Students' Academic Language Use when Constructing Scientific Explanations in an Intelligent Tutoring System

Haiying Li, Janice Gobert, Rachel Dickler, and Natali Morad

Rutgers University, New Brunswick NJ 08901, USA

Abstract. In the present study, we first examined the formality and use of academic language in students' scientific explanations in the form of written claim, written evidence, and written reasoning (CER). Middle school students constructed explanations within an intelligent tutoring system after completing a virtual science inquiry investigation. Results showed that students tended to use more formal, academic language when constructing their evidence and reasoning statements. Further analyses showed that both the number of words and pronouns used by students were significant predictors for the quality of students' written claim, evidence, and reasoning statements. The quality of claim statements was significantly reduced by the lexical density (type-token ratio) of student writing, but quality of reasoning significantly increased with lexical density. The quality of evidence statements increased significantly with the inclusion of causal and temporal relationships, verb overlap as captured by latent semantic analysis, and inclusion of descriptive writing. These findings indicate that students used language differently when constructing their claim, evidence, and reasoning statements. Implications for instruction and scaffolding within intelligent tutoring systems are discussed in terms of how to increase students' knowledge of and use of academic language.

Keywords: Science Inquiry, Scientific Explanations, Language Processing.

1 Introduction

The Common Core State Standards for English Language Arts [2] require that students develop academic writing skills. This requirement is expanded upon in the College and Career Readiness Anchor Standards for Reading and writing in Grades K-12 [23]. The specific abilities outlined in the standards include "write arguments to support claims in an analysis of substantive topics or texts, using valid reasoning and relevant and sufficient evidence" [23, p. 18]. Moreover, the CCSS-ELA demands that students use academic discourse in their writing. As of 2011, however, the National Assessment of Educational Progress (NAEP) in writing [22] reported that secondary students face considerable challenges in meeting these standards. Unfortunately, there is little research on secondary students' academic language use in the context of science inquiry.

Currently, the instruction and evaluation of academic writing in areas such as science has focused on students' understanding of scientific content and processes [30].

With the development of automated scoring tools, researchers have the potential to examine the role of academic language in student writing in science inquiry. Prior studies, however, have primarily used linguistic features to automatically score the quality of student writing [3, 19, 29] rather than examining the use of academic language. Therefore, the relationship between students' writing in the context of science and use of academic language required by the CCSS-ELA have yet to be investigated.

In this study, we examined students' academic language use within scientific explanations in the format of claim, evidence, and reasoning using the automated text analysis tool, Coh-Metrix [18]. Coh-Metrix automatically scores 100 linguistic features of written text, which is extremely valuable when examining multiple components of student language use. This study contributes to research on scientific explanations, as well as research on students' use of academic writing by addressing the gap in research on the relationship between students' academic language use and writing performance in the context of science inquiry.

This paper has four sections. First, we briefly review studies on academic language, the use of the automated text analysis tool, Coh-Metrix, and scientific explanations. Second, we describe the materials and measures used in the present study. Third, we display results and discuss the findings in terms of the overall level of academic language use and the individual language features used in students' writing. Fourth, we present implications for teachers and researchers, as well as how the results of the present study contribute to instruction and scaffolding of scientific writing within the intelligent tutoring system, Inq-ITS.

1.1 Academic Language

Academic language, also called scientific language, refers to the use of language in scholarly contexts (the classroom, textbooks, etc.) that is often more sophisticated and complex than language used in common day-to-day interactions [26-27]. Snow and Uccelli [26] developed a comprehensive pragmatic-based framework of academic language that groups linguistic features according to: interpersonal stance (e.g., informational or detached versus conversational), information load (e.g., concise information versus redundant repetition), organization of information (e.g., embedded clauses, connective metadiscourse markers), lexical choices (e.g., academic vocabulary versus colloquial expressions), and representational congruence (e.g., grammatical incongruence, such as nominalization, passive voice). In addition to these linguistic features, they included three core domains of cognitive accomplishment involved in academic-language performance: genre mastery (e.g., narration versus explanation), command of reasoning/argumentative strategies (e.g., ways of argumentation and persuasion), and disciplinary knowledge taxonomies (e.g., facts versus constructed knowledge).

This pragmatic-based framework provided a more comprehensive view and some possible measures for each level of the framework. Unfortunately, they failed to provide an automated text analysis tool to extract these linguistic features for educators and researchers. In summary, the current research on academic language still focuses on one or two shallow linguistic levels, such as lexical level and/or syntactic level [4-6]. Some measures were manually analyzed, such as the number of embedded clauses [5]. Even though this comprehensive framework exists and has been applied through manual efforts, it is lacking a corresponding automated tool that would lessen the amount of time and effort spent on text analysis. Both a panoramic theoretical framework and a related automated text analysis tool are needed to enable better research on language use.

