Different approaches to assessing the quality of explanations following a multiple-document inquiry activity in science

Jennifer Wiley[1], Peter Hastings[2], Dylan Blaum[3], Allison J. Jaeger[1], Simon Hughes[2], Patricia Wallace[3], Thomas D. Griffin[1], & M. Anne Britt[3]

[1] *University of Illinois at Chicago, Chicago, Illinois*

[2] *DePaul University, Chicago, Illinois*

[3] *Northern Illinois University, DeKalb, Illinois*

Abstract. This article describes several approaches to assessing student understanding using written explanations that students generate as part of a multiple-document inquiry activity on a scientific topic (global warming). The current work attempts to capture the causal structure of student explanations as a way to detect the quality of the students' mental models and understanding of the topic by combining approaches from Cognitive Science and Artificial Intelligence, and applying them to Education. First, several attributes of the explanations are explored by hand-coding and leveraging existing technologies (LSA and Coh-Metrix). Then, we describe an approach for inferring the quality of the explanations using a novel, two-phase machine learning approach for detecting causal relations and the causal chains that are present within student essays. The results demonstrate the benefits of using a machine learning approach for detecting content, but also highlight the promise of hybrid methods that combine ML, LSA and Coh-Metrix approaches for detecting student understanding. Opportunities to use automated approaches as part of Intelligent Tutoring Systems that provide feedback toward improving student explanations and understanding are discussed.

Keywords. Automatic Assessment, Mental Models, Explanations, Causal Structure, Causal Relations, Machine Learning, Natural Language Processing

INTRODUCTION

As part of instruction in many subject-matter areas, students are often asked to demonstrate their understanding by responding to open-ended questions. In science, students may be asked to learn about the causes of phenomena such as volcanic eruptions, ice ages, el Niño, skin cancer, coral bleaching, or global warming, so that they might construct mental models of how or why these things happen. From a Socratic perspective, one ideal educational context for this learning to take place in would be with a 1:1 teacher-to-student ratio, where each student could articulate their

Portions of this work were supported by the Institute of Education Sciences (R305B070460, R305F100007) and the National Science Foundation (1535299). Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of these organizations. The authors thank Karyn Higgs, Kristopher Kopp, Mike Mensink, Tegan Michl, Carlos Salas, Brent Steffens, and Andrew Taylor for their contributions on this project.

understanding to an instructor in a face-to-face setting, and the instructor could give the students feedback on their mental models, help them to repair or remediate their misconceptions, and prompt them to be more coherent, complete, or focused in their responses. Yet, the realities of instruction are far from this ideal. Our public educational system does not have the resources to provide 1:1 human tutoring for all students in all subjects all of the time. As a much more feasible alternative, student understanding is often assessed by closed-ended tests that only require recognition or verification of ideas on the part of the student, and can be easily scored. Another alternative for assessment is asking students to demonstrate their understanding in writing, by composing responses to open-ended questions, including explanations of how or why things happen. Because of the importance of developing student skills in written communication and explanation, and because prompting students to articulate an explanation can provide a sensitive measure of student understanding, explanation essays are a valuable way of assessing student learning. However, these explanations still need to be evaluated for their quality. Developing automated coding systems that can recognize the quality of student understanding in written responses following reading assignments is one possible way to close the teacher-to-student-ratio gap. New technologies offer the promise of better individualized assessment, which may allow for tailored feedback and support during the learning process, which in turn may ultimately support better student performance and understanding.

A substantial body of work has explored hand coding and automated coding for the quality of student writing in response to composition prompts (Crossley, Kyle, & McNamara, 2015; Crossley, & McNamara, 2010, 2011; McNamara, Crossley, Roscoe, Allen, & Dai, 2015). In this research, students are asked to write expositions on themes or persuasive essays on a topic. The main goal of this research has been finding reliable predictors for the quality of writing as assessed by independent expert raters (Hout, 1996). Students are not given texts to read or specific content to learn, but rather are asked to expound upon a topic based on their prior knowledge and opinion. This closely mimics what students experience in the classroom as well as outside of the classroom as part of placement and exit assessments for writing skills. This type of skills assessment is quite a different enterprise than using students' written responses to evaluate the quality of their understanding of a topic from a learning activity or from a particular set of readings. Findings from prior studies that have been concerned with predicting perceived writing quality in student compositions may or may not be relevant for predicting student understanding from written responses. At present, there is much less work that has explored the features of student writing that are predictive of their understanding of a topic. Correspondingly, there has been a recent push to consider the disciplinary context and the goals of the written product as part of the assessment process (Ferris, 2007; Hout, 1996; Sommers, 2008).

