Research Article

Statistical and Qualitative Analyses of Students' Answers to a Constructed Response Test of Science Inquiry Knowledge

Seohyun Kim, Minho Kwak, Lourdes Cardozo-Gaibisso, Cory Buxton, Allan S. Cohen, *University of Georgia*

Structured Abstract

• **Objective:** We report on a comparative study of the language used by middle school students in their answers to a constructed response test of science inquiry knowledge.

• **Background:** Text analyses using statistical models have been conducted across a number of disciplines to identify topics in a journal, to extract topics in Twitter messages, and to investigate political preferences. In education, relatively few studies have analyzed the text of students' written answers to investigate topics underlying the answers.

• **Methodology:** Two types of linguistic analysis were compared to investigate their utility in understanding students' learning of scientific investigation practices. A statistical method, latent Dirichlet allocation (LDA), was used to extract topics from the texts of student responses. In the LDA model, topics are viewed as multinomial distributions over the vocabulary of documents. These topics were examined for content and used to characterize student responses on the constructed response items. The change from pre-test to post-test in proportions of use of each of the topics was related to students' learning. Next, a qualitative method, systemic functional linguistic (SFL) analysis, was used to analyze the text of student responses on the same test of science inquiry knowledge. Student assessments were analyzed for two linguistic features that are

DOI: 10.37514/JWA-J.2017.1.1.05

important for convincing scientific communication: technical vocabulary usage and high lexical density. In this way, we investigated whether human judgement regarding the changes observed from texts based on the SFL framework agreed with the inference regarding the changes observed from the texts through LDA.

• **Research questions:** Two research questions were investigated in this study:

(1) What do the LDA and SFL analyses tell us about students' answers?

(2) What are the similarities and differences of the two analyses?

• **Data:** The data for this study were taken from an NSF-funded host study on teaching science inquiry skills to middle school students who were a mix of both native English speakers and English-language learners. The primary objective was to enable participants to learn to take ownership of scientific language through the use of language-rich science investigation practices. The LDA analysis used a sample of 252 students' pre-and post-assessments. The SFL analysis used a second sample of 90 students' pre- and post-assessments.

• **Results:** In the LDA analysis, three topics were detected in student responses: "preponderance of everyday language (Topic 1)," "preponderance of general academic language (Topic 2)," and "preponderance of discipline-specific language (Topic 3)." Students' use of topics changed from pre-test to post-test. Students on the post-test tended to have higher proportions of Topic 3 than students on the pre-test. In the SFL analysis, students tended to use more technical vocabulary and have higher lexical density in their written responses on the post-test than on the pre-test.

• **Discussion:** Results from the LDA and SFL analyses suggest that students responded using more discipline-specific language on the post-test than on the pre-test. In addition, the results of the two linguistic features from the SFL analysis, technical vocabulary usage and lexical density, were compared with the results from the LDA analysis.

• **Conclusion:** Results of the LDA and SFL analyses were consistent with each other and clearly showed that students improved in their ability to use the discipline-specific and academic terminology of the language of scientific communication.

Keywords: constructed response items, latent Dirichlet allocation, systemic functional linguistic analysis, text analysis, topic models, writing analytics

1.0 Background

Text analyses using statistical models have been conducted across a number of disciplines to extract meaningful textual-based information from documents. Erosheva, Fienberg, and Lafferty (2004) analyzed the abstracts and bibliographies of a journal using a mixed-membership model to identify topics in the journal. Paul and Dredze (2011) analyzed Twitter messages using an applied latent Dirichlet allocation model (LDA; Blei, Ng & Jordan, 2003) to extract health-related issues. Phan, Nguyen, and Horiguchi (2008) analyzed medical texts using the LDA model to verify hidden topics underlying texts. In political science, Grimmer (2009) introduced the expressed agenda model to find topics in the texts from politicians. Also, Lauderdale and Clark (2014) analyzed text and voting data by combining the LDA model and a multidimensional item response model to investigate political preferences.

In education there are also a number of studies analyzing texts, especially those written by students. These texts, however, are generally first scored using a rubric, and then the values obtained are analyzed. Huerta, Lara-Alecio, Tong, and Irby (2014) quantified students' writings in their notebooks to evaluate the use of academic language. Ruiz-Primo, Li, Ayala, and Shavelson (2004) evaluated students' quality of communication in science again using the text in students' notebooks. Rescorla, Mirak, and Singh (2000) used a vocabulary test to measure nominal word percentage in children's answers, and then fit a statistical model to the test scores to investigate growth in vocabulary.

Recently, a few studies have attempted to use statistical models to more directly analyze texts written by students. Chen, Yu, Zhang, & Yu (2016) analyzed journals from preservice teachers to investigate latent topics and patterns. Kim, Kwak, and Cohen (2017) analyzed student responses to constructed response (CR) items on tests using LDA to detect latent topics in the answers. The information was then used as a covariate to explain students' latent class membership obtained using a mixture item response model of the scores to the CR items. Kwak, Kim, and Cohen (2017) analyzed student written responses to CR items to investigate methods for explaining growth in use of academic vocabulary. These studies used latent topics obtained using statistical models for text analysis.

