The Relationship between Scientific Explanations and the Proficiencies of Content, Inquiry, and Writing

Haiying Li
Rutgers University
New Brunswick, U.S.A.
Haiying.li@gse.rutgers.edu

Janice Gobert
Rutgers University
New Brunswick, U.S.A.
Janice.gobert@gse.rutgers.edu

Rachel Dickler
Rutgers University
New Brunswick, U.S.A.
Rachel.dickler@gse.rutgers.edu

ABSTRACT

Examining the interaction between content knowledge, inquiry proficiency, and writing proficiency is central to understanding the relative contribution of each proficiency on students' written communication about their science inquiry. Previous studies, however, have only analyzed one of these primary types of knowledge/proficiencies (i.e. content knowledge, inquiry proficiency, and writing proficiency) at a time. This study investigated the extent to which these proficiencies predicted students' written claims, evidence for their claims, and reasoning linking their claims to the evidence. Results showed that all three types of proficiencies significantly predicted students' claims, but only writing proficiency significantly predicted performance on evidence and reasoning statements. These findings indicate the challenges students face when constructing claim, evidence, and reasoning statements, and can inform scaffolding to support these challenges.

Author Keywords

Content knowledge; scalability; science inquiry; scientific explanation; writing proficiency.

INTRODUCTION

The science education community has been working to develop materials that accurately and authentically capture and assess the science inquiry practices outlined in the Next Generation Science Standards [20], including asking questions, planning and carrying out investigations, analyzing and interpreting data, warranting claims, constructing explanations, and communicating findings. Communicating findings involves the expression of scientific understandings [19, 20] and is a central practice used by experts in the field of science [28]. Assessments that capture the practice of communicating findings include traditional paper-based assessments [6, 18] that often contain open response items that require students to express their understandings in writing. One type of open response item that has been commonly used for capturing multiple science inquiry practice proficiencies, including communicating findings, is scientific explanations [6, 11, 12, 15, 18]. Scientific explanations involve the presentation of scientific concepts and evidence underlying a particular phenomenon [6, 11, 12, 18] and can be done as part of a hands-on experiment or in "virtual assessments" [15] such as Inq-ITS [5, 11, 12].

Open response scientific explanations can be assessed either holistically, based on the inclusion of central scientific concepts [15] or according to the inclusion of three structural components [11, 12, 18]. These structural components include: a claim, evidence that supports the claim, and reasoning for how the evidence supports the claim (CER) [11, 12, 18]. The CER format facilitates the identification of students' difficulties for each structural component so that more precise assessment and individualized scaffolding can be provided. Scientific explanations have also been assessed via forced response items (i.e., [1]), but these items are unable to capture students' writing proficiencies at communicating findings.

Prior studies have examined how and if individual proficiencies such as content knowledge [6], proficiency at conducting experiments [1, 12], and familiarity with task demands [17] were predictive of student performance on scientific explanations done in the context of inquiry. These individual proficiencies were found to not fully or accurately predict written scientific explanation performance (i.e. [6, 11, 12]). Writing proficiencies, however, have yet to be examined in relation to student performance on constructing scientific explanations. Additionally, prior studies have yet to determine the relative predictive power of each type of knowledge/proficiency (i.e. content knowledge, inquiry proficiency, and writing proficiency) on the claim, the evidence, and the reasoning statement (CER). In doing so, we would able to determine which proficiencies predict student performance on written C, E, and R. These data, in turn, could be used to detect student difficulties with CER writing so that appropriate scaffolding or instruction could be provided.

Thus, the present study aimed to examine whether content knowledge, inquiry proficiency, and writing proficiency predicted the quality of written scientific explanations in the form of claim, evidence, and reasoning, respectively. Content knowledge refers to students' domain-specific

L@S 2018, June 26–28, 2018, London, United Kingdom
© 2018 Association for Computing Machinery.
ACM ISBN 978-1-4503-5886-6/18/06…$15.00
https://doi.org/10.1145/3231644.3231660

knowledge. Inquiry proficiency refers to students' performance on inquiry practices, including: generating a hypothesis/question, carrying out investigations, analyzing and interpreting data, and warranting claims with evidence. Writing proficiency refers to students' proficiencies at constructing formal explanations using academic writing.

This study will advance research on scientific explanations in terms of communicating findings in the context of science inquiry for two primary reasons. First, this study provides empirical evidence on how students' different proficiencies contribute to the quality of their scientific writing. This is the first study in which multiple proficiencies are integrated and examined at a fine-grained, sub-component level. These fine-grained analyses enabled exploration of the contribution of each proficiency to the quality of written CER constructed during science inquiry. These findings will inform the automatic generation of instruction and scaffolds for scientific explanations, thereby allowing for scaling up intelligent systems such as Inq-ITS. Second, this study uses advanced technologies to automatically measure both inquiry proficiency and writing proficiency, again allowing for scalability of both automated assessment of, and in turn, automated scaffolding of, the full complement of NGSS practices; we are currently implementing these technologies for student writing in Inq-ITS [11, 12, 13].

This paper has four sections. First, we briefly review current studies on proficiencies related to the quality of scientific explanations constructed in science inquiry contexts. Second, we describe the materials, measures, and analyses of the present study in the Method section, as well as outline the scalability of these measures. Third, we display results and discuss the findings in terms of claim, evidence, and reasoning. Fourth, we present implications for teachers and researchers, as well as how the results of the present study contribute to scaling-up automated assessment and feedback.

CONTENT KNOWLEDGE

Domain specific content knowledge is relevant to science inquiry investigations, as investigations take place in the context of specific science topics. For instance, an inquiry investigation may involve the physical science topic of density, so students will need certain conceptual understandings (content knowledge) related to density in order to meaningfully engage in an inquiry investigation. Gotwals and Songer [6] investigated the relationship between students' content knowledge related to food webs in the domain of ecology and their performance on scientific explanations. Students' content knowledge and explanations were measured using a 20-item assessment that involved multiple choice and open response items. Content knowledge was scored in relation to item difficulty and explanations were scored according to the quality of students' claim, evidence, and reasoning statements. The study found that many students often had some level of understanding of certain concepts, but struggled with explaining those concepts. Therefore, there seemed to be a discrepancy in content understanding relative to student performance on constructing explanations. Li et al. have found similar data about the mismatch between students' inquiry proficiencies and their proficiencies at describing their inquiry (described in more detail in the next section; [12]).