In the present work, we describe the results from a variety of approaches that were used to evaluate the quality of explanations that were written as part of a multiple-document inquiry unit on global warming, where students were tasked with understanding how and why recent patterns of average global temperature differ from those seen in the past. In overview, the main goal was to develop and compare various approaches to assessing the quality of the mental models that students had constructed from the reading activity, by coding responses to an open-ended explanation essay prompt, and using test scores on a closed-ended comprehension test as the criterion measure of their understanding of the material. The writing activity did not involve a

general assessment of writing quality. The coding attempts ranged from identification of specific sets of attributes present in the explanations (e.g., concepts from a causal model), to more global or holistic evaluations of explanation quality (e.g., causal language), and from hand-coded scores to attempts to automate the scoring process using both existing technologies (LSA and Coh-Metrix) as well as a tailored machine learning approach specific to this inquiry task. The main question of interest is which approaches to coding the explanations would best capture the quality of each student's understanding of the subject matter.

We first present results of prior studies that have used various methods of hand-scoring of explanation essays to provide background for the coding and analyses that are employed in the current study. Next, we review prior work using existing out-of-the-box technologies (LSA and Coh-Metrix) to outline how those systems may be used for automatic detection of the hand-coded features. Before presenting the results of both hand coding and LSA/Coh-Metrix measures in terms of simple correlations between features of the essays and student understanding as measured by the criterial test of understanding (the closed-ended comprehension test), we provide relevant details about the sample and methods. This sets the stage for the main analyses in which regressions are used to examine which coded features of the written responses best predict student understanding. Finally, we describe a machine-learning approach that was developed to capture information about the arguments that students wrote for this specific inquiry activity and document set, and the extent to which it and other automated approaches can be used in combination to best predict explanation quality and student understanding.

BACKGROUND

Hand Scoring of Explanation Quality and Student Understanding

Prior work done specifically on learning from multiple-document reading and writing activities has examined the extent to which students transform the original text when they are asked to write a response to an inquiry question and whether they attempt to develop integrated causal models as part of understanding the readings (Britt & Aglinskas, 2002; Voss & Wiley, 1997; 2000; Wiley, 2001; Wiley & Voss, 1996; 1999). Several general aspects of students' responses have been considered in prior work: the organization or structure of their answers, the selection of the information that is included in the answers, and the integration or transformation of that information. Specific analyses of students' written responses have included the following features: a) the length of the response, (b) references to sources, (c) the organization or macrostructure of the response in relation to the prompt (i.e., listing of ideas versus analytical essay, use of evidence to support a claim); (d) the completeness of the account (i.e., the extent to which idea units mentioned in the document set are included in students' essays, or key concepts from the causal model); e) the integration and transformation of information within the account (i.e., number of causal connections or connectives present in the essays; proportion of sentences taken directly or paraphrased from sources, versus transformed or completely novel information). These features are the attributes that are focused on in the present work.

One frequently analyzed aspect of written responses is their length. Essay length is generally operationalized as the number of words or number of sentences, and may positively predict essay quality as this feature can signify more complete understanding (Page, 1994).

However, length is not always an indicator of better understanding, especially when students are asked to summarize rather than just recall or report what they have read (Wade-Stein & Kintsch, 2004). We would expect that students who write very short explanations will be unlikely to provide coverage of the causal model in their essay, but it is unclear whether student understanding will always positively correlate with essay length.

