Can Text Features of Investigative Questions in Science Predict Students' Inquiry Competencies?

Introduction

It is the vision of the Next Generation Science Standards (NGSS, 2013) that students develop an understanding of disciplinary core ideas by engaging with key science practices. With this vision in mind, the intelligent tutoring system [ITS] engages students in inquiry, logs students' interactions, and uses patented, educational data-mined algorithms on the log data to provide rich, fine-grained assessment of students' competencies with science inquiry practices. These practices include: asking questions (NGSS Practice 1), planning and carrying out investigations (NGSS Practice 3), and analyzing and interpreting data (NGSS Practice 4).

To become competent with these science practices and develop a deeper understanding about science phenomena (Goldman et al., 2016a), students must be able to interact with scientific text in meaningful ways (Cervetti et al., 2013). However, science texts often introduce new concepts by using unfamiliar language (Goldman et al., 2016b; Snow, 2010). Therefore, it is important to capture the difficulty level of the text that the students encounter in [ITS] virtual labs and consider how that difficulty may influence their performance on science inquiry.

To analyze the lexical sophistication of several investigative questions in [ITS], we used the natural language processing text evaluator, TAALES (Kyle & Crossley, 2015). In prior work, a similar text evaluator was used to show how the linguistic features of what students *wrote* in response to Claim-Evidence-Reasoning explanation tasks in [ITS] could be used to understand and predict their competencies with NGSS practices (Authors et al., 2018). However, because interacting with text involves both reading and writing, our goal in the present study is to explore the relationship between the text features of the investigative questions that students *read* in the virtual lab and their inquiry performance across science topics.

Method

Participants

The participants in the present study included 3,009 middle school students from across the United States. Each student completed at least one [ITS] virtual lab during the Fall 2020 semester.

Materials

At the beginning of each [ITS] virtual lab, students are presented with an investigative question, which includes key vocabulary used throughout the lab (see Table 1 for examples). As students conduct an experiment aligned to the investigative question, students are guided through three inquiry practice stages: (1) asking questions/hypothesizing, (2) carrying out investigations/collecting data, and (3) analyzing and interpreting data. [ITS] uses educational data-mined algorithms to assess students' fine-grained competencies with the science inquiry practices in each stage of the virtual lab activity (Authors et al., 2013; Authors et al., 2018). These scores were extracted and aggregated to determine students' overall science inquiry performance score (see Measures).

Data reflecting students' performance on each virtual lab that they completed within three activity sets (i.e., Phase Change, Free Fall, or Gravity & Mass) were collected via log files. These three activity sets were chosen because these were the most popular activity sets in the [ITS] catalogue in the Fall 2020 semester that address disciplinary core ideas in the domain of Physical Science.

Each activity set includes three or four labs with investigative questions that are aligned to the NGSS disciplinary core ideas for the topic (see Table 1). The lexical sophistication of each investigative question was determined using the indices from the text evaluator, TAALES (Kyle & Crossley, 2015; see Measures).

Measures

Students' performance on the science inquiry practice stages in each virtual lab was automatically assessed and stored in [ITS] using previously validated educational data mining and knowledge engineering techniques (Authors et al., 2013). These fine-grained scores were then averaged to determine the overall science inquiry performance scores (continuously ranging from 0 to 1) for each student on each investigative question.

Because prior work has highlighted that students have certain difficulties with comprehending difficult vocabulary within science text (Goldman et al., 2016a; Hiebert et al., 2019), we compared students' inquiry scores to the lexical sophistication of the investigative questions, as measured by the following indices: frequency, familiarity, age of exposure (inverse), and concreteness. These indices are based on a variety of corpora (Kyle & Crossley, 2015) and were chosen because they targeted features at the word level.

Analyses

Students' average inquiry scores and the automatically computed TAALES indices for each investigative question were used for the analyses. First, descriptive statistics were used to explore the trend in student performance across topics and investigative questions (see Table 1). A stepwise regression model was then constructed for each activity set to identify whether the TAALES indices as well as the order in which students conduct the virtual labs within an activity set could significantly predict students' inquiry performances (and to identify the factors that could best explain the variance in student performance).

Table 1. Average overall inquiry scores and the investigative question text for each virtual lab within three activity sets

Activity Set Topic	Lab Order	Virtual Lab Investigative Question Text	Inquiry Score (Standard Dev.)	N
Phase Change	1	Determine how the amount of ice affects the boiling point of water	.75 (.21)	1196
Phase Change	2	Determine how the amount of ice affects the melting point of ice	.83 (.19)	1068
Phase Change	3	Determine how the amount of heat affects the boiling point of water	.88 (.17)	980
Phase Change	4	Determine how the size of the container affects the time the water takes to boil	.90 (.16)	958
Free Fall	1	Determine how the mass of the ball affects the mechanical energy as the ball hits the ground	.77 (.22)	677
Free Fall	2	Determine how the height of the drop affects the potential energy before the ball is dropped	.88 (.17)	721
Free Fall	3	Determine how the height of the drop affects the kinetic energy as the ball hits the ground	.91 (.17)	592
Gravity and Mass	1	Determine how the planetary body we are orbiting affects the mass of the gold	.74 (.22)	1136
Gravity and Mass	2	Determine how the amount of gold affects the weight of the gold	.89 (.17)	956
Gravity and Mass	3	Determine how the planetary body we are orbiting affects the weight of the gold	.88 (.17)	870