INQUIRY PROFICIENCY

Studies have also investigated the relationship between students' experimental proficiencies and scientific explanations. Some assessments for science inquiry only examine students' scientific explanations of phenomenon without having students actually collect the data and engage in an inquiry investigation themselves (i.e. [6]). Studies that capture both students' experimental and explanatory writing proficiencies provide a valuable opportunity to examine the relationship between these proficiencies. For instance, Baker et al. [1] conducted a study where students engaged in a virtual investigation on genetic mutations and then constructed scientific explanations based on their investigation. Students' inquiry performance was scored dichotomously depending on whether or not components of students' investigations lead to accurate results. The scientific explanations consisted of a claim and evidence that supported the claim. The explanations were scored according to the level of accuracy. The researchers found that they could model student performance on causal explanation construction based on their performance in a frog mutation virtual investigation with a modest correlation of r = .53. Explanations in this study, however, were constructed through forced response items rather than in an open response format.

A study by Li et al. [12] examined the relationship between inquiry proficiencies and written scientific explanation performance in an intelligent tutoring system, Inq-ITS [5]. Inquiry practice performance was captured using machine-learned, automated scoring techniques (see [5] for details; [4]) and written explanations were scored according to components of CER [11]. The study found that while inquiry proficiency was a significant predictor of students' written explanations, inquiry proficiency could only explain 28% of the variance in students' explanations. These findings imply that students' CER writing is likely to be associated with other proficiencies, namely, content knowledge and writing proficiency.

WRITING PROFICIENCY

The NGSS emphasizes that students should be able to communicate their science inquiry findings through writing [18, 19]. Researchers have emphasized the importance of providing instruction on writing in science contexts [28] and use of academic writing style in science [24], but have yet to attend to writing proficiency in relation to the construction of the C, the E, and the R in the context of science inquiry.

Wiley et al. [27] investigated the writing quality of students' explanatory essays on the topic of global warming, but these essays differed from the types of explanations referred to in the present study based on length and context. For their measure of writing quality, Wiley et al. [27] used Coh-Metrix, a text analysis tool that can extract over 100 language and discourse features as indices of various aspects of students' writing proficiencies [16]. The study found that cohesion, causality, and lexical diversity correlated with the student essays. Lexical diversity (e.g., the number of different words used in students' writing out of the total number of words used; also called type-token ratio) was the only index, however, found to significantly predict written essay performance and only explained 8% of variance in writing scores. It is therefore possible that student performance on written scientific explanations though only partially predicted by writing proficiencies, may be fully predicted using additional inquiry practice proficiencies (e.g. forming hypotheses, collecting data, analyzing data, etc.) as well as content knowledge or other attributes of writing proficiency.

CLAIM, EVIDENCE, AND REASONING

The structure of scientific explanations has been identified as a factor that influences student performance on constructing explanations. Specifically, explanations can be elicited in a general open response format or specifically in a claim, evidence, reasoning (CER) format [11, 12, 18]. The CER format is based on a modified version of Toulmin's [26] framework for argumentation where students make a claim, provide evidence for their claim, and provide a justification for how their evidence supports their claim. McNeill [18] found that a lack of familiarity with the CER structure and specific components involved in that structure impacted students' performance on written explanations. Similarly, Li et al. [13] developed and used a fully operationalized analytic rubric to score subcomponents of students' CER. Li et al. [13] found that students had the greatest difficulties in writing about evidence as compared to claim and reasoning.

RESEARCH QUESTIONS

This study investigated the following research question: To what extent do students' content knowledge, inquiry proficiency, and writing proficiency (according to five dimensions of Coh-Metrix; see Writing Proficiency section for details) predict their performance on claim, on evidence, and on reasoning? As a follow-up to the main research question, the present study also examined the question of: Which specific variables representing content knowledge (i.e. general content knowledge score), inquiry (i.e. score on each science inquiry practice), and writing proficiency (i.e. performance on five dimensions of Coh-Metrix) are the most robust predictors of student performance on claim, on evidence, and on reasoning?

We hypothesize that students' content knowledge, inquiry, and writing proficiency can all predict writing performance to some extent, but predictive power may vary with claim, evidence, and reasoning. Specifically, content knowledge may have more predictive weight for claim because content knowledge may help students correctly interpret the relationship between a target independent variable (IV) and dependent variable (DV). Inquiry proficiency may have more predictive weight for evidence statements because successful evidence statements are possible (but not guaranteed) only if a student has collected appropriate data via successful inquiry. Successful inquiry, however, does not guarantee that a student can successfully describe their data in words. Writing proficiency may have more predictive weight for reasoning, as reasoning requires higher level writing proficiencies to generate a coherent, causal explanation for how the evidence supports the claim.

METHOD

Participants and Materials

254 middle school students (Grades 7 and 8 with a mean of 7.69, *SD* = 0.46) were from six different public schools located in Oregon and Massachusetts. Students completed one Inq-ITS density virtual lab, in which they investigated whether the shapes of a container (narrow, square, and wide) affected the density of a liquid through four stages of inquiry. Students completed the virtual lab on computers during their regular science class periods. All students had received prior instruction during the school year on the concept of density.

The density virtual lab is representative of other Inq-ITS virtual labs in terms of the structure of each inquiry stage and the shape of the container activity is representative of the other density lab activities. The first three stages of each Inq-ITS lab involve doing science via widgets, whereas the last stage involves only writing. During the Hypothesizing stage (hereafter called Hypothesis), students used a widget (dropdown menu; see Figure 1) to formulate a hypothesis to address an inquiry goal. The options available within the dropdown widget (displayed within parentheses) are as follows:

- If I change the (amount of the liquid, density of the liquid, shape of the container, type of the liquid) so that it (goes from narrow to wide, goes from wide to narrow, goes from narrow to square, goes from square to narrow, goes from wide to square, goes from square to wide), then I will observe that the (amount of the liquid, density of the liquid, shape of the container, type of the liquid) will (increase, decrease, stay the same).

Based on the research goal, students chose an independent variable (IV) that was to be manipulated, two conditions that they would manipulate, a dependent variable (DV), and the hypothesized effect of the IV on the DV.

In the Collecting Data phase (hereafter called Data Collection), students used a widget (clickable buttons) to manipulate the IVs in a simulation (see left image in Figure 2) while a data table automatically recorded their data (see right image in Figure 2).