Grimmer and Stewart (2013) argue that using results from automatic content analyses, such as LDA, requires caution and that validation is necessary to justify the interpretation of the results. They noted that the automatic content analysis can mislead researchers about the content, or the analysis can provide incorrect results. This consideration motivated the comparison in this study of the LDA with systemic functional linguistic analysis (SFL; Halliday, 2004), a qualitative method of text analysis based on a theory of language developed by Halliday (2004). The focus in this study was on convergent evidence (AERA, APA, & NCME, 2014) for evaluating the validity of our results from the LDA analysis. If our interpretation of the LDA results is valid, then other methods assessing a similar construct should provide results that lead to the same interpretation. Thus, we investigated whether human judgement regarding the changes observed from texts based on the SFL framework agreed with the inference regarding the changes observed from the texts through LDA.

In this study, we evaluated the validity by analyzing students' written answers to CR items from both a pre-test and a post-test using the LDA model and the SFL framework. This study had two main research questions:
(1) What do the LDA and SFL analyses tell us about students' answers?
(2) What are the similarities and differences of the two analyses?
We first focus on the analysis using each of the two methods to investigate what we can learn from the two methods. Then, we compare results from the two analyses to examine the similarities and differences of the results from the two methods.

The data used for this study were taken from a larger NSF-funded host study, Language-Rich Inquiry Science for English Language Learners (LISELL; Buxton, Allexsaht-Snider, Suriel, Kayumova, Choi, Bouton, & Baker, 2013). Data were collected during the 2012-2013 academic year. The primary objective of this project was to enable middle school students, in schools with rapidly growing numbers of English learners, to take ownership of scientific language through the use of language-rich science investigation practices. As a part of the project, a pre-test and a post-test were administered to 1,581 and 1,767 middle school students, respectively. These students were a mix of native English speakers and English learners, with the vast majority (>95%) of English learners being native speakers of Spanish. The tests consisted of CR items designed to measure understanding of science investigation practices. Two non-overlapping samples were selected from the students for the LDA and SFL analyses. The sample of students for the LDA analysis included both native English speakers and English learners, while the sample for the SFL analysis was drawn only from the English learners in the population. As the tests were in a paper and pencil format, the written responses needed to be hand entered into a machine-readable

format for the LDA analysis. Students chosen for the pre- and the post-test in both samples took the same test.

This paper is organized as follows. We first introduce the LDA model and present the results of the LDA analysis. Next, we introduce the SFL framework, describe the linguistic features that were considered for recoding students' work using the SFL, and present the results of the SFL analysis. Finally, results from the two methods are compared in the Discussion.

2.0 Study 1: Latent Dirichlet Allocation Analysis

2.1 Data

The sample for Study 1 consisted of a sample of 115 middle grades students from the pre-test, drawn from the full sample of 1,581 students, and a sample of 137 students from the post-test, drawn from the full sample of 1,767 students. On the pre-test, the sample included 44 students (38.26%) in Grade 7 and 71 students (61.74%) in Grade 8. There were 61 males (53.04%) and 54 females (46.96%) in the pre-test sample. On the post-test, the sample included 52 students (37.96%) in Grade 7 and 85 students (62.04%) in Grade 8. There were 58 males (42.34%) and 79 (57.66%) females in the post-test sample. The sample of students consisted of native English speakers and English learners.

2.2 Model

LDA is a statistical model that has been used to detect latent topics in a corpus. It also provides proportions of use of each of the topics for each document. In this study, a document refers to the answers a student gave to the items on the test. The LDA model assumes each word in a sentence comes from one of K topics. The topics are defined as multinomial distributions over the vocabulary of a corpus with probabilities of $\gamma_k = (\gamma_{k1}, \gamma_{k2}, \dots, \gamma_{kV})$ for each topic, where V is the size of the vocabulary. The parameter $z_{d,n}$ indicates the membership on the topics for the nth word in document d. A given word can appear on more than one topic. The proportion of topics for document d is represented by the parameter $\eta_d = (\eta_{d1}, \eta_{d2}, \dots, \eta_{dK})$ in the LDA model. The membership variable $z_{d,n}$ follows a multinomial distribution with probabilities of η_d. The variable η_d follows a Dirichlet distribution with parameter α. The set of probabilities for each topic (γ_k) also follows a Dirichlet distribution with parameter β. (Here β is used simply to distinguish this Dirichlet distribution from the Dirichlet distribution of η_d.)

2.3 Estimation

Parameters in the LDA model were estimated using the *lda* package in R (Chang, 2015). This package uses a Monte Carlo Markov chain (MCMC) algorithm with collapsed Gibbs sampling (Chang, 2010; Heinrich, 2009). MCMC has two main phases, a burn-in phase and a post-burn-in phase. The burn-in phase is used to allow the MCMC chain to converge. Once convergence is obtained, the burn-in iterations are discarded and the post-burn-in iterations are used to estimate the parameters of interest. In this study, the number of burn-in iterations was 10,000, and the number of post-burn-in iterations was 20,000 to obtain the posterior means of the parameter estimates.

Before conducting the LDA analysis, preprocessing of data was performed. First, stop words were removed. Stop words are words that are considered as not contributing to the meaning of the document. Second, all words were converted to lower case. Third, all plural words were converted to singular, and all past tense verbs were converted to present tense to avoid inaccuracies in classification of words due to morphological variances. (A detailed discussion of preprocessing of data in LDA can be found in Boyd-Graber, Mimno, and Newman (2014).)