Results

Descriptive statistics revealed that the order of the investigative questions in each activity set had a potential effect on inquiry scores, as average inquiry scores primarily increased with each proceeding lab (see Table 1). Because students can become more familiar with the scientific language as they progress through the virtual labs within an activity set, this finding prompted the team to investigate how the order of the investigative question within an activity set—as well as the text indices—affect overall inquiry scores.

The results of the stepwise regression analyses revealed that the model for each lab topic was significant, which suggests that the selected text indices account for some of the variability in students' inquiry scores (see Table 2). For the Phase Change model ($F(2,4198) = 227.565$, $p < 0.001$), Order and Concreteness were significant predictors and together could explain 9.8% of the variance in student performance. It is intuitive that Order had an effect because students tend to improve on [ITS] virtual labs with increased experience (Authors et al., 2019). Additionally, Concreteness indicates the importance of presenting terminology that can be readily conceptualized and described by students. The significant predictor for the Free Fall model ($F(1,1988) = 179.018$, $p < 0.001$) was Age of Exposure (Inverse), which explained 8.3% of variance in student performance. This indicates that as the Age of Exposure (Inverse) index decreased, students' inquiry performance improved, suggesting that students performed better on inquiry when lower-level vocabulary was used. The significant predictor for the Gravity and Mass model ($F(1, 2960) = 468.332$, $p < 0.001$) was Familiarity, which could explain 13.7% of the variance in student inquiry performance. Similar to the findings from Free Fall, students' inquiry performance improved when the words in the investigative question received a higher Familiarity index by the text evaluator.

Upon further examination, the technical vocabulary found in Free Fall and Gravity and mass was less common compared to the vocabulary found in Phase Change, causing the activities to be explained by Age of Exposure (Inverse) and Familiarity. Furthermore, the vocabulary found in the Phase Change questions were of relatively similar Familiarity, which meant that Familiarity could not explain variance to the degree that it did for Gravity and Mass. However, these differences did not result in a difference in overall student performance, as seen in Table 1.

Table 2. TAALES regression analysis predicting overall inquiry scores

Topic	Coefficient(s)	Beta	Sig.	R-squared
Phase Change	Order Concreteness	.305 .171	<.001 <.001	.098
Free Fall	Age of Exposure (Inverse)	.287	<.001	.083
Gravity and Mass	Familiarity	.370	<.001	.137

Discussion

This study suggests that exposure to and familiarity with scientific terminology can impact performance on science inquiry tasks, which aligns with findings from prior studies (Authors et al., 2018). This work is important in helping science educators understand why their students may be struggling with science inquiry tasks. The insight gained from text evaluators may suggest a more nuanced approach for instructional designers in crafting science text that is more accessible to students at a particular grade band and within a science domain. It is essential to keep the relationship between text features and science inquiry performance in mind when interpreting assessment results and to be aware that the extent to which language interacts with students' inquiry competencies as measured during assessment may vary by science topic.

References

Authors. (2013, 2018, 2019).

Cervetti, G. N., Pearson, P. D., Greenleaf, C., & Moje, E. B. (2013). Science? Literacy? Synergy! In W. Banko, M. L. Grant, M. E. Jabot, A. J. McCormack, & O'Brien (Eds.), *Science for the next generation: Preparing for the new standards* (pp. 99-124). NSTA & STANYS.

Goldman, S. R., Britt, M. A., Brown, W., Cribb, G., George, M., Greenleaf, C., & Project READI. (2016a). Disciplinary literacies and learning to read for understanding: A conceptual framework for disciplinary literacy. *Educational Psychologist*, *51*(2), 219-246.

Goldman, S. R., Snow, C., & Vaughn, S. (2016b). Common themes in teaching reading for understanding: Lessons from three projects. *Journal of Adolescent & Adult Literacy*, *60*(3), 255-264.

Hiebert, E. H., Scott, J. A., Castaneda, R., & Spichtig, A. (2019). An analysis of the features of words that influence vocabulary difficulty. *Education Sciences*, *9*(1), 8.

Kyle, K., & Crossley, S. A. (2015). Automatically assessing lexical sophistication: Indices, tools, findings, and application. *Tesol Quarterly*, *49*(4), 757-786.

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

Snow, C. E. (2010). Academic language and the challenge of reading for learning about science. *Science*, *328*(5977), 450-452.