During the Analyzing and Interpreting Data stage, students stated their claim based on the data that they collected through a widget in the same format as the Hypothesis widget (dropdown menu), and identified whether or not their claim supported their hypothesis (dropdown menu). Students also warranted their claims by selecting evidence from their data table (clickable buttons) (see left image in Figure 3).

The Explaining/Communicating Findings stage (hereafter called Scientific Explanation) was the final inquiry stage in which students responded to three open response questions in order to explain their claim (which corresponded with students' interpretation of data through widgets in the virtual lab), their evidence (which corresponded with the warranting of claims via the data selection widget), and describe their reasoning for how their evidence supported their claim (see right figure in Figure 3). The following are the open response prompts given for claim, evidence, and reasoning (also shown in the right image in Figure 3):

- Claim: Write a sentence that states what you found out about the scientific question you just investigated. Provide enough detail so that a friend who did not do the experiment could learn from your description.
- Evidence: Provide and describe scientific evidence from your data table that supports (or refutes) your claim. Remember to provide enough detail so that a friend who did not do the experiment could learn from your description.
- Reasoning: Explain why your evidence (what you wrote in Box 2) supports your claim (what you wrote in Box 1). Remember to provide enough details so that a friend who did not do the experiment could learn from your description.

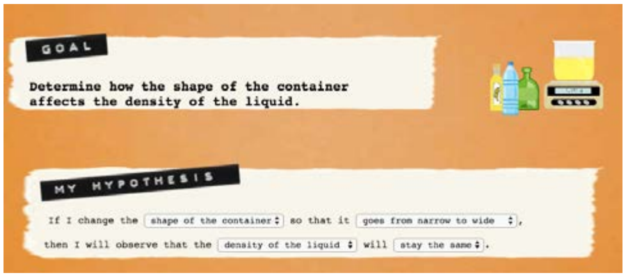
Measures

Content Knowledge

Students' content knowledge was measured before the inquiry investigation in the virtual lab using 10 multiple-choice questions. Researchers collaborated with a middle school science teacher consultant to construct these 10 questions. The questions were designed to address the NGSS strands related to density for middle school. The questions were also sent to middle school science teachers for feedback. The final 10 multiple-choice items covered content related to inquiry processes involved in calculating density (i.e. mass and volume), the scientific principle of how density is related to volume and mass, and the scientific principle of how density is a property of a substance. These questions were used to capture students' baseline density content knowledge that would then be applied when they used the Density virtual lab.

Inquiry Proficiency

Inquiry proficiency was measured according to four components using patented educational data mining techniques in Inq-ITS (see [5] for details; [4]) that evaluate whether students demonstrated the following four inquiry proficiencies in the Inq-ITS environment: hypothesis generation, data collection, data interpretation, and warranting a claim. Each practice was conceptualized and measured based on corresponding sub-components:



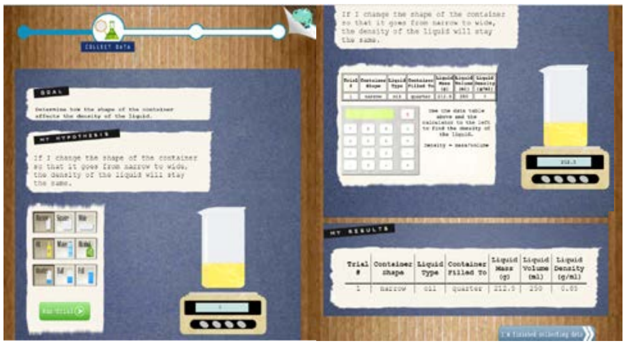
GOAL

Determine how the shape of the container affects the density of the liquid.

MY HYPOTHESIS

If I change the [shape of the container ⬍] so that it [goes from narrow to wide ⬍], then I will observe that the [density of the liquid ⬍] will [stay the same ⬍].

Figure 1. Hypothesizing stage.



GOAL

Determine how the shape of the container affects the density of the liquid.

MY HYPOTHESIS

If I change the shape of the container so that it goes from narrow to wide, the density of the liquid will stay the same.

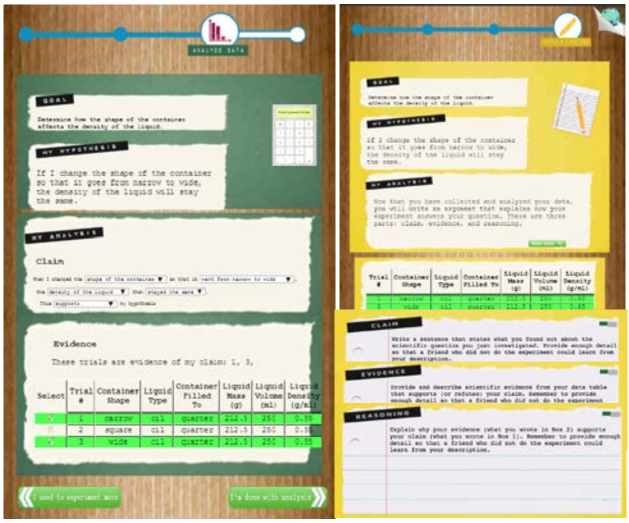
Use the data table above and the calculator to the left to find the density of the liquid.

Density = mass/volume

MY RESULTS

Trial #	Container Shape	Liquid Type	Container Filled To	Liquid Mass (g)	Liquid Volume (ml)	Liquid Density (g/ml)
1	narrow	oil	quarter	212.5	250	0.85

Figure 2. Collect data stage.



GOAL

Determine how the shape of the container affects the density of the liquid.

MY HYPOTHESIS

If I change the shape of the container so that it goes from narrow to wide, the density of the liquid will stay the same.

MY ANALYSIS

Claim

When I changed the [shape of the container ▼] so that it [went from narrow to wide ▼], the [density of the liquid ▼] then [stayed the same ▼].

This [supports ▼] my hypothesis

Evidence

These trials are evidence of my claim: 1, 3,

Select	Trial #	Container Shape	Liquid Type	Container Filled To	Liquid Mass (g)	Liquid Volume (ml)	Liquid Density (g/ml)
✓	1	narrow	oil	quarter	212.5	250	0.85
□	2	square	oil	quarter	212.5	250	0.85
✓	3	wide	oil	quarter	212.5	250	0.85

GOAL

Determine how the shape of the container affects the density of the liquid.

MY HYPOTHESIS

If I change the shape of the container so that it goes from narrow to wide, the density of the liquid will stay the same.

MY ANALYSIS

Now that you have collected and analyzed your data, you will write an argument that explains how your experiment answers your question. There are three parts: claim, evidence, and reasoning.

Trial #	Container Shape	Liquid Type	Container Filled To	Liquid Mass (g)	Liquid Volume (ml)	Liquid Density (g/ml)
1	narrow	oil	quarter	212.5	250	0.85
3	wide	oil	quarter	212.5	250	0.85

CLAIM

Write a sentence that states what you found out about the scientific question you just investigated. Provide enough detail so that a friend who did not do the experiment could learn from your description.

EVIDENCE

Provide and describe scientific evidence from your data table that supports (or refutes) your claim. Remember to provide enough detail so that a friend who did not do the experiment

REASONING

Explain why your evidence (what you wrote in Box 2) supports your claim (what you wrote in Box 1). Remember to provide enough detail so that a friend who did not do the experiment could learn from your description.

Figure 3. Analyze data stage and explaining findings stage.

- Hypothesis generation: (1) the identification of the IV and (2) the identification of the DV.
- Data collection: (1) designing a controlled experiment, (2) testing the hypothesis, and (3) running a pair of trials where the target IV was changed and other variables were controlled.
- Data interpretation: (1) examining the target IV, (2) examining the target DV, (3) interpreting the IV-DV relationship, and (4) identifying whether the findings supported the initial hypothesis.
- Warranting the claim: (1) warranting the IV-DV relationship with appropriate evidence, (2) selecting more than one trial to warrant the claim, (3) selecting trials that

support or refute the hypothesis, and (4) selecting all controlled trials.

Each sub-component was automatically scored according to whether students behaved correctly (1 point) or not (0 points). The score for each practice was calculated by generating the sum of sub-component scores underlying hypothesis generation, data collection, data interpretation, and warranting the claim, respectively.

Writing Proficiency

Coh-Metrix is one of the broadest and most sophisticated, textual assessment tools used for the automated evaluation of students' writing quality [16]. Different individual indices were identified in prior studies as sufficient to evaluate student writing in forms such as essays, self-explanations [16], and summaries [10]. Previous studies have used individual indices related to cohesion, lexical diversity, and causality (e.g., [27]) extracted by Coh-Metrix to predict the quality of explanatory essays. Even though some of these indices can individually predict the quality of writing to some extent, it is important to use multiple textual levels that represent overall writing proficiency.

Writing proficiency should reflect the mastery of language use at multiple-textual levels [8] as found in academic writing style [24]. Five major dimensions of Coh-Metrix have been used to represent language style ranging from informal (conversational) to formal (academic) according to: word use (Word Concreteness), syntax (Syntactic Simplicity), the explicit textbase (Referential Cohesion), the referential situation model (also called the mental model; Deep Cohesion), and the discourse genre and rhetorical structure (the type of discourse and its composition; Genre) [7, 10]. These measures have effectively differentiated writing in different genres (narrative versus scientific) and at different grade levels [7, 14]. This study adopts the following five Coh-Metrix dimensions to measure students' writing proficiency:

- Word concreteness: Concrete words can evoke mental images and are thus assumed to be more meaningful to the writer relative to abstract words. In scientific writing, students are expected to use more academic language as indicated by the use of more abstract words.
- Syntactic simplicity: Sentences are constructed with few words and simple, familiar syntactic structures. Complex sentences have structurally embedded syntax. For scientific explanations, better writing is expected to have greater syntactic complexity.
- Referential cohesion: High-cohesion writing contain words and ideas that overlap across sentences and the text as a whole, forming threads that connect the explicit textbase. For written scientific explanations, higher levels of referential cohesion indicate better writing.
- Deep cohesion: Causal, intentional, and other types of connectives are taken as evidence that writing reflects a more coherent and deeper understanding. For written scientific explanations, higher levels of deep cohesion indicate higher quality writing.
- Narrativity: Narrative texts tell stories that are familiar to the reader and are closely associated with everyday oral conversation. The opposite end of the spectrum is informational texts. In the context of written scientific explanations, students are expected to use higher levels of informational writing (i.e. lower narrativity).

Scientific Explanations

Students' written claim, evidence, and reasoning were manually graded according to a scoring rubric that was modified based on prior, fine-grained rubrics [11, 12]. In Inq-ITS' Analyzing and Interpreting Data stage of a virtual lab, students construct a claim using a widget with four components: IV (shape of the container), IVR (IV relationship: change from one shape to another among narrow, square, and wide), DV, and DVR (DV relationship, namely, changes or stays the same). The written claim was graded according to the same four components. The score for claim was the sum of these four sub-components, ranging from 0 to 4 points.

Written evidence was graded in terms of its sufficiency and its appropriateness [22]. Sufficiency is a measure of whether students provided enough evidence, i.e., whether students specified changing the shape of container from one shape to another (narrow to wide, wide to square, square to wide, etc.). Mentioning only one specific shape was considered insufficient evidence and not mentioning any shapes was considered incorrect. Appropriateness is a measure of whether students provided data related to understanding the IV and DV relationship. These data included the values of the mass of liquid, volume of the container, and density of the liquid. The score for evidence was the sum of these three sub-components, ranging from 0 to 4 points.

Reasoning was composed of three sublevel components: theory, connection of data to theory, and data. Theory referred to whether students understood that density was a property of the liquid substance or that it was represented by the ratio of mass to volume. Data referred to whether students drew accurate conclusions from the data, such as "The shape of the container does not affect the density of the liquid." The data-theory connection referred to whether students specified that their data supported or refuted their claim.

Two expert raters discussed the rubrics and then graded each sublevel component. The maximum score for claim and evidence was 4 points, respectively. The maximum score for reasoning was 6 points. Inter-rater reliability was assessed by the intraclass correlation coefficient with a two-way random model and absolute agreement type [23]. The interrater-reliabilities by Cronbach's alphas were .99, .99, .94 and the intraclass correlations were .99, .99, .88 for claim, evidence, and reasoning, respectively. Then two raters discussed the disagreements and generated agreement scores. The agreement scores were used to compute the total score for written scientific explanation performance in the form of claim, evidence, and reasoning.

ANALYSES AND FINDINGS

Analyses

The structure of the written scientific explanations was used to split the data into three subsets: claim, evidence, and reasoning. Three hierarchical regression analyses for these three subsets of data, respectively, were conducted to examine the two research questions. The dependent variable was explanatory writing scores for claim, evidence, and reasoning, respectively. The independent variables were: (1) content proficiency: total score of content knowledge as measured by the 10-item pretest; (2) inquiry proficiency: scores of the four inquiry components including hypothesis formation, data collection, data interpretation, and warranting a claim; and (3) writing proficiency measured by the five dimensions of Coh-Metrix including narrativity, syntactic simplicity, word concreteness, referential cohesion, and deep cohesion. A series of relevant assumptions were tested before conducting analyses. The criteria for interpreting magnitude of correlations was: small (r = 0.1), medium (r = 0.3), and large (r = 0.5) [3].

First, an examination of the correlations revealed that warranting a claim was highly correlated with data collection (r = .72) and data interpretation (r = .85), which exceeded the limit of the assumption that correlations between each pair of independent variables should be less than .70. Therefore, we removed variables of data collection and interpretation due to their high correlation with warranting (as measured via widget data). The sample size was deemed adequate, given eight independent variables (N = 254) [25]. The collinearity statistics were all within acceptable limits, as tolerance was greater than 0.10 and the variance inflation factor was below 10; thus, the assumption of multicollinearity was satisfied [2, 9, 21]. A value of Cook's distance less than 1 met the assumption of outliers. Residual and scatterplots indicated that the assumptions of normality, linearity, and homoscedasticity were all satisfied [8, 21].

We used Z-scores to standardize scores of all variables for the convenience of comparisons for all analyses. We conducted a 3-step hierarchical regression analysis for claim, evidence, and reasoning, respectively. For each analysis, content knowledge was entered at Step 1 to control for prior domain-specific knowledge (Model 1). Inquiry proficiency (Hypothesis and Warranting) was entered at Step 2 (Model 2), and writing proficiency was entered as Step 3 (Model 3). The order of these three types of proficiency followed the same order as encountered within the stages of inquiry in the virtual lab.

Findings

Table 1 displays the correlations of all variables. Table 2 displays the statistics related to the change in R^2 at each step in terms of claim, evidence, and reasoning. Table 3 shows the coefficients of each variable in the best model, which ended up occurring at Step 3 for claim, evidence, and reasoning, respectively.

Three Types of Proficiency and Scientific Explanations

To answer the primary research question, to what extent do students' content knowledge, inquiry proficiency, and writing proficiency predict their writing of claim, of evidence, and of reasoning, the changes in variance explained by the models (R^2) were compared across the three models. Specifically, we examined whether adding proficiency of content knowledge, inquiry proficiency, or writing proficiency step by step to the regression model would significantly improve the model. The proficiency that is shown to significantly improve the model would be considered significantly predictive of claim, evidence, and reasoning performance.

Var.	Write	Cont.	Hyp.	War.	Word	Syn.	Ref.	Deep
	Claim							
Cont.	0.30**							
Hyp.	0.35**	0.14*						
War.	0.42**	0.35**	0.49**					
Word	0.22**	0.11	0.08	0.14*				
Syn.	-0.39**	-0.18**	-0.15*	-0.27**	-0.55**			
Ref.	0.07	0.02	0.10	0.18**	0.07	-0.21**		
Deep	-0.06	-0.11	-0.04	-0.05	0.29**	-0.12	0.09	
Nar.	0.03	-0.07	0.06	0.14*	0.01	-0.02	0.59**	0.30**
	Evidence							
Cont.	-0.08							
Hyp.	-0.02	0.14*						
War.	-0.13*	0.35**	0.49**					
Word	0.09	0.06	0.01	-0.04				
Syn.	-0.22**	-0.05	-0.02	0.09	-0.40**			
Ref.	0.22**	-0.01	-0.12	0.01	-0.14*	-0.03		
Deep	0.10	-0.05	-0.02	-0.04	0.07	0.17**	0.20**	
Nar.	-0.13*	0.01	-0.01	0.02	-0.29**	0.23**	0.37**	0.35**
	Reasoning							
Cont.	0.08							
Hyp.	0.02	0.14*						
War.	-0.04	0.35**	0.49**					
Word	0.11	-0.05	-0.01	-0.05				
Syn.	-0.37**	0.04	-0.04	0.05	-0.32**			
Ref.	0.12	0.00	0.02	-0.05	0.03	0.14*		
Deep	0.02	0.08	0.08	0.02	0.16**	0.25**	0.06	
Nar.	-0.03	0.03	-0.01	-0.03	-0.04	0.04	0.52**	0.37**

Table 1. Pearson correlations between variables.

Note. *** p < .001. ** p < .01. * p < .05. † p < .10 (Same for tables below). Var. = Variables. Write = Writing Score, Cont. = Content Knowledge. Hyp. = Hypothesis. War. = Warranting a Claim. Word = Word Concreteness. Syn. = Syntactic Simplicity. Ref. = Referential Cohesion. Deep = Deep Cohesion. Nar. = Narrativity.

Claim

Results of the hierarchical regression analysis revealed that at Step 1, content knowledge significantly contributed to the regression model, accounting for 9% of the variance in written claim performance, $F(1,252) = 24.66$, $p < .001$, R^2 = 0.10. At Step 2, adding two variables that measured inquiry proficiency (hypothesis and warranting) explained an additional 14% of the variance in written claim performance, and this change was significant. Proficiency of content knowledge and inquiry together significantly explained 23% of the total variance in written claim performance, $F(3,250) = 25.11$, $p < .001$, $R^2 = 0.26$. At Step 3, adding the five Coh-Metrix variables that measured writing proficiency (Word Concreteness, Syntactic Simplicity, Referential Cohesion, Deep Cohesion, and Narrativity) explained an additional 8% of variance in written claim performance, and this change was also significant. The proficiency of content knowledge, inquiry, and writing together significantly explained 31% of the total variance in written claim performance, $F(5,243) = 14.34$, $p < .001$, $R^2 = 0.37$.

Proficiency	***R^2***	***R^2 Change***	***df***	***F Change***
Claim				
Content	0.09	0.09	1,252	24.66***
Inquiry	0.23	0.14	2,250	23.16***
Writing	0.31	0.08	5,245	5.68***
Evidence				
Content	0.01	0.01	1,252	1.67
Inquiry	0.02	0.01	2,250	1.78
Writing	0.17	0.14	5,245	8.50***
Reasoning				
Content	0.01	0.01	1,252	1.81
Inquiry	0.02	0.01	2,250	1.01
Writing	0.17	0.16	5,245	9.46***

Table 2. Unique contribution of three proficiencies to writing.