Graesser and McNamara [11] developed another framework, used primarily for reading comprehension and writing, to analyze texts on many different levels. The multilevel components of the framework included the: surface code (e.g., words, syntax), explicit textbase (e.g., overlapping of propositions and idea units), situation model (e.g., causal, intentional, temporal, relationships), the genre (e.g., narrative, expository), rhetorical structure, and pragmatic communication level. While utilizing different labels, there is significant overlap between this framework and Snow and Ucelli's [26] framework. Both frameworks cover multiple textual levels, including words (e.g., pronouns, concrete/abstract words, lexical density/diversity), syntax (e.g., embedded words, phrases, and clauses), referential cohesion (e.g., overlapping of propositions and ideas), deep cohesion (e.g., cohesion represented by connectives), and genre (narrative versus informational). Based on their multilevel framework, Graesser and McNamara [11] developed an automated text analysis tool, Coh-Metrix.

1.2 Automated Text Analysis Tool: Coh-Metrix

Coh-Metrix (cohmetrix.com) is a computer-based tool that automates many language- and text-processing mechanisms for hundreds of measures at multiple levels of linguistic analysis, including word characteristics, sentence characteristics, and discourse relationships between ideas [18]. Specifically, Coh-Metrix (3.0) measures include descriptive (e.g., the number of syllables, words, sentences,), word information (e.g., incidence of noun, word frequency, concreteness, imageability), syntactic pattern density (e.g., density of noun phrase), syntactic complexity (e.g., left embedded words before main verbs), connectives (e.g., causal, temporal), lexical density (e.g., type/token ratio), latent semantic analysis, referential cohesion (e.g., noun overlap, argument overlap), and readability (e.g., Flesch-Kincaid grade level). Examination of these individual indices can often provide valuable information in addition to examining whole dimensions. For instance, Wiley et al. [29] found that cohesion, causality, and lexical diversity were correlated with the quality of students' explanatory science essays, but only lexical diversity was a significant predictor. Li, Graesser, and Cai [15] used Coh-Metrix individual indices to automatically evaluate the quality of summaries after reading scientific texts.

Coh-Metrix also includes five major components identified through a principal component analysis performed on 52 individual indices [18]. These five dimensions include: word concreteness (concrete words can evoke mental images and are thus assumed to be more meaningful to the writer relative to abstract words), syntactic simplicity (sentences are constructed with few words and simple, familiar syntactic structures), narrativity (narrative texts tell stories that are familiar to the reader and are closely associated with everyday oral conversation), referential cohesion (high-cohesion writing contains words and ideas that overlap across sentences and the text as a whole, forming threads that connect the explicit textbase), and deep cohesion (causal, intentional, and other types of connectives are taken as evidence that writing reflects a more coherent and deeper understanding)[18].

The five Coh-Metrix dimensions can be examined together in order to uncover information about the overall language style of a text, also referred to as formality [12,13]. Text formality provides information on the overall difficulty of a piece of writing based on the structure and content of the text. Specifically, formality increases with more abstract words, syntactic complexity, high cohesion, and more informational text. Li et al. [13] found that formality as captured by the five Coh-Metrix dimensions was better able to distinguish between text difficulty relative to a more traditional index of formality measured at the surface language level [15]. Overall, Coh-Metrix can provide valuable information related to student writing and language use based on the individual indices, individual dimensions, and comprehensive measure of formality.

1.3 Scientific Explanations

Scientific explanations have been used to evaluate students' science inquiry competencies for analyzing and interpreting data and engaging in argument from evidence [24]. Researchers have assessed written scientific explanations in various forms including within the structure of claim, evidence, and reasoning (CER) [10, 21, 25]. The CER format is based on a modified version of Toulmin's [28] framework for argumentation in which students make a claim, provide evidence for their claim, and provide a justification for how their evidence supports their claim. CER has traditionally been scored according to the accuracy of content, as well as the completeness of claim, sufficient and appropriate use of evidence, and justification for how the evidence supports the claim [20, 16, 17, 25]. To date, however, no studies have comprehensively examined students' claims, evidence, and reasoning statements in terms of academic language used within each component. It is necessary to address this gap in the literature, as student performance on each component of CER has been found to vary according to content quality [20]. It would be valuable to understand how or if language use relates to this variation in performance for each component of CER. Also, even though the content may vary across claim, evidence, and reasoning statements, students should still use consistent academic language across each component.

This study examined students' writing of claim, evidence, and reasoning statements constructed within an intelligent tutoring system according to formality and academic language use. Specifically, this study investigated two research questions:

RQ1: To what extent do students use academic (formal) language to write claim, evidence, and reasoning, respectively?

RQ2: What specific language and discourse features are actually used in students' claim, evidence, and reasoning, respectively?

The first question examines whether students used academic language when they wrote their claim, evidence, or reasoning statements. It is important that students use formal language at the appropriate grade level for each of these statements. Students' claim, evidence, and reasoning writing should also meet the academic language requirements for each component of formality (i.e., high levels of deep cohesion, etc.) [2].

The second question investigates which specific Coh-Metrix indices predict claim, evidence, and reasoning in order to determine the features that contribute to the construction of high quality writing for each component of C, E, and R [16-17]. The measures of academic language and the Coh-Metrix indices that were used in this study are reported in detail in the Methods section.