Other work has been concerned with whether students explicitly cite sources in the essays (Britt & Aglinskas, 2002; Rouet, Britt, Mason, & Perfetti, 1996; Rouet, Favart, Britt, & Perfetti, 1997), include information from many documents (Britt, Wiemer-Hastings, Larson, & Perfetti, 2004), or use information from multiple texts to support their claims on a controversy (Rouet et al., 1996). The presence of citations when writing from multiple documents in history is usually related to better quality essays (e.g., Britt & Aglinskas, 2002). We included this feature in the hand-coding to examine the extent to which citations would predict understanding from this science unit.

Another aspect of written essays that has been explored in prior work is the organization or top-level structure (c.f. Wiley & Voss, 1999). Using Meyer's (Meyer, 1985) taxonomy, essays can be classified as either having a collective structure (that is, the essay consists of a listing of ideas with minimal focus) or a more analytic or causal structure (having a main claim, thesis, or conclusion, with information organized in relation to that main claim). Studies have found that students who demonstrate better understanding of the material on comprehension tests write essays that are more likely to have an analytic or causal macrostructure (Voss & Wiley, 1997; 2000; Wiley, 2001; Wiley & Voss, 1999). Because in the present study students were explicitly prompted to write an explanation about how and why recent patterns of temperature differ from the past, examining the macrostructure provides a measure of whether students attempted to write an essay that directly answered the question.

To code for coverage, researchers may engage in a discourse analysis of the original reading material to identify a finite set of idea units that are present (Perfetti, Britt, & Georgi, 1995; Rouet et al., 1996; Wiley & Voss, 1999). Alternatively, researchers may identify a set of core causal concepts or a subset of idea units that are most important or critical for developing an appropriate mental model of the phenomenon (Griffin, Wiley, Britt, & Salas, 2012; Jaeger & Wiley, 2015; Sanchez & Wiley, 2006, 2009, 2010, 2014; Wiley, Ash, Sanchez, & Jaeger, 2011; Wiley et al., 2009). Sometimes the concepts from the *a priori* causal model are further differentiated into proximal versus distal causes (Wiley et al., 2014) and often codes are created to document the number of misconceptions or erroneous causes included in the essay (Wiley et al., 2009, 2011, 2014). Other coverage codes can identify non-central content including discussion of background information or non-essential details (irrelevant elaborations) as part of the essay (Perfetti et al., 1995; Wiley et al., 2014). Coverage (number of overall idea units) generally does not predict learning, but significant correlations are typically observed between comprehension test scores and coverage of the key ideas identified as part of an *a priori* causal model (Wiley et al., 2009; Wiley et al., 2011). Negative correlations can be seen when essays include misconceptions (Hemmerich & Wiley, 2002). In the present study, the inclusion of key ideas from the causal model can represent an index of the quality of a student's mental model, and we would expect it to predict performance on the comprehension test.

The final dimension of essay quality considered in this study builds on the work of Scardamalia and Bereiter (1987) and Spivey (1990), who made a distinction between knowledge-telling and knowledge-transforming when students compose essays to demonstrate their understanding of a topic. Telling is regarded as a passive transfer of information from text to paper, whereas transformation is regarded as a more active and constructive process in which the writer relates the contents of sources in new ways by making novel connections within source material, as well as connections to the writer's knowledge. Knowledge-telling involves a relatively superficial interaction with the text, whereas knowledge transforming involves more active construction of a mental model from the text contents. Several measures have been developed with the goal of assessing the extent to which students attempt to integrate and transform information as they write. One measure has been the incidence of connections and connectives included in each essay (Britt & Aglinskas, 2002; Voss & Wiley, 1997; Wiley & Voss, 1999). This serves as an index of the extent to which students attempt to connect or integrate ideas, rather than just reporting what they read. Students who demonstrate better understanding of the material on comprehension tests tend to write essays that have more connected ideas, and more causal connections (Voss & Wiley, 1997; 2000; Wiley, 2001; Wiley & Voss, 1999).