The Dirichlet parameters, α and β, have an effect on topic proportion distribution (η_d) and the distribution of words in a topic (γ_k), respectively. A small α will cause documents to be dominated by a fewer number of topics. Likewise, a small β will cause topics to be distinctive. The LDA model assumes that the α and β parameters are known; that is, they are not estimated during the LDA analysis. There are several suggestions regarding the choice of the parameters (e.g., Chang, 2010; Griffiths & Steyvers, 2004; Kwak et al., 2017). In this study, the values chosen for α and β were $1/K$ (Chang, 2010) and 0.01 (Griffiths, Steyvers, & Tenenbaum, 2007), respectively.

2.4 Empirical Data Analysis Results

The pre- and post-test data were combined and then analyzed using the LDA model. In this study, an exploratory approach was used to determine the best fitting LDA model. Candidate LDA models with two-, three-, four-, five-, and six-topics were compared. After choosing the best fitting model, each of the topics in the model used for this study was described, and students' use of words from each topic was compared between the pre-test and the post-test.

Table 1 presents descriptive statistics for the data from the pre- and post-test. The term "All" in the heading of the first column indicates the combined pre-test and post-test data. *V* and *L* refer to the vocabulary size and the average

number of words per document, respectively. The descriptive statistics indicate that both vocabulary size and average number of words used per document increased from pre-test to post-test.

Table 1

Descriptive Statistics for the LISELL Data

	Sample size	*V*	*L* (*SD*)
Pre-test	115	358	94.94 (30.71)
Post-test	137	404	98.72 (34.30)
All	252	532	98.56 (32.99)

Notes. *V* is the vocabulary size, *L* is the average number of words per document, and *SD* is the standard deviation of the mean.

Determining the best fitting LDA model is an open question (Glynn, Tokdar, Banks, & Howard, 2015). Several studies have used fit indexes to determine the number of topics (e.g., Blei et al., 2003; Griffiths & Steyvers, 2004; Lauderdale & Clark, 2014). However, Chang, Gerrish, Wang, Boyd-Graber, and Blei (2009) have shown that the statistically best fitting model of a corpus often did not agree with human judgement regarding the corpus. Grimmer and Stewart (2013) suggested use of human judgement to decide the final model, rather than relying on fit statistics. Further, Quinn, Monroe, Colaresi, Crespin, & Radev (2010) considered substantive and conceptual meanings of topics as the primary criteria for selecting the number of topics. In this study, we set the number of topics as three after examining the results from the candidate LDA topic models. The interpretability of topics was considered as the primary criterion for selection of the model. Topics from the LDA models that had more than 3 topics appeared to duplicate one of the topics in the 3-topic model.

Table 2 presents the 15 words that had the highest proportions of occurrence in a document for each topic. Most of the words in Topics 1, 2, and 3 can be characterized as everyday language, general (non-content area specific) academic language, and discipline-specific language, respectively. The three topics were therefore labeled as “preponderance of everyday language,” “preponderance of general academic language,” and “preponderance of discipline-specific language,” respectively. Everyday language refers to words that students use in their everyday lives. General academic language consists of terms that are commonly used across disciplines, and discipline-specific language

consists of terms that are associated with a particular discipline (Nagy & Townsend, 2012).

Table 2

Proportions of Top 15 Words Used in Each Topic

Rank Order	Topic 1 Preponderance of Everyday language	Proportion in Topic 1	Topic 2 Preponderance of General Academic Language	Proportion in Topic 2	Topic 3 Preponderance of Discipline Specific Language	Proportion in Topic 3
1	fish	0.040	variable	0.086	fish	0.050
2	weight	0.036	change	0.062	energy	0.042
3	water	0.036	fish	0.052	water	0.037
4	salt	0.033	boil	0.046	weight	0.035
5	because	0.030	independent	0.042	salt	0.034
6	if	0.028	water	0.033	small	0.032
7	more	0.027	dependent	0.029	same	0.021
8	eat	0.026	different	0.029	boil	0.020
9	lift	0.025	effect	0.028	increase	0.020
10	make	0.021	cause	0.028	amount	0.020
11	algae	0.019	salt	0.028	decrease	0.019
12	shadow	0.016	disease	0.026	kinetic	0.019
13	small	0.015	small	0.023	algae	0.018
14	hold	0.015	bottle	0.021	potential	0.017
15	stronger	0.015	die	0.019	because	0.016

Table 3 presents an example sentence that shows which word came from which topic. Each word in the sentence belongs to one of the topics or stop word categories. This example illustrates that each word belongs to one of the topics and that a sentence can consist of words from various topics.

Table 3

Topic Membership from an Example Sentence

Sentence	Index	Words
"The dependent variable is the	Topic 1	if, work, harder, time
outcome of your exercising, so	Topic 2	dependent, variable, change
if you work harder next time,	Topic 3	outcome, exercise
this will change."	Stop words	the, is, so, you, next, this, will, of

Figure 1 illustrates where students fell with respect to Topics 1 and 3 when answering the CR items on both the pre- and the post-test. Each dot in the two plots represents one student. The values on the X and Y axes represent the proportions of words for Topic 1 (η_1) and Topic 3 (η_3), respectively. The proportions for Topic 2 (η_2) can be obtained as $\eta_2 = 1 - \eta_1 - \eta_3$. Each corner of Figure 1, except the upper right corner, represents strong use of the corresponding topic. So, a student located in the upper left corner would have topic proportions of (0,0,1), indicating the student always used the words from Topic 3 when answering the CR items. Similarly, a student located in the lower right corner would have topic proportions of (1,0,0), indicating the student always used the words from Topic 1. It should be noted that the dots in the figure need to be interpreted as joint proportions. Therefore, if one topic proportion is high, then the other topic proportions should be low.