2 Method

2.1 Participants and Materials

The participants were 293 students in grades seventh through eighth from public middle schools located in Massachusetts, Minnesota, or Oregon. All participants had completed a Density Virtual Lab in Inq-ITS, an inquiry intelligent tutoring system [7-9]. Inq-ITS virtual labs are for science topics in the domains of Life, Earth, and Physical Science. Inq-ITS uses machine learned detectors to automatically score students' inquiry practice competencies as they engage in virtual labs [7]. Each virtual lab involves three to six activities in which students conduct an investigation to address a directed question or goal.

The Density Virtual Lab contains three different activities. The data used in the present study is from the Density, Shape of the Container activity. The goal of the Shape of the Container activity was to "determine how the shape of the container affects the density of the liquid." The three possible container shapes included wide, narrow, or square. In the first stage of the virtual lab, students formed a hypothesis as to whether the density of a liquid would change or not when the container was changed. Students then manipulated a simulation to determine whether the shape of a container impacted a liquid's density. Data from the students' trials were automatically recorded and stored in a data table. In the third stage, students interpreted the data from their data table, made a claim regarding the relationship between the density of a liquid and the shape of a container using a widget (dropdown menu), and selected data to warrant their claim using a separate widget (clickable buttons). In the final stage of the activity, students responded to each of the following three open response prompts (i.e. CER) in order to explain the findings from their investigation: write a sentence that states what you found out about the scientific question you just investigated (Claim), provide and describe scientific evidence from your data table that supports (or refutes) your claim (Evidence), and explain why your evidence supports your claim (Reasoning).

The present study only uses student data from the final stage of the Shape of the Container Lab. Specifically, the data consisted of students' written claim, evidence, and reasoning statements. Even though 293 students completed the activity, some students wrote gibberish responses for their claim, evidence, or reasoning. These gibberish responses were strings of letters such as "jhhhhhhhhhhhh" or "eciu3ghf." Gibberish responses were removed for analyses resulting in: 288 written claim statements, 288 written evidence statements, and 287 written reasoning statements.

2.2 Measures

Scientific Explanations of Claim, Evidence, and Reasoning. Students' written scientific explanations from the final stage of the Density Shape of the Container Activity in the form of a claim, evidence, and reasoning were scored according to fine-grained rubrics developed by Li et al. [16-17].

Students' written claims were scored according to four sub-components, including identifying: the correct independent variable (IV; i.e. shape of the container), the specific shapes of the containers (IVR; i.e. narrow, square, or wide), the correct dependent variable (DV; i.e. density), and whether the density was affected by changing the container shape (DVR; i.e. the density stayed the same). Each subcomponent of claim was worth a maximum of 1 point. Students could receive a maximum total score of 4 points for claim.

Written evidence statements were scored based on three sub-components of: describing at least two relevant trials (Sufficiency), stating numeric data for the mass and volume of a liquid (Appropriate Mass and Volume), and stating numeric data for the density of a liquid (Appropriate Density). Sufficiency was scored out of 2 points. Appropriate Mass and Volume, and Appropriate Density were scored out of 1 point. Students could receive a maximum total score of 4 points for evidence.

Students' written reasoning statements were scored according to five sub-components of: interpreting whether or not the data described in the evidence supports (or refutes) the claim (Connection), describing the IV data and how it was changed (Data IV/IVR), describing the DV (Data DV), describing and interpreting data for the DV (Data DVR), and backing up conclusions with a scientific theory (Theory). The subcomponent of Theory was scored out of 2 points. Sub-components of Connection, Data IV/IVR, Data DV, and Data DVR were scored out of 1 point. Students could receive a maximum total score of 6 points for reasoning.

Academic Language. Academic language of students' written claim statements, written evidence statements, and written reasoning statements was measured by Coh-Metrix formality (12-14, 16-17]. Formality increases with word abstractness, syntactic complexity, expository texts, and high referential cohesion and deep cohesion. The formula used to calculate formality is listed below:

$$\text{Formality} = (\text{deep cohesion} + \text{referential cohesion} - \text{narrativity} - \text{word concreteness} - \text{syntactic simplicity})/5 \qquad (1)$$

Coh-Metrix formality uses the reference corpus TASA (Touchstone Applied Science Associates, now renamed Questar Assessment Inc.) [18], with numbers higher than 0 representing more formal discourse, and numbers below 0 representing more informal discourse. Graesser et al. [12] compared formality in three genres across grades K-12 using 37,650 texts in the TASA corpus and displayed that formality of science reading for middle school students (Grade 6-8) was slightly below 0. This implies that if student formality in writing reaches 0, their use of academic language is equivalent to the academic language used in their informational reading materials.

Individual Coh-Metrix Indices. Students' actual use of language for written claim, written evidence, and written reasoning was measured using individual linguistic features from Coh-Metrix. Coh-Metrix captures and evaluates over 100 individual indices. Not all indices were used to evaluate students' written claim, evidence, and reasoning statements, as there were too few data to apply such a large number of indices. In order to avoid overfitting of our regression model as a result of the small number of data ($N = 285$-288) available in the present study, we followed two regression assumptions. First, if the correlation between each pair of Coh-Metrix indices exceeded the limit of the assumption (less than .70), we removed one variable. We followed the rule that if one index, such as sentence count (the number of sentences in writing) was highly correlated with more than one index, we kept sentence count, but removed other variables. If two indices were highly correlated, we kept the one that was used as a predictor in previous studies. Second, we removed indices whose correlations with the dependent variable were smaller than .30.

The remaining independent indices for written claim scores were: lexical diversity measured by the type-token ratio calculated by the proportion of unique words out of all words (LDTTRa), the density of preposition phrases (DRPP), the incidence of pronouns (WRDPRO), the incidence of first person singular pronouns (WRDPRP1s), word frequency based on the CELEX word data base (WRDFRQc), the mean of polysemy for content words (WRDPOLc), and the Flesch-Kincaid grade level (RDFKGL). The independent indices for written evidence scores were: the number of sentences (DESSC), the average number of words per sentence (DESSL), the standard deviation of the average number of words per sentence (DESSLd), LDTTRa, the incidence of causal verbs and particles (SMCAUSvp), verb overlap based on latent semantic analysis (SMCAUSlsa), temporal cohesion of the text (SMTEMP), the syntactic simplicity based on the number of words that occur before the main verb (SYNLE), syntactic simplicity based on the average number of adjacent sentences with similar syntactic structures (SYNSTRUTa), WRDPRO, the average number of visually descriptive words (WRDIMGc), and the Coh-Metrix readability level (RDL2). The indices for written reasoning scores were: DESSC, DESSL, LDTTRa, a measure of textual lexical diversity based on the MTLD word data base (LDMTLD), the average number of modifiers per noun phrase (SYNNP), and WRDPRO [Graesser et al., 2011; McNamara et al., 2012].

3 Findings and Discussion

3.1 Formality of Claim, Evidence, and Reasoning

To answer the first research question, a One-Way ANOVA was performed to compare the formality of language used in students' claim, evidence, and reasoning statements. An analysis of variance showed a significant effect of explanation component on formality, $F(2, 860) =27.09, p < 0.001, \eta^2 = 0.06$. Post hoc analyses using the Bonferroni criterion for significance with adjustment for multiple comparisons indicated that claim formality (M = -.33, SD = .50) was significantly lower than evidence formality (M = -.03, SD = .82, p < .001, Cohen's d = .43) and reasoning formality (M = .08, SD = .72, p < .001, Cohen's d = .67). No significant difference was found between evidence formality and reasoning formality.

These findings indicate that students used more academic language when they constructed their evidence and reasoning statements relative to their claim statements. The average formality score for evidence and reasoning was about 0 points and the average claim formality score was about -.33 points. Graesser et al [12] indicated that the formality score of science reading texts for middle school students (Grade 6-8) was slightly below 0, whereas a formality score around -.30 points was for grades 2-3. Therefore, when the middle school students (grades 7-8) in the present study generated evidence and reasoning, they used academic language to the same extent encountered in their formal science readings (i.e. textbooks). Students were less likely, however to use academic language when constructing their claims relative to evidence and reasoning. Claims only involve stating a conclusion, whereas evidence involves describing at least two pieces of data and reasoning involves using data to support a claim. This difference between explanatory components may lead to different levels of formality between claim and evidence and reasoning.

A simple, stepwise linear regression with 10-fold cross-validation was calculated to predict the written claim statement scores based on the formality scores. Results showed that formality was not a significant predictor for written claim scores. The same regression analysis was conducted for the written evidence scores and was significant with an R^2 of 0.119. The same regression analysis for the written reasoning scores was also significant with an R^2 of 0.005. The predicted scores of written evidence and written reasoning based on formality are displayed in equations (2) and (3), respectively. These findings further indicated that the more formal language that students used when they generated evidence and reasoning, the higher the scores they received based on content.

$$\text{Written evidence scores} = 1.33 + 0.55 \times \text{Formality} \quad (2)$$

$$\text{Written reasoning scores} = 2.45 + 0.29 \times \text{Formality} \quad (3)$$

Formality scores were computed by five major Coh-Metrix dimensions (word concreteness, syntactic simplicity, referential cohesion, deep cohesion, and narrativity). To further examine the extent to which student language differed across claim, evidence, and reasoning, three multiple linear regressions with stepwise 10-fold cross-validation were performed. Specifically, each of the five Coh-Metrix dimensions was used to predict scores on written claim, evidence, and reasoning, respectively. A significant regression equation for written claim was found with an R^2 of 0.073. Syntactic Simplicity, Deep Cohesion, and Narrativity were significant predictors, but Referential Cohesion and Word Concreteness were not. Results showed a significant regression equation with an R^2 of 0.273. Significant predictors for evidence were Syntactic Simplicity, Referential Cohesion, Deep Cohesion, and Narrativity, but not Word Concreteness. Results of the analyses for reasoning showed the same significant predictors as evidence and a significant regression equation with an R^2 of 0.169. The predicted scores of written claim, written evidence, and written reasoning based on the five Coh-Metrix dimensions are displayed in equations (4), (5), and (6), respectively.

$$\text{Written claim scores} = 2.61 - 0.30 \times \text{Syntactic Simplicity} - 0.07 \times \text{Deep Cohesion} - 0.10 \times \text{Narrativity} \quad (4)$$

$$\text{Written evidence scores} = 0.84 - 0.29 \times \text{Syntactic Simplicity} + 0.23 \times \text{Referential Cohesion} + 0.09 \times \text{Deep Cohesion} - 0.15 \times \text{Narrativity} \quad (5)$$

$$\text{Written reasoning scores} = 2.33 - 0.44 \times \text{Syntactic Simplicity} + 0.12 \times \text{Referential Cohesion} + 0.09 \times \text{Deep Cohesion} - 0.11 \times \text{Narrativity} \quad (6)$$

Findings from these find-grained analyses were consistent with those from analyses for overall formality. Students' quality of written evidence and reasoning increased with more complex sentence structures, cohesion, and expository/informational writing style. Word concreteness was not a significant predictor of evidence or reasoning performance because the density activity did not involve many extremely abstract or concrete words. Written claim showed a different pattern relative to evidence and reasoning. Students' quality of written claim increased with more complex sentence structures and expository style, but less deep cohesion.

The finding that the quality of claim decreases with high deep cohesion is contradictory to written evidence and written reasoning. We examined the written claims where students achieved high scores on claim content, but low scores on the dimension of deep cohesion. We found that claims that received low content scores contained some causal connectives (e.g., so, cause), which dramatically increased the scores of deep cohesion to above 6.00 points (e.g. "Cause I said so and it worked"), even though the content in these claims was inaccurate, irrelevant, or mistakenly spelt. On the other hand, we examined written claims that received high scores and found that students did not specify the causal relationship between the IV and DV (e.g., "i found out that the shape of container going from narrow to wide doesn't change its density"), which led to low deep cohesion scores (less than −4.00 points). We found that only 11% (*N* = 33) of students' written claims (*N* = 288) showed high Deep Cohesion above the average score of 0 points. These findings reveal that students were able to generally articulate the relationship between the IV and DV in their claims, but did not use causal connectives to state this causal mechanism. Students need to be instructed on how to use appropriate causal connectives in order to effectively and explicitly express causal relationships when they generate a claim. It is important that claims involve causal language as used in evidence and reasoning, as claims specifically involve drawing a conclusion in regard to the relationship between variables.

3.2 Student Language use for Claim, Evidence, and Reasoning

In order to investigate the second research question, individual Coh-Metrix indices were used to predict students' writing performance on claim, evidence, and reasoning, respectively. The process used to extract individual indices was based on regression assumptions, as detailed in the Individual Coh-Metrix Indices measures section.

A multiple linear regression with stepwise 10-fold cross-validation was used to predict written claim performance based on seven individual Coh-Metrix indices (LDTTRa, DRPP, WRDPRO, WRDPRP1s, WRDFRQc, WRDPOLc, and RDFKGL). Five indices were significant predictors; DRPP and WRDPOLc were not significant. A significant regression model was found with an R^2 of 0.332. The same analyses for written evidence were conducted based on 12 individual Coh-Metrix indices (DESSC, DESSL, DESSLd, LDTTRa, SMCAUSvp, SMCAUSlsa, SMTEMP, SYNLE, SYNSTRUTa, WRDPRO, WRDIMGc, and RDL2). Eight indices were significant predictors; DESSLd, LDTTRa, SyNSTRUTa, and RDL2 were not significant. A significant regression equation was found with an R^2 of 0.44. The same analyses for written reasoning were performed based on six individual Coh-Metrix indices (DESSC, DESSL, LDTTRa, LDMTLD, SYNNP, and WRDPRO). Five indices were significant predictors; LDTTRa was not significant. A significant regression equation was found with an R^2 of 0.52. The predicted scores of written claim, written evidence, and written reasoning based on Coh-Metrix individual indices are displayed in equation (7), (8), and (9), respectively.

$$\text{Written claim scores} = 6.32 - 2.35 \times \text{LDTTRa} - 0.002 \times \text{WRDPRO} - 0.004 \times \text{WRDPRP1s} - 0.89 \times \text{WRDFRQc} + 0.11 \times \text{RDFKGL} \tag{7}$$

$$\text{Written evidence scores} = - 0.38 + 0.23 \times \text{DESSC} + 0.04 \times \text{DESSL} - 0.001 \times \text{SMCAUSvp} + 0.90 \times \text{SMCAUSlsa} + 0.26 \times \text{SMTEMP} + 0.03 \times \text{SYNLE} - 0.001 \times \text{WRDPRO} + 0.002 \times \text{WRDIMGc} \tag{8}$$

$$\text{Written reasoning scores} = - 0.29 + 0.74 \times \text{DESSC} + 0.08 \times \text{DESSL} + 0.01 \times \text{LDMTLD} + 0.65 \times \text{SYNNP} - 0.003 \times \text{WRDPRO} \tag{9}$$

These findings indicated that the use of pronouns (WRDPRO) negatively predicted claim, evidence, and reasoning. Therefore high quality claim, evidence, and reasoning statements contained less pronouns. Pronoun use has been found to be highly correlated with conversational language [1, 15], so the minimal use of pronouns in high quality CER implies that students were using more formal language. The following is an example of a claim that received a full score for quality and did not contain any pronouns: "*when the shape was changed from narrow to wide and eventually square, the density of the liquid stayed the same.*" Here is an example of a claim that received 0 points and included the pronoun "it": "*It didn't support my answer.*" For this claim, the reader cannot determine what "it" means without additional context.

Other word information was also found to predict claim performance, but not performance on evidence or reasoning. Specifically, when students used less first-person singular pronouns (WRDPRP1s) and less frequently used words (WRDFRQc), their claim scores were higher. This phenomenon was not found in either evidence or reasoning. It is likely that students more often used structures in claims such as "*I found*", "*my claim/hypothesis*", and "*I changed*", compared to in their evidence and reasoning. Even though this pattern was not found in evidence, the use of content words that evoked mental images (WRDIMGc) increased the quality of evidence. The following two evidence statements demonstrate the value of including descriptive language: "*When using the oil in both containers (wide and narrow) the mass equaled 425 and the volume equaled 500, it gives you 0.85 for the density, there for the density remains the same*," and "*My experiment proves my evidence.*" The former had an extremely high WRDIMGc score relative to the latter, which received only 0 points. It is hard to obtain any useful information from the latter example, relative to the first example.

Scores of evidence and reasoning increased with both the number of sentences (DESSC) and the number of words (DESSL). However, these two features were not selected as predictors of claim performance because written claims are usually generated within one sentence. Thus, the number of sentences (DESSC) had a low correlation with written claim (r = 0.003) and DESSL was not selected because it was highly correlated with the Flesch-Kincaid Grade Level (RDFKGL) (r = 0.745), which was partially computed based on the mean number of words per sentence. Thus, RDFKGL was used to present DESSL and RDFKGL was a significant predictor of claim performance. For this reason, we could conclude that number of words was a significant predictor for claim, evidence, and reasoning.

Lexical density significantly predicted claim and reasoning, but not evidence. However, the method used to compute lexical density for claim was different from reasoning. Lexical density for claim was computed by the type-token ration

(LDTTRa), namely the number of unique words in writing (i.e., types) divided by the overall number of words (i.e., tokens) in writing. The measure of Textural Lexical Density (LDMTLD) used to predict reasoning was computed based on the mean length of sequential word strings in writing that maintained a given type-token ratio. LDTTRa was a Lexical density measure that was entered as a predictor for reasoning, but was not a significant predictor based on the step-wise regression procedure. LDTTRa was also used as a predictor for evidence, but it could not significantly predict evidence. One possible explanation is that its function may be counterbalanced by other variables, such as syntactic complexity and the situation model.

Additionally, lexical density largely decreased the scores of claim statements, but increased scores of reasoning statements. High lexical density indicates that there are more unique words, which involves the introduction of more new information. Low lexical density, on the other hand, implies repetition of words and redundancy. In claims with high scores, students tended to elaborate more, which likely led to the repetition of functional words, such as the article "the" in the following example: "*My hypothesis said that if I changed the container from narrow to wide that the density of the liquid would increase*" (LDTTRa = 0.59). On the contrary, in claims with low scores, students articulated general ideas, such as "*My hypothesis was wrong*," in which each word was unique and resulted in a LDTTRa score of 1.00 point.

Syntactic complexity was related to both evidence and reasoning, but different individual features were selected to predict evidence and reasoning, respectively. Specifically, for evidence, left embedded words before main verbs (SYNLE) was used as a predictor and was found to significantly increase evidence scores. For example, a high quality evidence statement involved left embeddedness with a large number of words before the main verb, such as "When using…" in the following example, "*When using the oil in both containers (wide and narrow) the mass equaled 425 and* ...". However, the low quality evidence statements did not contain a large number of words before main verb, such as in the example "*stay the same in both container.*"

SYNLE was not used as a predictor for reasoning because its correlation with written reasoning score was below 0.30. The mean number of modifiers per noun phrases (SYNNP) was used as a predictor and it significantly increased reasoning scores. For example, the following reasoning statement had a high score, *"This shows that the container shapes, wide and narrow, will allow the densities of the liquid to stay the same, which supports the claim"*, which contained some modifiers before and after the noun phrases, such as "the container" before "shapes," "wide and narrow" after "shapes," "the liquid" as a modifier of "densities," and the "which" clause was at the end of the statement. Low SYNNP caused reasoning scores to decrease. For example, in this reasoning statement, "*when u changed the container it got bigger*", no modifiers occurred before or after the noun "container," which meant the reader did not know what about the container was changed (i.e. shape, size, etc.). SYNNP was not used as a predictor for evidence because its correlation with evidence scores was below 0.30.

The indices used only to predict performance on evidence statements included causal verbs and causal particles (SMCAUSvp), LSA verb overlap (SMCAUSlsa), and temporal cohesion based on tense and aspect repetition (SMTEMP). These indices contribute to situation models that represent deep cohesion and clear causality. Here is an example that best illustrates a high quality evidence statement, "*I used a* *narrow container of oil getting a mass of 212.5 and volume of 250 getting .85 as my density. Then I used the wide container getting mass 212.5 and volume of 250 and getting .85 again as the density.*" In this example, the causality measure (SMCAUSvp) was zero, meaning no casual verbs or causal particles such as "impact" or "as a result" were used in this statement. However, this statement was replete with verbs that have clear links to actions, events, and states (e.g., "used" and "getting"), which is called LSA verb overlap. This example also showed high temporal cohesion (SMTEMP) as represented by the use of temporal particles such as "then," and consistency of tense (e.g., past tense of "used") and aspect (e.g., progressive "getting"). However, these three indices had low correlations (below .30) with written claim and reasoning statement performance. Therefore, they were not included in the claim or reasoning predictive models.

4 Conclusions and Implications

This study examined students' written scientific explanations from two perspectives: formality and actual language use. Results showed that students used academic language when they generated evidence and reasoning statements, but not claim statements. The analyses for the five major Coh-Metrix dimensions showed that students used more complex syntax and informational text when they generated high quality claim, evidence, and reasoning statements. Moreover, high quality evidence and reasoning statements tended to include more referential cohesion and deep cohesion. The quality of claim, however, decreased with deep cohesion. These findings imply that students need to be instructed on how to use deep cohesion when generating a high quality claim statement.

Analyses with individual Coh-Metrix indices showed that students need to be instructed to use causal verbs or causal connectives to explicitly specify relationships in their claims. Additionally, students need to be instructed not to use first person pronouns in their professional, academic writing. Similarly, they need to be instructed to avoid using vague pronouns to refer to a person or a thing when writing a claim, evidence, and reasoning statement. Although students demonstrated deep cohesion and use of descriptive words in their evidence statements, they require support in order to transfer this language use to their claim and reasoning statements. These various gaps in linguistic formality and use could be addressed through integrating automated assessments and scaffolding within intelligent tutoring systems. While the design of automated assessment and scaffolding of students' CER in terms of content is under way [16-17], researchers have yet to design automated scoring and feedback specifically to address language use.

The present study unpacks the academic language used when students generate a claim, evidence, and reasoning statement at both the macro-level and micro-level. The findings provide valuable information for teachers and researchers that can be used to enhance students' academic writing. A limitation of this study is that the data came from just one density activity. Future studies may include more activities within the same topic or across topics in order to investigate whether familiarity with a particular domain is related to academic language use in writing scientific explanations in the form of claim, evidence, and reasoning.

References

1. Biber, D.: Variation Across Speech and Writing. Cambridge University Press, New York (1988)
2. Common Core State Standards for English Language Arts (2010) http://www.corestandards.org/ELA-Literacy/
3. Crossley, S.A., Kyle, K., McNamara, D.S.: The tool for the automatic analysis of text cohesion (TAACO). Automatic assessment of local, global, and text cohesion. Behav. Res. **48**, 1227–1237 (2015). doi: 10.3758/s13428-015-0651-7
4. Galloway, E.P., Uccelli, P.: Modeling the relationship between lexico-grammatical and discourse organization skills in middle grade writers: insights into later productive language skills that support academic writing. Reading and Writing, **28**(6), 797-828 (2015). doi: 10.1007/s11145-015-9550-7
5. Gámez, P.B., Lesaux, N.K.: The relation between exposure to sophisticated and complex language and early-adolescent english-only and language minority learners' vocabulary. Child Development **83**(4), 1316-1331 (2012). doi: 10.1111/j.1467-8624.2012.01776.x
6. Gámez, P.B., Lesaux, N.K.: Early-adolescents' reading comprehension and the stability of the middle school classroom-language environment. Dev Psychol. **51**(4), 447-58 (2015). doi: 10.1037/a0038868
7. Gobert, J.D., Sao Pedro, M., Raziuddin, J., Baker, R.S: From log files to assessment metrics: measuring students' science inquiry skills using educational data mining. J. Learn. Sci. **22**, 521-563 (2013). doi:10.1080/10508406.2013.837391
8. Gobert, J.D., Baker, R.S., Sao Pedro, M.A.: Inquiry skills tutoring system. U.S. Patent No. 9,373,082. U.S. Patent and Trademark Office, Washington, DC (2016)
9. Gobert, J., Sao Pedro, M., Betts, C., Baker, R.S.: Inquiry skills tutoring system (alerting system). U.S. Patent No. 9,564,057. Washington, DC: U.S. Patent and Trademark Office (2016)
10. Gotwals, A.W., Songer, N.B.: Reasoning up and down a food chain: using an assessment framework to investigate students' middle knowledge. Sci. Educ. **94**(2), 259-281 (2010). doi: 10.1002/sce.20368
11. Graesser A.C., McNamara D.S., Kulikowich J.M.: Coh-Metrix: providing multilevel analyses of text characteristics. Ed. Res. **40**, 223-234 (2011). doi: 10.3102/0013189X11413260
12. Graesser, A.C., McNamara, D.S., Cai, Z., Conley, M., Li, H., Pennebaker, J.: Coh-Metrix measures text characteristics at multiple levels of language and discourse. The Elem School J. **115**, 210-229 (2014). doi: 10.1086/678293
13. Li, H., Cai, Z., Graesser, A.C.: Comparing two measures for formality. In: Proceedings of the Twenty Sixth International FLAIRS Conference, pp. 220-225 (2013)
14. Li, H., Cheng, C., Graesser, A.C.: A measure of text formality as a human construct. In: I. Russel, I. Eberle, B. (eds.) Proceedings of the Twenty-eighth International Florida Artificial Intelligence Research Society Conference, pp. 175–180. AAAI Press, Palo Alto, California (2015)
15. Li, H., Cai, Z., Graesser, A.C.: How good is popularity? Summary grading in crowdsourcing. In: 9th International Conference on Educational Data Mining. EDM Society, Raleigh, pp. 430-435 (2016)
16. Li, H., Gobert, J., Dickler, R.: Dusting off the messy middle: Assessing students' inquiry skills through doing and writing. In: André, E., Baker, R., Hu, X., Rodrigo, M., du Boulay, B. (eds.) Artificial Intelligence in Education. AIED 2017. Lecture Notes in Computer Science, vol 10331. Springer, Cham (2017). doi: 10.1007/978-3-319-61425-0_15

17. Li, H., Gobert, J., Dickler, R.: Automated assessment for scientific explanations in on-line science inquiry. In: Hu, X., Barnes, T., Hershkovitz, A., Paquette, L. (eds.) Proceedings of the 10th International Conference on Educational Data Mining. EDM Society, Wuhan, China, pp. 214-219 (2017)
18. McNamara, D.S., Graesser, A.C., McCarthy, P.M., Cai, Z.: Automated evaluation of text and discourse with Coh-Metrix. Cambridge University Press, New York (2014)
19. McNamara, D.S., Crossley, S.A., Roscoe, R.D., Allen, L.K., Dai, J.: A hierarchical classification approach to automated essay scoring. Assessing Writing **23**, 35–59 (2015). doi: 10.1016/j.asw.2014.09.002
20. McNeill, K., Lizotte, D.J., Krajcik, J., Marx, R.W.: Supporting students' construction of scientific explanations by fading scaffolds in instructional materials. J. Learn. Sci. **15**, 153-191 (2006). doi:10.1207/s15327809jls1502_1
21. McNeill, K.L.: Elementary students' views of explanation, argumentation, and evidence, and their abilities to construct arguments over the school year. J of Res in Sci Teaching **48**, 793-823 (2011). doi: 10.1002/tea.20430
22. National Center for Education Statistics: National Assessment of Educational Progress (NAEP) 2015 Reading Assessment [Data file] (2015). http://nces.ed.gov/nationsreportcard/ subject/publications/stt2015/pdf/2016008AZ4.pdf
23. National Governors Association Center for Best Practices & Council of Chief State School Officers: Common Core State Standards. Authors, Washington D.C. (2010)
24. National Research Council: A framework for K-12 science education: practices, crosscutting concepts, and core ideas. National Academies Press, Washington (2012)
25. Ruiz-Primo, M., Li, M., Shin-Ping, T., Schneider, J.: Testing one premise of scientific inquiry in science classrooms: examining students' scientific explanations and student learning. J of Res in Sci Teaching **47**, 583-608 (2010). doi: 10.1002/tea.20356
26. Snow, C.E., Uccelli, P.: The challenge of academic language. In: Olson, D.R., Torrance, N. (eds.) The Cambridge Handbook of Literacy vol 121, pp. 112-133. Cambridge University Press, Cambridge (2009). doi:10.1017/CBO9780511609664.008
27. Snow, C.E.: Academic language and the challenge of reading for leaning about science. Science 328 (2010). doi: 10.1126/science.1182597
28. Toulmin, S.: The Uses of Argument. Cambridge University Press, Cambridge (1958)
29. Wiley, J., Hastings, P., Blaum, D. et al.: Different approaches to assessing the quality of explanations following a multiple-document inquiry activity in science. Int J Artif Intell Educ **27**, 758 (2017). doi: https://doi.org/10.1007/s40593-017-0138-z
30. Yore L.D., Hand, B.M., Prain, V.: Scientists as writers. Sci Education **86**, 672-692 (2002). doi:10.1002/sce.10042.

Acknowledgements
This research is funded by the Department of Education (R305A120778). Any opinions expressed are those of the authors and do not necessarily reflect those of the funding agencies.