Var.	***B***	***SE B***	***β***	***t***	***R^2***	***F***
Claim					0.31	13.85***
(Constant)	0.00	0.05		0.00		
Content	0.15	0.06	0.15	2.54*		
Hypothesis	0.20	0.06	0.20	3.26***		
Warranting	0.18	0.07	0.18	2.77**		
Word Concreteness	0.04	0.07	0.04	0.55		
Syntactic Simplicity	-0.29	0.07	-0.29	-4.27***		
Referential Cohesion	-0.08	0.07	-0.08	-1.14		
Deep Cohesion	-0.09	0.06	-0.09	-1.45		
Narrativity	0.07	0.07	0.07	0.93		
Evidence					0.17	6.07***
(Constant)	0.00	0.06		0.00		
Content	-0.04	0.06	-0.04	-0.66		
Hypothesis	0.07	0.07	0.07	1.10		
Warranting	-0.13	0.07	-0.13	-1.81†		
Word Concreteness	-0.04	0.07	-0.04	-0.55		
Syntactic Simplicity	-0.18	0.07	-0.18	-2.74**		
Referential Cohesion	0.28	0.06	0.28	4.32***		
Deep Cohesion	0.16	0.06	0.16	2.44*		
Narrativity	-0.25	0.07	-0.25	-3.62***		
Reasoning					0.17	6.47***
(Constant)	0.00	0.06		0.00		
Content	0.11	0.06	0.11	1.79†		
Hypothesis	0.01	0.07	0.01	0.13		
Warranting	-0.06	0.07	-0.06	-0.90		
Word Concreteness	-0.06	0.06	-0.06	-0.86		
Syntactic Simplicity	-0.40	0.07	-0.40	-6.09***		
Referential Cohesion	0.13	0.07	0.13	1.82†		
Deep Cohesion	0.16	0.07	0.16	2.37*		
Narrativity	-0.14	0.07	-0.14	-1.93†		

Table 3. Coefficients in the full model (*df* (8, 245)).

Note. $^{***}p < .001$. $^{**}p < .01$. $^{*}p < .05$. $^{\dagger}p < .10$

These findings indicated that each type of proficiency was related to the quality of the written claim, with inquiry proficiency having more predictability than content knowledge and writing proficiency. The predictive power of content knowledge and writing was almost equivalent. The larger predictive weight of inquiry proficiency implies that students rely more on information that they obtained through inquiry when constructing a claim. This is consistent with our assumption that successful inquiry facilitates the generation of a better, more accurate claim. Moreover, mastery of content knowledge fosters students in constructing a better claim as well. Perhaps, content knowledge enables students to confirm that what they write in their claim is correct. Furthermore, students' higher level of writing proficiency increases the quality of their written claims.

Evidence

The second hierarchical regression analysis was conducted for evidence and the results showed a different pattern relative to the pattern for the written claim results. At Step 1, content knowledge did not significantly predict written evidence. At Step 2, the two variables added that measured inquiry proficiency did not significantly predict written evidence on its own or along with content knowledge. At Step 3, the addition of the five Coh-Metrix variables that measured writing proficiency significantly explained an additional 14% of variance in students' written evidence. The predictors of content knowledge, inquiry, and writing proficiency together significantly explained 17% of the total variance in evidence writing, $F(8,245) = 6.07$, $p < .001$, $R^2 = 0.17$.

These findings indicated that content knowledge and inquiry proficiency did not contribute to students' writing of evidence. Only writing proficiency was substantially predictive of evidence writing. These findings imply that students' written evidence statements depend primarily on their writing proficiency, but not content knowledge or inquiry proficiency. It makes sense that content knowledge does not significantly contribute to written evidence performance because evidence requires students to report the data that they collected during inquiry. The data are not based on students' content knowledge, but on the results of their inquiry investigation. However, it is surprising that inquiry proficiency is not related to evidence writing because students are specifically expected to write about the data collected during their inquiry investigation. Further examination of students' data collected and selected during inquiry showed that 207 students (81%) selected sufficient and appropriate data to warrant the claim during inquiry, but only six of those students fully reported these data in their written evidence. 131 out of 207 students (63%) did not report any data. For example, some students described their work, saying "I found this by doing the work" or provided slightly more but vague information "I changed the variable…" These 207 students should know which data support the claim, but most of the students did not report these data in their evidence writing. Other students were not clear on the difference between claim and evidence. For instance, some students wrote "I already did this part." These findings demonstrate that students have difficulty demonstrating their inquiry proficiencies in writing. This is commensurate with previous studies [12, 16] who found that students needed support regarding what information to include in their evidence statements.

Reasoning

A third hierarchical regression analysis was conducted for reasoning. Results were similar to those for evidence statements. Only at Step 3 with the addition of five Coh-Metrix variables that measured writing proficiency did we yield statistically significant results, whereby an additional 16% of variance was explained in students' reasoning statements. Proficiency of content knowledge, inquiry, and writing together significantly explained 17% of the total variance in reasoning writing, $F(8,245) = 6.47$, $p < .001$, $R^2 = 0.17$.

These findings indicate that only writing proficiency was predictive of reasoning performance, as was the case for evidence performance. It is no surprise that the findings in evidence statements were consistent with those for reasoning statements. This is because half of the scores for reasoning statements were in some way related to data that would be reported in the evidence statements (i.e. explaining how data of the DV was impacted by changing the IV). However, it is a surprise that content knowledge was not a significant predictor because the inclusion of a scientific theory in one's statement accounted for two points out of the total reasoning score. Moreover, if proficiency of content knowledge and inquiry were predictive of claim performance, they would be expected to be predictive of reasoning performance as well since part of students' reasoning involved reiterating the claim. For instance, components of the claim (IV, DV, the IV-DV relationship) also accounted for three points out of the total reasoning score. However, we found that neither content knowledge nor inquiry could significantly explain reasoning writing performance. The same 207 students who collected and selected appropriate data during inquiry also demonstrated high performance on warranting their claims with evidence using the widget, but only 18 of those students (9%) pointed out how and what data supported their claim in their writing. The remaining students failed to articulate this relationship using data. The most frequent responses for reasoning included: partially relevant responses, such as "my hypothesis worked/is wrong"; incomplete responses, such as "See above", "Look at it and read just like I can I'm older than 5"; some off-topic responses, such as "I love you so much fun…"; some metacognitive responses, such as "I don't know"; or some gibberish writing where they repeated the same letters. These findings further demonstrate that students may not be fully aware of what information they should provide when they generate reasoning or of the differences between claim, evidence, and reasoning.

Robust Features

To answer the follow-up research question of, "Which specific variables representing content knowledge (i.e. general content knowledge score), inquiry (i.e. score on each science inquiry practice), and writing proficiency (i.e. Five-Coh-Metrix indicators) are the most robust predictors for the writing of claim, of evidence, and of reasoning?" we identified the most predictive variables from the best performing hierarchical regression model for claim, for evidence, and for reasoning.

Predictors for claim

Coefficients for four out of the eight variables in the final Step 3 model for claim writing were significant predictors of performance on the claim statement. Syntactic Simplicity (i.e. use of few, familiar words and/or writing sentences with basic structures) was the most robust predictor, followed by generating a hypothesis, warranting a claim, and content knowledge. Specifically, the quality of written claim significantly improved if students had higher content knowledge, were better at generating a hypothesis and warranting a claim, and used more complex syntactic structures. The overall quality of written claim improves with an increase in hypothesis performance by 0.20 units, warranting a claim by 0.18 units, and content knowledge by 0.15 units; and decrease in syntactic simplicity by 0.29 units.

These findings imply that when students construct their claim, they may be relying on inquiry proficiencies. Specifically, high inquiry proficiency may facilitate the generation of an investigable hypothesis, collection of useful data, accurate analysis and interpretation of data, and warranting the claim with sufficient and appropriate data.

Our findings showed that students also potentially wrote the claim with some dependence on content knowledge, since content knowledge was the second most robust predictor of student performance on claim. Specifically, results of the model indicated that the more content knowledge students had, the better they performed on their written claim.

Performance on the claim was also related to students' writing proficiency, but this proficiency was restricted to the sentence level. This is because a good claim can be explicitly stated in one sentence. Thus, it is unnecessary for students to use cohesive devices to bridge inference gaps for readers. Since writing a claim involves stating a scientific fact, students should write their claims using an informational style. The results of the model, however, indicated that narrativity was not a robust predictor. Moreover, students would be expected to use more abstract words to generate a high quality claim, but word concreteness was not a robust predictor. These findings imply that students did not use expository style writing with abstract words when writing their claims.

Predictors for Evidence

The final Step 3 model was the best model for written evidence, and showed that only the four variables for writing proficiency robustly predicted evidence writing performance. These four variables from most to least predictive include: Referential Cohesion, Narrativity, Syntactic Simplicity, and Deep Cohesion. Specifically, the quality of written evidence significantly improved when students repeated content words instead of using pronouns to replace content words, used more causal connectives, used less narrative language, and used less simple sentence structures. The overall quality of written evidence improved, with the increase in Referential Cohesion by 0.28 units and Deep Cohesion by 0.16 units; and with the decrease in Narrativity by 0.25 units, and Syntactic Simplicity by 0.18 units.

These findings imply that when students construct evidence statements, they do not refer to the data obtained during inquiry. This may explain why students' evidence scores were very low with an average of 1.28 points out of the total score of 4 points (*SD* = 1.22).

Findings indicated that students' evidence statements reflected a proficient level of writing, as compared to their claim statements. Besides using more complicated syntactic structures, students used more referential cohesive devices, such as repeatedly using content words to bridge the gaps that require readers' inference-making. They also used more causal connectives to explicitly state causal relationship(s) so that readers could easily understand the writing. Moreover, students used a more expository style of writing instead of conversational writing. Only one dimension, Word Concreteness was not a significant predictor. Thus, the four significant dimensions that represent writing proficiency indicate that if students use more complex sentences, cohesion, and less narrativity in their evidence writing, the quality of evidence writing improves. There is still the question, however, of why students demonstrated higher writing proficiency on their evidence compared to their claim statements? One possible reason is that a claim does not involve as much repeated information, whereas evidence writing requires students to provide at least two pieces of evidence. Thus, students used more words and sentences in evidence writing relative to claim writing: *M* = 17.90 (*SD* = 7.48) and *M* = 22.33 (*SD* = 13.82) average number of words for claim and evidence, respectively; *M* = 1.07 (*SD* = 0.27) and *M* = 1.36 (*SD* = 0.71) average number of sentences for claim and evidence, respectively.

Predictors for Reasoning

The final Step 3 model for reasoning was the best model and showed that two variables for writing proficiency robustly predicted the quality of reasoning writing. Syntactic Simplicity was the most predictive of reasoning scores, followed by Deep Cohesion. Specifically, the quality of written reasoning significantly improved if students used more complicated syntactic structures and more causal connectives. The overall quality of written reasoning improved with the increase in Deep Cohesion by 0.16 units, and with the decrease in Syntactic Simplicity by 0.40 units.

These findings were similar to evidence writing. Students did not utilize information that they acquired during inquiry to elaborate on the relationship between the claim and evidence; nor did make a theoretical connection based on their prior content knowledge. Similar to evidence, students used more complex sentences and causal connectives in reasoning. Different from evidence statements, the variables of Narrativity and Referential Cohesion were only marginally significant predictors for reasoning. One reason is that students were not required to present two pieces of evidence in reasoning, so there was less repetition of phrases in terms of referential cohesion. Students could also use slightly more informal language in their reasoning to explain the relationship between the claim and evidence, so narrativity was not a necessary indicator of reasoning quality.

CONCLUSIONS AND IMPLICATIONS

In this study, we explored the extent to which content knowledge, inquiry proficiency, and writing proficiency predicted the quality of scientific explanations during inquiry according to the three components of: claim, evidence, and reasoning. The findings indicated that the combination of content knowledge, inquiry proficiency, and writing proficiency could significantly predict only students' claim performance. Specifically, the robust predictors for claim consisted of content knowledge, Hypothesis, Warranting, and Syntactic Simplicity. The findings further indicated that writing proficiency alone could predict students' evidence and reasoning performance. Robust predictors for evidence included the four Coh-Metrix dimensions, Narrativity, Syntactic Simplicity, Referential Cohesion, and Deep Cohesion. Robust predictors for reasoning were Syntactic Simplicity and Deep Cohesion. These findings suggest that each structural component of the scientific explanation (i.e. claim, evidence, and reasoning) involved different proficiencies as well as challenges.

Findings suggest that students used three types of proficiency (i.e., content knowledge, inquiry, and writing) for generating their claim statements, but writing proficiency alone could predict performance on evidence and reasoning statements. These findings may be primarily a result of students' difficulties with constructing written evidence and reasoning statements. In the present study, the scores of evidence and reasoning were very low, with average scores below 50% for evidence (*M* = 1.28, *SD* = 1.21; Total Points Possible = 4) and reasoning (*M* = 2.41, *SD* = 1.43; Total Points Possible = 6). If students do not understand how to construct these statements or are unfamiliar with the content that is supposed to be included in these statements, then the quality of their writing would not be improved regardless of their prior content knowledge or inquiry proficiency.

These findings further indicate the need to teach students the information that should be included in each component of: claim, evidence, and reasoning. This work also sets the stage for scaling up the assessment of written CER in terms of using automated scoring and text analysis tools to capture the quality of student writing in online inquiry environments. Automated scoring that takes into account students' content, inquiry, and writing proficiencies is an important step towards assessing and supporting the full complement of inquiry practices at scale.

ACKNOWLEDGEMENTS

This research was supported by the IES (R305A120778) and NSF (1629045) to Janice Gobert, PI at Rutgers University.

REFERENCES

1. R. S. Baker, J. Clarke-Midura, J. Ocumpaugh. 2016. Towards general models of effective science inquiry in virtual performance assessments. *J Comp Assist Learn* 32: 267-280.
2. S. J. Coakes, L. Steed. 2009. *SPSS: Analysis without anguish using SPSS version 14.0 for windows.* Wiley.
3. J. Cohen. 1992. A power primer. *Psychological Bulletin* 112: 155-159.
4. J. D. Gobert, R. S. Baker, M. A. Sao Pedro. 2014. Inquiry skills tutoring system, U.S. Patent 9,373,082, Filed February 1, 2013, issued January 29, 2014.
5. J.D. Gobert, M. Sao Pedro, J. Raziuddin, R. S. Baker. 2013. From log files to assessment metrics: measuring students' science inquiry skills using educational data mining. *J Learn Sci* 22: 521–563.
6. A. W. Gotwals, N. B. Songer. 2010. Reasoning up and down a food chain: using an assessment framework to investigate students' middle knowledge. *Sci Educ* 94: 259-281. DOI: 10.1002/sce.20368.
7. A. C. Graesser, D. S. McNamara, Z. Cai, M. Conley, H. Li, J. Pennebaker. 2014. Coh-Metrix measures text characteristics at multiple levels of language and discourse. *Elem School J* 115: 210–229. DOI: 10.1086/678293
8. A. C. Graesser, D. S. McNamara, J. Kulikowich. 2011. Coh-Metrix: providing multilevel analyses of text characteristics. *Educ Research* 40: 223–234.
9. J. F. Hair Jr., R. E. Anderson, R. C. Tatham, W. C. Black. 1998. *Multivariate data analysis.* Prentice-Hall.
10. H. Li, Z, Cai, A. C. Graesser. 2016. How good is popularity? Summary grading in crowdsourcing. In *Proceedings of the 9th International Conference on Educational Data Mining,* 430-435.
11. H. Li, J. Gobert, R. Dickler. 2017. Automated assessment for scientific explanations in on-line science inquiry. In *Proceedings of the 10th International Conference on Educational Data Mining,* 214-219.
12. H. Li, J. Gobert, R. Dickler. 2017. Dusting off the messy middle: assessing students' inquiry skills through doing and writing. In *Artificial Intelligence in Education* (AIED '17), 175-187.
13. H. Li, J. Gobert, R. Dickler. 2017. Analyzing the sub-skills underlying students' scientific claims, evidence, and reasoning during inquiry. Paper presented at the *Twenty-seventh Annual Meeting of the Society for Text & Discourse*.
14. H. Li, A. C. Graesser, Z. Cai. 2013. Comparing two measures of formality. In *Proceedings of the Twenty-sixth International Florida Artificial Intelligence Research Society Conference,* 220–225.
15. O. L. Liu, J. A. Rios, M. Heilman, L. Gerard, M. C. Linn. 2016. Validation of automated scoring of science assessments. *J Res Sci Teach* 53: 215-233.
16. D. S. McNamara, A. C. Graesser, P. M.McCarthy, P. M., Z. Cai, Z. 2014. *Automated evaluation of text and discourse with Coh-Metrix*. Cambridge University Press.
17. K. L. McNeill. 2011. Elementary students' views of explanation, argumentation, and evidence, and their abilities to construct arguments over the school year. *J Res Sci Teach* 48, 7: 793-823.
18. K. McNeill, D. J. Lizotte, J. Krajcik, R. W. Marx. 2006. Supporting students' construction of scientific explanations by fading scaffolds in instructional materials. *J Learn Sci* 15: 153-191.
19. National Research Council. 2012. *A framework for K-12 science education: practices, crosscutting concepts, and core ideas.* National Academies Press.
20. Next Generation Science Standards Lead States. 2013. *Next generation science standards: for states, by states*. National Academies Press.
21. J. Pallant. 2013. *SPSS survival manual.* McGraw-Hill Education.
22. W. A.Sandoval, K. A. Millwood. 2005. The quality of students' use of evidence in written scientific explanations. *Cognition and instruction* 23, 1: 23-55.
23. P. E. Shrout, J. L. Fleiss. 1979. Intraclass correlations: uses in assessing rater reliability. *Psychol Bull* 86: 420–428. DOI: 10.1037/0033-2909.86.2.420
24. C. E. Snow, P. Uccelli, P. 2009. The challenge of academic language. In *The Cambridge handbook of literacy,* D. R. Olson and N. Torrance (eds.). Cambridge University Press, 112−133.
25. B. G. Tabachnick, L. S. Fidell. 1996. *Using multivariate statistics* (3rd. ed.). HarperCollins.
26. S. Toulmin. 1958. *The Uses of Argument.* Cambridge University Press.
27. J. Wiley, P. Hastings, D. Blaum, A. J. Jaeger, S. Hughes, P. Wallace, T. D. Griffin, M. A. Britt. 2017. Different approaches to assessing the quality of explanations following a multiple-document inquiry activity in science. *International Journal of Artificial Intelligence in Education*, 1-33.

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.