In Figure 1, the majority of students were located along the diagonal of the plot on both the pre-test and the post-test. This indicates that they used words mostly from Topics 1 and 3, particularly those students in the center part along the diagonal. Students that were off the diagonal used words from Topic 2 for some portion of their answers. Thus, dots in the upper left part of the plot that are off the diagonal indicate students who mostly used words from Topics 2 and 3, and dots in the lower right part of the plot that are off the diagonal indicate students who mostly used words from Topics 1 and 2. It is evident from the plots that students were more spread out toward the upper left corner on the post-test than students on the pre-test. This indicates that on the post-test students tended to use more words from Topic 3 (discipline-specific words) compared to the two other topics.

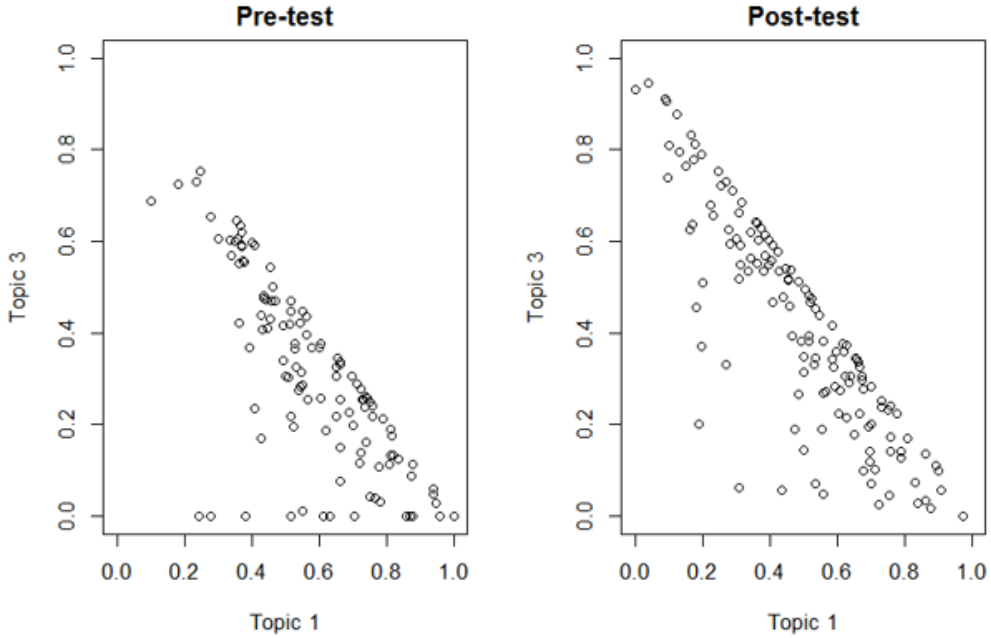

Figure 1. Topic proportions for students on the pre- and post-test. (The students in the pre-test plot and the students in the post-test plot are different.)

2.5 Conclusions for Study 1

The LDA model with three latent topics was used to describe topics underlying students' responses to CR items. Results in Figure 1 show how the topic proportions changed from pre-test to post-test. The three latent topics were characterized as preponderance of everyday language (Topic 1), preponderance of general academic language (Topic 2), and preponderance of discipline-specific language (Topic 3). Students differed in the extent to which they used each of these topics. Overall, students tended to use words from Topics 1 and 3 together. On the post-test, students tended to use more words from Topic 3 compared to usage on the pre-test.

3.0 Study 2: Systemic Functional Linguistic Analysis

3.1 Data

The data for Study 2 were taken from the same NSF-funded host study as for study 1. The sample for Study 2 consisted of a sample of 45 students from the pre-test, drawn from the full sample of 1,581 students, and a sample of 45

students from the post-test, drawn from the full sample of 1,767 students. The sample was restricted to students who self-identified as native speakers of Spanish, as this is the focal population for the host study. On the pre-test, the sample included eight students in grade 6 (18%), 19 students (42%) in Grade 7 and 18 students (40%) in Grade 8. There were 20 males (44%) and 25 females (56%) in the pre-test sample. On the post-test, the sample included six students in grade 6 (13%), 19 students (42%) in Grade 7 and 20 students (45%) in Grade 8. There were 20 males (44%) and 25 (56%) females in the post-test sample. Each student assessment consisted of six CR items used to assess three science investigation practices: coordinating hypothesis, observation and evidence; controlling variables; and explaining cause and effect relationships. Three of the CR items (one for each practice) were selected for the SFL analysis.

3.2 Method

The focus in Study 1 on students' word choices represents one important aspect of students' learning to use the language of science. In Study 2, systemic functional linguistic analysis was used to investigate the degree to which the students were adopting additional key features of scientific language. Building on the work of the linguist M.A.K. Halliday (1994), systemic functional linguistics uses predominantly qualitative approaches to explore how people make linguistic choices to better communicate their ideas to their intended audience. In addition, SFL theory is concerned with analyzing how language is organized for use in different contexts and how it is mediated by disciplinary and institutional discourses (Eggins, 1994). Consequently, SFL "sees the language system as a set of options available for construing different kinds of meaning" (Schleppegrell, 2004, p. 7). Linguists who have applied this theory to the language of science (e.g., Fang & Schleppegrell, 2008) have identified the following four linguistic features as particularly important for effective scientific communication: technical vocabulary usage; grammatically stable linguistic classes; high lexical density; and rheme to theme structure. We limit our discussion here to the two features that are most closely related to Study 1, technical vocabulary usage and high lexical density. Details of how these features were analyzed in the student samples are discussed below.

3.3 Results

3.3.1 Technical vocabulary usage. Technical vocabulary from an SFL perspective corresponds closely to the discipline-specific vocabulary discussed in Study 1, and is considered to include vocabulary that is unique to the context of

science (e.g., photosynthesis, insulator), as well as vocabulary that has both an everyday and a science-specific usage if used appropriately in the scientific sense (e.g., matter, conduct). To analyze the technical vocabulary used by students on the assessment items, we began by counting all uses of technical vocabulary on each focal question. Next, we conducted frequency counts of each technical vocabulary word used in responses to each focus item. Table 4 shows the three most common technical vocabulary words used in student responses to each question (in decreasing frequency of use), as well as the means and standard deviations of technical vocabulary words used by students for that item on both pre- and post-test.

Table 4

Patterns of Technical Vocabulary Usage

Question category	Pre-test (*M*; SD)	Post-test (*M*; SD)
Controlling variables	Weight Strength Experiment (3.9; 2.8)	Weight Variable Independent (4.9; 3.0)
Coordinating hypothesis and evidence	Temperature Dissolve Experiment (1.7; 1.8)	Temperature Hypothesis Observation (3.5; 1.9)
Explaining cause & effect	Algae Decrease Effect (2.0; 1.9)	Algae Effect Decrease (4.1; 2.9)

We noted two patterns in students' uses of technical vocabulary that seemed to be consistent across the different questions, and therefore, across the different investigation practices. First, we noted a difference in the nature of the technical vocabulary words used most frequently between pre-test and post-test. On the pre-test, the most common technical vocabulary words used were those related to the science content of the question, for example, words related to weight training for the variables question, words related to states of matter for the hypothesis question, and words related to a pond ecosystem for the cause and effect question. In contrast, on the post-test, the most frequently used technical vocabulary words were more often related to science investigation practices, for example, variable, hypothesis, and effect. Second, there was growth in the average number of technical vocabulary words used from the pre-test to the post-test. For two of the three items analyzed (coordinating hypothesis and evidence & explaining cause and effect), the mean number of technical vocabulary words was

more than doubled. Further, a two-tailed *t*-test showed that the differences in the mean number of technical vocabulary words for the two items were statistically significant; $t(88) = -4.61,\ p < .01$ for the coordinating hypothesis and evidence item and $t(88) = -4.06,\ p < .01$ for the explaining cause and effect item.

3.3.2 Lexical density. Lexical density is a measure of the ratio of content words to grammatical words in any given text (spoken or written). Content words are defined as the words that provide fundamental meaning to an utterance by describing the content of what is being said or written. These include most nouns (except pronouns), most adjectives, most verbs (except auxiliary verbs), and most adverbs. Grammatical words (also sometimes called functional words) include pronouns, prepositions, conjunctions, auxiliary verbs (e.g., can, could, will), pro-form adverbs and adjectives, determiners (e.g., "a," "the," "my") and interjections (e.g., "wow"). These grammatical words are closely related to the "stop words" that were removed in the Study 1 (LDA) analysis.

The average ratio of content words to grammatical words, known as lexical density, is approximately 0.5 or 50% in standard non-technical written text in both English and Spanish. The ratio for everyday spoken conversation is typically less than 50%, and for technical or specialized academic writing (such as science texts), it is typically greater than 50%. Thus, an increase in lexical density can be interpreted as a shift from more conversational or everyday language to more technical or academic language. We calculated the lexical density for each student's response to each focal question by dividing the number of content words in the response by the total number of words and expressing it as a percentage. Table 5 shows the mean and standard deviation of lexical density as a percentage for each question at pre-test and post-test. A small but consistent increase in the lexical density can be seen across students' responses between the start and the end of the year. A two-tailed *t*-test showed that only the difference in explaining cause and effect category was statistically significant; $t(88) = -2.67,\ p < .01$. While a change of 2-4% in lexical density is modest, it is still meaningful both because it was consistent across questions and because lexical density has a fairly restrictive range, rarely varying outside of the 45%-60% range, meaning that large changes are unlikely. Thus, when comparing the pre- and post-tests, students' written responses on all three questions analyzed showed a trend from language usage that was more conversational to language usage that was more academic.

Table 5

Patterns of Lexical Density as a Percentage of the Sample (Standard Deviation in Parentheses)

Question category	Lexical Density on Pre-test	Lexical Density on Post-test
Controlling variables	50.7 (7.1)	52.7 (7.2)
Coordinating hypothesis and evidence	51.6 (5.9)	53.7 (6.0)
Explaining cause & effect	50.6 (5.0)	54.7 (9.0)

3.4 Conclusions for Study 2

Through the SFL analysis we were able to observe and interpret students' shifts in technical vocabulary usage and lexical density between pre- and post-test. In terms of technical vocabulary usage, the SFL analysis shows that students moved from using science words that are commonly associated with science content to technical words that are associated with scientific processes. This seems to indicate a shift in what students feel is important to communicate when describing a science investigation, with less emphasis placed on describing science as a body of knowledge and more emphasis placed on describing science as a process of inquiry. Regarding lexical density, results show that students' written responses present an increasing usage of content words (words that communicate the content of the utterance). Thus, students' written responses on the post-test were more closely aligned with a primary discursive pattern of the language of science, namely, a high lexical density. We can conclude that both the vocabulary and the structure of these students' science writing were better aligned with the communicative norms of science on the post-test than they were on the pre-test.

4.0 Discussion

In this study, two methods of analysis were performed using student written responses to CR items. One was based on a statistical analysis using the latent Dirichlet allocation model (LDA); the other was based on a qualitative analysis, systemic functional linguistic analysis (SFL).

In the LDA analysis, students tended to have higher proportions of use of Topic 3 (discipline-specific words) on the post-test than on the pre-test. As noted earlier, discipline-specific vocabulary refers to language that is specifically related

to the discipline being discussed. In the case of the LISELL host study, the discipline being studied was a focus on science and engineering practices used to engage in science investigations. Increase in use of Topic 3 words from pre-test to post-test relative to the other two topics reflected students' increased application of discipline-specific terms they had learned during the year. This was consistent with the instructional intervention in the host study. Previous research has demonstrated that use of words from Topic 3 and overall achievement score on the test were related (Kim et al., 2017). Results from that study found a high relationship ($r = .70$) between number of words from Topic 3 and students' scores on the post-test. Correlations for the two other topics were low ($r = .24$) for Topic 1 or negative ($r = -.22$) for Topic 2. The scoring of the student responses for all tests in the larger host study, i.e., those used in both Study 1 and Study 2, was done by trained raters based on a rubric prior to the LDA or SFL analyses.

Results from a previous SFL analysis of LISELL data from the prior year in Buxton, Allexsaht-Snider, Aghasaleh, Kayumova, Kim, Choi, & Cohen (2014) indicated findings similar to those in the present analysis. On the pre-test, the most common technical vocabulary used related to the science content of the questions, while on the post-test, students increased the quantity of technical vocabulary used and improved the quality of the technical vocabulary they selected to more clearly express ideas related to the science inquiry practices that were the focus of the questions (e.g., hypotheses, effect and observation). In the present study, the following were two key findings from the SFL analysis of students' written responses: 1) an increase at the end of the year in the appropriate use of technical vocabulary useful for engaging in science investigations; and 2) an increase in lexical density at the end of the year, representing a more academic register in students' writing after the intervention.

LDA is a soft clustering method, meaning that each word can belong to more than one topic. For example, the words *fish*, *water*, and *salt* appeared in both Topic 1 and Topic 3. Topic 3 was largely dominated by discipline-specific words, but some words classified in that topic were not discipline-specific. On the other hand, the SFL analysis requires manual recoding, and thus involves human judgement throughout the entire process. Consequently, SFL results tend to be more directly related to the different themes associated with meaning making detected in the data.

In this study, as results from both Studies 1 and 2 suggest, students responded using more discipline-specific language on the post-test than on the pre-test. In addition, the results of the two linguistic features from the SFL analysis, technical vocabulary usage and lexical density, can be explained using the results from the LDA analysis. For the technical vocabulary usage, the three

most common technical vocabulary words used on the post-test determined by the SFL analysis were often the words that characterize Topic 3 from the LDA analysis. Many of the technical vocabulary words on the post-test (see Table 4) had high proportions of occurrence in Topic 3. When *hypothesize* was used, for example, 84% of the time it was assigned to Topic 3, 16% of the time to Topic 2, and 0% of the time to Topic 1. Similarly, more than 80% of the time the words *temperature* and *observation* were assigned to Topic 3, and the rest of the time they were assigned to Topic 2. Further, the word *weight* had the third highest proportion in Topic 3, and the word *decrease* only appeared in Topic 3.

The lexical densities of the three different types of question on the test increased from pre-test to post-test. The increment of lexical density in responses to *Explaining cause & effect* questions, for example, was higher than that of the other types of question. This increasing pattern could also be found in the LDA analysis results. Most of the words that had the highest proportions in Topic 3 were key words in responding to the questions related to *Explaining cause & effect*. Among the top 15 words for Topic 3 in Table 2, 60% of the words (including *fish*, *energy*, *water*, *small*, *increase* … etc.) could be used for describing the answers to the questions. Thus, the increasing trend in lexical density of responses to *Exploring cause & effect* questions was in line with an increasing trend of use of Topic 3 in student answers.

The relationship between the results from the LDA and SFL analyses discussed above shows that the interpretation of the LDA analysis results corresponded to the results of the SFL analysis. The interpretation of the LDA analysis results was based on our labeling of the three topics, and the changes in topic proportions were interpreted based on these topic labels. This interpretation was supported by the SFL analyses of the technical vocabulary usage and lexical density, which were closely related to the discipline-specific language topic.

Compared to previous work with LDA (e.g., Blei et al., 2003; Griffiths & Steyvers, 2004; Lauderdale & Clark, 2014), the sample size used for the LDA analysis in this study was relatively small. One result of the small sample size was that the rank of the words in Table 2 appeared to be influenced by the list of stop words, although the general interpretation of each topic was consistent with the current ones. Further, Kwak et al. (2017) used paired data from the same host study with a sample size similar to that used in this study. The contents of topics found in Kwak et al. (2017) were also very similar to the contents of topics in this study. Further, Kwak et al. detected significant differences in the change in use of topics from the pre-test to the post-test even with the small sample. Thus, although the number of documents considered in this study was small, the results were stable.

5.0 Conclusions

In this study, we explored what we could learn from student written responses to CR items by using LDA and SFL analyses with samples from the same host study. Using LDA analysis, we found topics underlying the written responses, and we examined how the use of topics changed from pre-test to post-test. Using SFL analysis, we examined two linguistic features of the written responses – technical vocabulary usage and lexical density – and how the two linguistic features changed from pre-test to post-test. The LDA analysis showed that student responses in the post-test tended to have a higher discipline-specific topic proportion than the responses in the pre-test. The SFL analysis showed that students moved from using words related to science content to using technical words related to scientific processes. Further, the lexical density of student answers was higher on the post-test than on the pre-test. The technical words were often the words that characterized the discipline-specific language topic (Topic 3) detected by the LDA, and the item category that showed a statistically significant increase in lexical density was also closely related to the discipline-specific language topic from the LDA.

The two approaches, LDA and SFL, revealed method-specific aspects of the written response data, but both also showed that students shifted from using less technical or academic words to more technical or academic words. The consistency of the results from the two approaches provided evidence for the inferences that we made using the LDA results. The types of evidence required for validation are determined by the arguments made (Kane, 2016). We noted in the LDA analysis that students in the post-test tended to use more discipline-specific words in their answers than students in the pre-test. We provided support for the validation arguments to this effect in the comparison of the results of the LDA and SFL analyses as described in the Discussion section. LDA analysis can be used to summarize a large amount of text data in a shorter time than is possible with the manual coding required by the SFL, but the analysis is not as fine-grained as that which can be obtained from manual coding. Grimmer and Stewart (2013) suggest that automated content analysis has the potential to be either misleading or incorrect. On the other hand, because SFL analysis requires human coding, it provides more detailed information about the meaning making use of language, thereby producing finer grained results than LDA. As used in this study, the SFL analysis provided a useful criterion against which to compare the validity of the results from the LDA.

Finally, we note that in science education there is ongoing debate about the relative value of teaching technical vocabulary or teaching other aspects of scientific language, such as the role of lexical density. Combining LDA and SFL

analyses allowed us to consider how both the vocabulary and the structure of students' writing are intertwined meaning-making resources. Thus, findings from the present study seem to imply that the two are mutually supportive and should both be taught in the science classroom.

Author Biographies

Seohyun Kim is a doctoral student of Educational Psychology specializing in Quantitative Methodology at the University of Georgia. Her research focuses on statistical modeling using latent variables in education.

Minho Kwak is a doctoral student of Quantitative Methodology at the University of Georgia. His research focuses on analyzing constructed response items using topic modeling and item response theory.

Lourdes Cardozo-Gaibisso is a former middle school teacher in her native Uruguay. Lourdes is currently completing her Ph.D. at the University of Georgia's College of Education in the Department of Language and Literacy Education. She is a Research Assistant for the National Science Foundation funded Language-Rich Inquiry Science with English Language Learners through Biotechnology project, and currently serves as President of the Georgia Association of Multilingual and Multicultural Education.

Cory Buxton is a professor of Science Education in the Department of Educational Theory and Practice at the University of Georgia. His research focuses on fostering more equitable science learning opportunities for all students and especially for English learners and immigrant students.

Allan S. Cohen is a professor of Educational Psychology specializing in Quantitative Methodologies at the University of Georgia. His research focuses on use of quantitative methodologies for understanding the response processes of examinees in educational and psychological examinations.

Acknowledgements

This research was supported in part by the National Science Foundation (NSF) grant at the University of Georgia under No. 1019236.

References

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education (2014). *Standards for educational and psychological testing*. Washington, DC: American Educational Research Association.

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent Dirichlet allocation. *Journal of Machine Learning research*, *3*(Jan), 993-1022.

Boyd-Graber, J., Mimno, D., & Newman, D. (2014). Care and feeding of topic models. In E.M. Airoldi, D.M. Blei, E.A. Erosheva, and S.E. Fienberg (Eds.) *Handbook of mixed membership models and their applications* (pp. 225-254). Chapman and Hall/CRC. doi:10.1201/b17520-16

Buxton, C., Allexsaht-Snider, M., Suriel, R., Kayumova, S., Choi, Y. J., Bouton, B., & Baker, M. (2013). Using educative assessments to support science teaching for middle school English-language learners. *Journal of Science Teacher Education*, *24*(2), 347-366.

Buxton, C., Allexsaht-Snider, M., Aghasaleh, R., Kayumova, S., Kim, S. H., Choi, Y. J., & Cohen, A. (2014). Potential benefits of bilingual constructed response science assessments for understanding bilingual learners' emergent use of language of scientific investigation practices. *Double Helix*, *2*, 1-21.

Chang, J. (2010). Not-so-latent Dirichlet allocation: Collapsed Gibbs sampling using human judgments. In C. Callison-Burch & M. Dredze (Eds.), *Proceedings of the North American Chapter of the Association for Computational Linguistics Human Language Technologies 2010 Workshop on Creating Speech and Language Data with Amazon's Mechanical Turk* (pp. 131-138). Stroudsburg, PA: Association for Computational Linguistics.

Chang, J. (2015). lda: Collapsed Gibbs sampling methods for topic models. R package version 1.4. 2.

Chang, J., Gerrish, S., Wang, C., Boyd-Graber, J. L., & Blei, D. M. (2009). Reading tea leaves: How humans interpret topic models. In Y. Bengio, D. Schuurmans, J. D. Lafferty, C. K. I. Williams, & A. Culotta (Eds.), *Advances in neural information processing systems* (pp. 288-296). Red Hook, NY: Curran Associates, Inc.

Chen, Y., Yu, B., Zhang, X., & Yu, Y. (2016, April). Topic modeling for evaluating students' reflective writing: A case study of pre-service teachers' journals. In *Proceedings of the Sixth International Conference on Learning Analytics & Knowledge* (pp. 1-5). ACM.

Eggins, S. (1994). *An introduction to systemic functional linguistics*. London: Pinter Publishers.

Erosheva, E., Fienberg, S., & Lafferty, J. (2004). Mixed-membership models of scientific publications. *Proceedings of the National Academy of Sciences*, *101*(suppl 1), 5220-5227.

Fang, Z. & Schleppegrell, M. J. (2008). *Reading in secondary content areas: A language-based pedagogy*. Ann Arbor: University of Michigan Press.

Glynn, C., Tokdar, S. T., Banks, D. L., & Howard, B. (2015). Bayesian analysis of dynamic linear topic models. *arXiv preprint arXiv:1511.03947*.

Griffiths, T. L., & Steyvers, M. (2004). Finding scientific topics. *Proceedings of the National Academy of Sciences*, *101*(suppl 1), 5228-5235.

Griffiths, T. L., Steyvers, M., & Tenenbaum, J. B. (2007). Topics in semantic representation. *Psychological Review*, *114*(2), 211-244.

Grimmer, J. (2009). A Bayesian hierarchical topic model for political texts: Measuring expressed agendas in Senate press releases. *Political Analysis*, *18*(1), 1-35.

Grimmer, J., & Stewart, B. M. (2013). Text as data: The promise and pitfalls of automatic content analysis methods for political texts. *Political Analysis*, *21*(3), 267-297.

Halliday, M. A. K. (1994). *An introduction to functional grammar*. London, UK: Edward Arnold.

Halliday, M. A. K. (2004). *The language of science.* London, UK: Continuum.

Heinrich, G. (2009). Parameter estimation for text analysis. *University of Leipzig, Technical Report*.

Huerta, M., Lara-Alecio, R., Tong, F., & Irby, B. J. (2014). Developing and validating a science notebook rubric for fifth-grade non-mainstream students. *International Journal of Science Education*, *36*(11), 1849-1870.

Kane, M. (2016). Validation strategies: Delineating and validating proposed interpretations and uses of test scores. In S. Lane, M. Raymond, & T. M. Haladyna (Eds.) *Handbook of test development* (2nd ed., pp. 64-80). New York, NY: Routledge.

Kim, S., Kwak, M., & Cohen, A. S. (2017). A Mixture Partial Credit Model Analysis Using Language-Based Covariates. In *Quantitative Psychology* (pp. 321-333). Springer International Publishing.

Kwak, M., Kim, S., & Cohen, A. (2017, January) *Mining students' constructed response answers.* Paper presented at the International Conference on Writing Analytics, Tampa, FL.

Lauderdale, B. E., & Clark, T. S. (2014). Scaling politically meaningful dimensions using texts and votes. *American Journal of Political Science*, *58*(3), 754-771.

Nagy, W., & Townsend, D. (2012). Words as tools: Learning academic vocabulary as language acquisition. *Reading Research Quarterly*, *47*(1), 91-108.

Paul, M. J., & Dredze, M. (2011). You are what you Tweet: Analyzing Twitter for public health. *ICWSM*, *20*, 265-272.

Phan, X. H., Nguyen, L. M., & Horiguchi, S. (2008, April). Learning to classify short and sparse text & web with hidden topics from large-scale data collections. In *Proceedings of the 17th International Conference on World Wide Web* (pp. 91-100). ACM.

Quinn, K. M., Monroe, B. L., Colaresi, M., Crespin, M. H., & Radev, D. R. (2010). How to analyze political attention with minimal assumptions and costs. *American Journal of Political Science*, *54*(1), 209-228.

Rescorla, L., Mirak, J., & Singh, L. (2000). Vocabulary growth in late talkers: Lexical development from 2; 0 to 3; 0. *Journal of Child Language*, *27*(2), 293-311.

Ruiz-Primo, M. A., Li, M., Ayala, C., & Shavelson, R. J. (2004). Evaluating students' science notebooks as an assessment tool. *International Journal of Science Education*, *26*(12), 1477-1506.

Schleppegrell, M. (2004). *The language of schooling: A functional linguistics perspective.* Mahwah, N.J: Lawrence Erlbaum.