Another measure of integration and transformation (based on Greene, 1994) considers the origin of information included in each sentence of an essay. In this approach, each sentence is scored as to whether or not it contains a connection between idea units that were presented in the reading materials. This measure represents the extent to which students recognize possible relations among factors. The connections that are generally included in this analysis are attributions, correlations, temporal links, simple conjunctions, and causal links. Ideas that co-occur in the same sentence, even without a connective term, can also be coded as connected. These sentences show that the reader has connected and integrated information within a sentence. This is similar to coding for the incidence of connections, but doing it on a per sentence basis. In this approach, the content of each sentence is classified into one of three categories: transformed, added, or borrowed (Wiley, & Voss; 1996; 1999). Sentences that combine some presented information with a new claim or fact, or that integrate two bits of presented information that were not previously connected, are classified as transformed. A sentence is coded as added when it contains only novel information. Sentences that are taken directly from, or are paraphrased from, the original material are classified as borrowed. Students who demonstrate better understanding of the material on comprehension tests write essays that contain a lower proportion of borrowed or copied sentences (Voss & Wiley, 1997; 2000; Wiley, 2001; Wiley & Voss, 1999). Thus, in the present study we would expect the number of connections that students include in their essays to positively predict understanding, while borrowing or copying of information might be a negative predictor.

Automated Scoring of Explanation Quality and Student Understanding from Existing Technologies

Given the kinds of features of student explanations that have been explored using hand-scoring, an obvious question is whether there might be existing technologies that can provide automated metrics for each of them. One simple approach used in many automatic scoring approaches has

been to use the length of the essay as a measure. Length is easily obtained from automated systems, as well as from basic text editors and word processing programs. Another measure that can be easily automated using simple pattern matching approaches is computing the frequencies of citations or references to documents (Britt et al., 2004; Foltz, Britt, & Perfetti,, 1996).

The more difficult features to automatically generate are those that attempt to capture the quality of student explanations, especially in terms of their macrostructure or causal structure. Much of the previous work that has attempted to detect student understanding of subject matter from written responses has been done within Intelligent Tutoring Systems (ITS) such as AutoTutor and MetaTutor. In these cases, students provide written responses as part of a tutoring dialogue, and the goal of assessment is determining which feedback or instructional scaffolds should be given to the tutee by the ITS (e.g. Graesser, Wiemer-Hastings, Wiemer-Hastings, Person et al., 2000; Graesser, McNamara, & VanLehn, 2005; Lintean, Rus, & Azevedo, 2011). In this work, a common assessment method has been to assess the similarity of each student response to a set of idealized target responses using Latent Semantic Analysis (LSA, Landauer, Foltz, & Laham, 1998). This type of approach generally does well at identifying the content material present in written responses. In the right type of discourse context, feedback based on this automated assessment of similarity to idealized responses can be very effective for helping students learn (VanLehn et al., 2007). It must be noted, however, that work done in these contexts generally requires students to only write very short responses (a word, phrase or sentence), so the difficulties of identifying larger elements of structure from the responses do not apply.

There have been some attempts to use LSA and ITS methods with essays and longer texts. For instance, LSA has been used to analyse the quality of student understanding by comparing student essays to expert essays (Foltz et al., 1996), or to sentences judged important by experts (Foltz et al., 1996), or to idealized peer essays (Ventura et al., 2004). A similar approach is attempted here using an idealized peer essay. This essay is referred to as idealized to emphasize that the fact that the text we use for automatic similarity scoring in LSA is not an actual essay written by an individual student, but a compilation essay made from combining several peer responses that provide full coverage of an *a priori* model. In this way, LSA can be used to provide an index for the quality of a student's mental model, and this index should positively predict student understanding as assessed by performance on the comprehension test.

Similarly, LSA can be used to assess the amount of transformation present in student essays by directly comparing student essays to the source documents that they read (Britt et al., 2004; Foltz et al., 1996). In these studies, student sentences that had an LSA cosine with a source sentence above an empirically determined threshold were identified as borrowed, copied or plagiarized unless proper citation was detected (using pattern matching). This approach also enabled calculation of a "coverage" score: the extent to which the student essay said something like what was in each of the source documents, using a lower cosine threshold (Hastings, Hughes, Magliano, Goldman, & Lawless, 2012).

Coh-Metrix (McNamara, Graesser, McCarthy, & Cai, 2014) is another system that has been previously employed in attempts to evaluate student writing, primarily with the goal of assessing the composition quality of persuasive essays written in response to SAT-style prompts

such as “Do images and impressions have a positive or negative effect on people?”. In a study attempting to identify which Coh-Metrix indices best predict expert evaluations of student writing, Crossley and McNamara (2011) reported that both essay length and lexical diversity were positive predictors. Students who wrote longer essays and used more diverse vocabulary were given better scores by human raters. Further, these authors have also found that similarity among adjacent sentences and the presence of causal language can serve as *negative* predictors of expert ratings for composition quality (Crossley & McNamara, 2010). However, these Coh-Metrix indices (lexical diversity, similarity among sentences, and presence of causal language) may predict student understanding differently than they predict expert ratings of writing quality.

As noted above, prior work predicts positive relations may be found between these indices (similarity and causal language) and comprehension test performance, because the extent to which students integrate ideas across sentences and use causal connectives has already been shown to relate to better student understanding (e.g. Wiley & Voss, 1999). Similarly, although using diverse vocabulary may relate to higher grades on writing assignments that may be similar to those given in an English composition class, the use of too many different terms in response to an inquiry activity for learning in science may mean that a student is not focussing on creating a coherent explanation from the sources. This suggests that lexical diversity might not have a strong positive relationship in this context. One could imagine that too little lexical diversity could be a sign of poorly developed explanations, but that too much diversity could be a sign of a lack of coherence or lack of focus on the inquiry goal of explaining a particular phenomenon or outcome using a particular set of source documents. Lexical diversity has been suggested to reflect the coherence of a writer’s mental representation about an event or topic (Pennebaker, 1993; Wade-Stein & Kintsch, 2004). That is, a person writing coherently about a single topic will use more of the same words than someone whose writing is more scattered. Thus, in the present study, to the extent that these indices relate to integration and transformation of information, and represent a focus on the development of a causal mental model of the topic, we would expect lexical diversity to negatively predict student understanding, and similarity and causal language to positively predict understanding.

Although LSA and Coh-Metrix may both provide some useful indices of writing quality, in general it has been suggested that generic cohesion-based approaches without grounding in content have not fared as well as more specific content-based approaches (Graesser & McNamara, 2012; Magliano & Graesser, 2012). Because of this, one might assume that out-of-the-box analyses using LSA or Coh-Metrix will be unable to capture students’ mental models very well, and that a specially tailored machine learning approach is likely to be needed to robustly predict student understanding from this inquiry activity. However, rather than making this assumption, the current study first tested the extent to which these available technologies might be able to detect student understanding from the written responses, before proceeding to develop and test a machine-learning approach.

METHODS

Learning Context and Learning Outcome Measures

The dataset consisted of 178 explanation essays generated by middle school and high school students who learned about the causes of global warming as part of a multiple-document inquiry

task, and who also completed the learning outcome measure following reading and writing. Students were asked to write an essay, "explaining how and why recent patterns in global temperature are different from what has been observed in the past." All participants were given a set of 7 documents containing information related to the causes of global temperature change. Five text-based documents covered several main topics including Ice Ages, the Carbon Cycle, The Greenhouse Effect, Solar Radiation, and Energy from Fossil Fuels. The document set also included a graph of CO_2 Concentrations over the last 400,000 years, presented as its own document. In addition, students were provided with a seventh document, titled "Changes in Global Temperatures", which provided textual background on the methods used to assess global temperatures. This document also included a graph of average global temperatures over the last 400,000 years, and a second graph showing the increases in average global temperatures from 1870 to 2010. The texts were excerpted from several online sources from the United States Geological Survey, the Public Broadcasting Service, the NASA earth observatory, the Environmental Protection Agency, as well as an extension module from an earth science textbook series (Bennington, 2009). On average, the text-based documents were 326 words long (range: 208-475), with a Flesch Reading Ease of 62.36, and an average Flesch-Kincaid grade level of 7.9.

The document set was designed to include all information necessary to construct a coherent representation of the topic, based on an *a priori* causal model, but the text set also required integration of ideas across documents in order to achieve an understanding and answer the question of "how and why recent patterns in global temperature are different from what has been observed in the past". Figure 1 gives a graphical representation of the concepts that were available to explain recent changes in average global temperatures (the target outcome is represented by the parallelogram). