease* (DVR). IV and DV were graded with binary scores: 1 for presence of the sub-skill and 0 for the absence. IVR was classified into four levels: (1) correct answers in which students reported two controlled conditions of the target IV, (2) general answers in which students stated IVR using general expressions rather than specifically stating the conditions of change (e.g., I found that the *change (IVR) of type of liquid (IV) changes (DVR) the density (DV)*, (3) partial answers in which students only reported one controlled target condition (e.g., The density of *water* is the largest), and (4) incorrect answers. Therefore, correct IVR was given 1 point; general IVR, 0.8 points; partially correct IVR, 0.5 points; and incorrect IVR, 0 points. DVR in the type-density activity was scored according to three levels: correct (1 point), general (0.8 as shown in IVR example), and incorrect. DVR in the shape-density activity was scored dichotomously, correct (1 point) versus incorrect (0 points). The DVR (shape of the container) did not affect the DV (density), so responses were either correct or incorrect and no general expressions were involved.

Evidence was scored by two sub-skills: sufficiency and appropriateness [6]. Sufficiency was a measure of whether students provided sufficient evidence. If two controlled target conditions were stated, then 2 points were given. Mentioning only one controlled target condition was insufficient and was given 1 point. Using general expressions was given 0.5. Not mentioning any controlled target condition was incorrect and was given 0 points. Appropriateness was a measure of whether students provided appropriate data, such as the data of mass, volume, and density, as displayed in students' data tables in Inq-ITS. This sub-skill was consistent with the sufficiency of evidence, but focused on the data. Here is an example of a good answer in the shape-density activity: No matter what the container shapes are, narrow or wide, and the *mass of oil was 212.5* (data of mass) while the *volume was 250* (data of volume). *The density resulted in 0.85 g/ml* (data of density). If students specified the data of density, they were given 1 point for DVR in appropriate evidence; otherwise, 0 points. If they reported both the data of mass and volume, they were given 1 point. If they only reported the data of either mass or volume, they were given 0.5 points. If they did not report any data of mass or volume, they were given 0 points.

Reasoning was measured by three sub-skills: theory, connection between data and the claim, and data that supports or refutes the claim. Theory referred to whether students stated a scientific principle related to density, here being: the properties of a substance (based on the type of liquid) affect the density, not the shape of the container. Four categories were classified: (1) complete theory for 2 points (e.g., When looking at the data chart, it is noticeable that *the mass and volume don't change so the density doesn't change*.), (2) partial but closer to complete for 1 point (e.g., only mentioning two of three properties), (3) partial but closer to none for 0.5 points (e.g., only mentioning one property), and (4) incorrect or no theories for 0 points (e.g., no property was mentioned). Connection between data and claim referred to whether students specified that their data supports or refutes their claim. If they did, 1 point was given (e.g., *My evidence supports my claim...*). If they only partially stated the connection, 0.5 points were given (e.g., *It will support my claim…*) because the student did not specify whether the data or evidence supported the claim. If there were no expressions specified, 0 points were given. Data in the reasoning task were similar to the claim task with one main difference. In scoring reasoning data, mentioning *either* IV or IVR was accepted as correct (1 points) and mentioning only one condition of change was considered partially correct (0.5 points).

Two expert raters scored students' CER according to the fine-grained rubric. The interrater-reliabilities by Cronbach's α were .993, .994, .938 and the intraclass correlations were .986, .988, .882 for claim, evidence, and reasoning, respectively, higher than human agreement in prior studies (e.g., [14]). Disagreements were discussed until agreement was reached and agreed upon scores were used for analyses.

2.3 Automated Scoring

The target sub-skills were extracted using regular expressions (RegEX) based on the rubrics used by human raters in section 2.2. RegEX is a natural language processing technique that often applies algorithms to search for specific phrases or phrases that are semantically equivalent to a target concept [26]. In ITSs, RegEX has been used to accurately identify the presence of target concepts in students' responses [12]. Table 1 displays some examples of the RegEX that we used to extract features. RegEXs were generated based on semantically similar phrases that corresponded to a particular concept noted in the rubric.

Table 1. Examples of RegEX in the Shape-Density Activity.

CER	Sub-Skill		RegEX
Claim (0~4)	IV		shape
	IVR		(narrow.*square| (square.*wide)
	DV		density
	DVR		(^((?!n[o']t|doesn(')?t.)* (same|constant))
Evidence (0~4)	Sufficient		Same as IVR
	IVR		((mass.*volume).*250)| ((volume*mass).*250)
	DVR		1|85|78
Reasoning (0~6)	Theory		((mass.*volume).*density)
	Connection		(data|evidence).*(support|prove|indicate| |show|refute).*(claim|hypothesis|theory))
	Data	IV/IVR	shape|(narrow.*(wide|square))
		DV	Same as DV claim
		DVR	Same as DVR claim

If the sub-skill was binary, RegEX was used to detect the presence or absence of the content with Python programming language. If the sub-skill contained more than two levels, RegEx was used to detect the presence or absence of the sub-skill with a higher score first, and then with a lower score. Each sub-skill at each level was assigned to a binary score, 1 for the presence and 0 for absence of the sub-skill. If the sub-skill had more than two scales, each scale was assigned to a binary score first and then transformed into the true scores. Take IVR in claim as an example (e.g., *I found that the change of the container shape does not change the density*). RegEX

matched two conditions first and assigned this category a score of 0 because the two specific shapes were not mentioned. Then RegEX matched general expressions and found the target expression, *change of the container shape*, so 1 was assigned to the general expression category and matching stopped when the target content was found. In the analysis, this claim IVR was given a score of 0.8 points.

In this study, we used an if-then algorithm to search for a particular word or phrase, as is done in AutoTutor [12]. Take the IV (e.g., shape of the container) in the claim as an example. First, RegEX "shap" was generated to match the word "shape." Second, this RegEX was used to search a written claim. Third, if there was the word "shape" in the claim, then IV was present and scored as "1". If no word "shape" existed in the claim, then IV was considered absent and scored as "0". Moreover, before searching the target work, the misspelt target words were corrected to avoid a decrease in agreement [14]. If-then algorithms enhanced the performance especially for the more complex sub-skills, such as IVR, by matching the higher-level features first and then filtering down to the lower-level features. The modification of RegEx and algorithms typically took about 10 iterations for complicated sub-skills, such as theory, IVR, but fewer iterations for simple sub-skills, such as IV and DV. Each iteration took about 1-30 minutes, depending on the complexity of the sub-skills.

2.4 Statistical Analyses

Linear regression analyses were conducted using M5-prime method to assess whether sub-skills were predictive of human scores of CER. We used two methods to validate the model. The first method was 10-fold cross-validation. The second method was to further validate the model with an independent data set in a different inquiry, shape-density activity. If the model yields good performance with similar statistics as the cross-validation analyses, our confidence in model stability is increased and the model could be generalized to different Inq-ITS activities. We used the Pearson correlations as previous studies [14] did to evaluate automated scores and followed the same rules for describing their magnitude [2]: none (0.00–0.09), small (0.10–0.30), moderate (0.31–0.50), and large (0.51–1.00).

3. RESULTS

3.1 Performance of Automated Scores

A linear regression analysis for automated claim scoring with 10-fold cross-validation yielded a significant model in the type-density activity, $r = .97$, $p < .001$. The four sub-skills of claims were combined to account for 94% of the variance in the human claim scores, with correlation coefficients (β) of 1.02, 1.04, 1.07, 0.86 ($p < .001$) for IV, IVR, DV, and DVR, respectively. When this model was validated in the shape-density data set, it was also significantly correlated with human scores, $r = .94$, $p < .001$, which explained 88% of the variance in the human scores.

The same procedures were applied to the automated evidence scores. The cross-validation analysis showed a significant model, $r = .97$, $p < .001$, with three sub-skills accounting for 94% of the variance in the human evidence scores, with βs of 0.99, 0.87, and 0.90 ($p < .001$) for sufficiency, appropriateness IVR, and DVR, respectively. When this model was validated in the shape-density data set, the automated scores were also almost perfectly correlated with human scores, $r = .97$, $p < .001$, which explained 94% of the variance in the human evidence scores.

Finally, the same analysis was conducted for automated reasoning scores. The cross-validation analysis indicated a significant model, $r = .84$, $p < .001$, with five sub-skills accounting for 71% of the variance in the human reasoning scores, with βs of 0.21, 0.94, 0.85, 1.09, and 0.96 ($p < .001$) for theory, connection between data and claim, data of IV/IVR, DV, and DVR, respectively. When this model was validated in the shape-density data set, the automated scores were highly correlated with human reasoning scores, $r = .85$, $p < .001$, which explained 72% of the variance in the human reasoning scores.

These findings imply that the automated CER scores could best capture human CER scores in the independent sets of data from both the same inquiry task/question and data from a different inquiry task/question ($r = .84$~$.97$, larger than threshold of .50) [2]. These findings imply that the automated methods with the sub-skills of CER are a promising approach to automatically score scientific explanations respective of CER in science inquiry. This automated method with regular expressions and if-then algorithms enables automated scoring to be generalized to different inquiry activities without additional training and testing of the model, and yields satisfactory performance.

3.2 Analyses of Errors

Across three components of scientific explanations, automated claim and evidence scores almost perfectly predicted human claim and evidence scores when validated using independent sets of data from both the same inquiry task/question, as well as when using data from a different inquiry task/question. Reasoning showed a very good correlation between automated scores and human scores in both data sets, but this correlation was relatively low as compared to claim and evidence. This section, therefore, analyzes the errors of reasoning in the type-density data set. Table 2 displays the confusion matrix of automated rating and human rating for reasoning, which explicitly demonstrated a discrepancy for disagreement in scores between humans and automated scores. Results showed a high discrepancy for scores 2 – 4. Specifically, when the human score was 2, only 40% were given a score of 2 by automated methods. Almost half of the remaining responses were given 1 and the other half were given 3 points or more. Similarly, when the human score was 3, only 44% was scored 3 by automated methods. More than 30% was scored 2 and about another 30% was scored 4 – 5. It is the same for the human score of 4. Less than 40% of responses were scored 4 by automated methods, while more than half was scored 3 by automated methods.

Table 2. Confusion Matrix for Reasoning.

Scores	**Automated (Column)**							
Human (Row)	0	1	2	3	4	5	6	*N*
0	**19**	5	1					25
1		**23**	7					30
2	1	**13**	**18**	9	1	2	1	45
3		6	**34**	**47**	6	**11**	2	106
4			5	**27**	**20**		1	53
5			1	1	1	**2**	1	6
6						1	**2**	3
N	20	47	66	84	28	16	7	268

Note. 0–6 are the total reasoning scores rated by humans and automated methods based on the analytic rubrics.

This relatively lower agreement may have been largely due to inaccuracy that was caused by simple regular expressions. As constructed reasoning responses involve more complex causal relationships and different levels of sub-skills, the simple regular expressions may not completely cover all alternative expressions in students' responses. To examine which sub-skill showed high discrepancy between human rating and automated rating, we compared the agreement for the five sub-skills of reasoning

between automated scores and human scores. Results showed very high agreement for the first four sub-skills: 85% for theory, 85% for connection, 92% for data IV/IVR, and 95% for data DV, whereas the agreement for data DVR was only 46%. The confusion-matrix analyses for data DVR revealed that the automated scores used the binary score for this sub-skill (i.e., incorrect versus correct), whereas the humans rated DVR on the four levels mentioned in section 2.3. Binary scoring for DVR in the reasoning of the type-density activity was used to remain consistent with the scoring used in the shape-density activity. In the shape-density activity, there were no partially correct or general answers. Only correct answers (i.e. "density of liquid *is the same"* or "density of liquid *doesn't change*") or incorrect answers were considered. In the type-density activity, responses for DVR included correct answers (i.e. "density of the liquid *decreases from water to oil*"), general answers (i.e. "density of liquid *changes* due to the change of liquid"), partial answers (i.e. "density of *water* is largest"), and incorrect answers. With the rule of least effort, we did not change the algorithms from one activity to another to satisfy the multiple categories of students' responses accounted for by humans. Thus, a large disagreement arose due to the binary scoring used by the automated method versus the four level scoring used by humans.

Even though the criteria that humans and automated methods used to score DVR in reasoning were different, automated scores still yielded pretty good performance. The performance can be improved if the automated method scores reasoning using the same criteria as humans. A future study may explore whether the consistency in DVR between automated and human rating would improve the performance of reasoning scores overall.

4. DISCUSSION

These findings demonstrate that using regular expressions to match key sub-skills of CER with if-then algorithms is a very promising approach to effective and efficient automated scoring of open response scientific explanations. This assertion can be confirmed based on two key factors. First, the automated methods showed very good correlations with human scores for CER in the independent sets of data with the 10-fold cross-validation analyses in the same inquiry task/question as well as in a different inquiry task/question. Previous studies on automated scoring of constructed response items showed that good correlations between automated scores and human scores ranged from .60 to .91 (e.g., [14]). In our study, automated scores for claim and evidence reached .97 in the cross-validation analyses in the same inquiry task/question. When transferred to a different inquiry task/question, results remain .97 for evidence and .94 for claim. These results greatly exceed the current state of research on automated scoring of scientific explanations, as they are almost perfectly correlated with human scoring of claim and evidence scores. Even for reasoning using evidence, a more complex task, results were good as well, ranging from .84 to .85. One explanation for the slightly lower performance of automated reasoning scores is that the agreement between humans was lower relative to agreement for claim and evidence (.88 versus .99) due to the complexity of the reasoning task. Another explanation is that the regular expressions and algorithms applied across different tasks were the same. If we modify regular expressions to satisfy each activity, the performance of automated scoring for reasoning will likely increase.

Second, the sub-skill features that were extracted by regular expressions along with if-then algorithms not only consistently predicted human scores, but also were simple to implement. A central factor to the success of this method was that experts were able to generate accurate regular expressions to identify sub-skills of explanations in science inquiry. More specifically, experts knew how to identify the sub-skills of CER, how to develop a fine-grained rubric to guide human and machine scoring, and how to generate nearly-complete regular expressions to capture as many alternative expressions as possible in students' responses. The use of appropriate regular expressions was key to the success of our automated scores. Regular expressions were easier and quicker to generate for simple sub-skills such as IV, IVR, and DV for claim and data in evidence. For more complex sub-skills, such as DVR and theory, more time was needed to develop sets of alternative expressions. However, once the algorithms yield good performance, only a slight modification is needed for different activities. Compared to manual scoring, the time and effort that was spent on the development of automated scoring was worthwhile. Another key to the success of our automated scoring method was the development of the fine-grained rubric. Our rubric was finalized over many iterations. When we used more general rubrics, the inter-rater reliabilities for reasoning were very low ($r = .50$). With the fine-grained rubric, the reliabilities increased to .88. The high agreement between human coders guaranteed the possibility of high agreement between human scores and machine scores.

The success of automated scoring for open responses in science inquiry will greatly contribute to science education by making possible immediate individualized feedback on students' explanations, as well as adaptive instruction and scaffolding. The implementation of automated scoring in computer-assisted learning and assessment systems will provide students with instant feedback on their constructed CER, which will allow students to immediately know their strengths and weaknesses with regard to scientific explanations. Teachers could then use the explicit feedback from automated scoring to adapt instruction based on what students need. In addition, the automated scoring of CER in science inquiry will advance the development of computer-assisted systems for inquiry, such as Inq-ITS. Inq-ITS has used automated scoring to implement immediate feedback and scaffolding for inquiry skills involved in "doing" science, such as formulating a question/hypothesis, collecting data, analyzing data, and warranting claims. Automated scoring could also be used to align students' "doing" science skills with their science "writing" skills. The alignment of sub-skills involved in "doing" with "writing" during inquiry will allow for comparison of students' conceptual knowledge with their ability to communicate such knowledge. Thus, this automated scoring approach truly advances science education by meeting the comprehensive assessment criteria that NGSS [21] demands: science assessments that include both students' understandings of core ideas, their skills at conducting inquiry, as well as their skills at effectively articulating what they know by generating a claim and evidence for that claim, and articulating their reasoning linking their claim to their evidence.

Even though the automated methods for scientific CER demonstrated good performance, there is one limitation that needs to be addressed in future studies. Namely, regular expressions for reasoning may be modified to adapt to each task/question to align with criteria used by humans. In doing so, the accuracy may be improved.

5. ACKNOWLEDGMENTS

The research reported here was supported by Institute of Education Sciences (R305A120778) to Janice Gobert. Any opinions are those of authors and do not reflect the views of these funding agencies, cooperating institutions, or other individuals.

6. REFERENCES

[1] Bejar, I. I. 2012. Rater cognition: implications for validity. *Educ. Meas.* 31 (Sept. 2012), 2-9.

[2] Cohen, J. 1968. Weighted kappa: nominal scale agreement provision for scaled disagreement or partial credit. *Psychol. Bull*. 70 (Oct. 1968), 213-220.

[3] Gerard, L. F., Ryoo, K., McElhaney, K. W., Liu, O. L., Rafferty, A. N., and Linn, M. C. 2016. Automated guidance for student inquiry. *J. Educ. Psychol.* 108 (Jan. 2016), 60-81.

[4] Gobert, J. D. 2016. Op-Ed: Educational data mining can be leveraged to improve assessment of science skills. (May 2016). Retrieved from US News & World Report: https://www.usnews.com/news/articles/2016-05-13/op-ed-educational-data-mining-can-enhance-science-education

[5] Gobert, J. D., Baker, R. S., and Sao Pedro, M. A. 2014. Inquiry skills tutoring system. (Jan. 2014). US Patent no. 9,373,082, Filed Feb. 1st., 2013, Issued Jan. 29th., 2014.

[6] Gobert, J. D., Pallant, A. R., and Daniels, J. T. 2010. Unpacking inquiry skills from content knowledge in geoscience: a research and development study with implications for assessment design. *Int. J. Learn. Technol.* 5 (June. 2010), 310-334.

[7] Gobert, J. D., Sao Pedro, M. A., Baker, R. S., Toto, E., and Montalvo, O. 2012. Leveraging educational data mining for real-time performance assessment of scientific inquiry skills within microworlds. *J. Educ. Data Min.* 4, 111-143.

[8] Gotwals, A. W., and Songer, N. B. 2010. Reasoning up and down a food chain: using an assessment framework to investigate students' middle knowledge. *Sci. Educ.* 94 (Oct. 2010), 259-281. DOI= http://dx.doi.org/10.1002/sce.20368.

[9] Ha, M., Nehm, R. H., Urban-Lurain, M., and Merrill, J. E. 2011. Applying computerized-scoring models of written biological explanations across courses and colleges: prospects and limitations. *CBE-Life Sci. Educ.* 10 (Sept. 2011), 379-393.

[10] Lee, H. S., Liu, O. L., and Linn, M. C. 2011. Validating measurement of knowledge integration in science using multiple-choice and explanation items. *Appl. Meas. Educ.* 24 (Mar. 2011), 115-136.

[11] Li, H., Gobert, J., and Dickler, R. 2017. Dusting off the messy middle: Assessing students' inquiry skills through doing and writing. In *Artificial Intelligence in Education: Lecture Notes in Computer Science* (Wuhan, China, June 25-28, 2017). AIED '17. Spinger, China.

[12] Li, H., Shubeck, K., and Graesser, A. C. 2016. *Using Technology in Language Assessment.* Bloomsbury Academic, London, UK.

[13] Liu, O. L., Brew, C., Blackmore, J., Gerard, L., Madhok, J., and Linn, M. C. 2014. Automated scoring of constructed-response science items: prospects and obstacles. *Educ. Meas.* 33 (Mar. 2014), 19-28.

[14] Liu, O. L., Rios, J. A., Heilman, M., Gerard, L., and Linn, M. C. 2016. Validation of automated scoring of science assessments. *J. Res. Sci. Teach.* 53 (Jan. 2016), 215-233.

[15] Matthews, K., Janicki, T., He, L., and Patterson, L. 2012. Implementation of an automated grading system with an adaptive learning component to affect student feedback and response time. *J. Inform. Syst. Educ.* 23 (Apr. 2012), 71-83.

[16] McNeill, K., Lizotte, D.J., Krajcik, J., and Marx, R.W. 2006. Supporting students' construction of scientific explanations by fading scaffolds in instructional materials. *J. Learn. Sci.* 15 (Nov. 2006), 153-191.

[17] Moharreri, K., Ha, M., and Nehm, R. H. 2014. EvoGrader: an online formative assessment tool for automatically evaluating written evolutionary explanations. *Evol.* 7 (Aug. 2014), 15.

[18] Moussavi, R., Gobert, J., and Sao Pedro, M. 2016. The effect of scaffolding on the immediate transfer of students' data interpretation skills within science topics. In *Proceedings of the 12th International Conference of the Learning Sciences.* (Singapore, June 20-24, 2016). ICLS '16. Scopus, Ipswich, MA, 1002-1005.

[19] Myford, C., and Wolfe, E. 2009. Monitoring rater performance over time: a framework for detecting differential accuracy and differential scale category use. *J. Educ. Meas.* 46 (Dec. 2009), 371-389.

[20] Nehm, R. H., Ha, M., and Mayfield, E. 2011. Transforming biology assessment with machine learning: automated scoring of written evolutionary explanations. *J. Sci. Educ. Technol.* 21 (Feb. 2012), 183-196.

[21] Next Generation Science Standards (NGSS) Lead States. 2013. *Next Generation Science Standards: For States, by States*. The National Academies Press, Washington, DC.

[22] Sao Pedro, M. 2013. Real-time assessment, prediction, and scaffolding of middle school students' data collection skills within physical science simulations. Ph.D. Dissertation. Worcester Polytechnic Institute, Worcester, MA.

[23] Sao Pedro, M. A., Baker, R. S., Gobert, J. D., Montalvo, O., and Nakama, A. 2013. Leveraging machine-learned detectors of systematic inquiry behavior to estimate and predict transfer of inquiry skill. *User Model. User-Adap.* 23 (Mar. 2013), 1-39.

[24] Sao Pedro, M. A., Gobert, J. D., & Baker, R. S. 2014. The impacts of automatic scaffolding on students' acquisition of data collection inquiry skills. Roundtable presentation at *2014 American Educational Research Association Annual Meeting* (Philadelphia, Pennsylvania, April 03-07, 2014).

[25] SPSS Inc. 2006. *SPSS Text Analysis for Surveys 2.0 User's Guide.* SPSS Inc, Chicago, IL.

[26] Thompson, K. 1968. Regular expression search algorithm. *Commun. ACM.* 11 (June. 1968), 419–422.

[27] Toulmin, S. 1958. *The Uses of Argument.* Cambridge University Press, Cambridge, MA.

[28] Wainer, H., and Thissen, D. 1993. Combining multiple-choice and constructed-response test scores: toward a Marxist theory of test construction. *Appl. Meas. Educ.* 6 (Apr. 1993), 103-118.

[29] Weston, M., Parker, J., and Urban-Lurain, M. 2013. Comparing formative feedback reports: human and automated text analysis of constructed response questions in biology. In *Annual Conference of the National Association on Research in Science Teaching* (Rio Grande, Puerto Rico, April 06-09, 2013). NARST '13.

[30] Williamson, D. M., Xi, X., and Breyer, F. J. 2012. A framework for evaluation and use of automated scoring. *Educ. Meas.* 31 (Mar. 2012), 2-13.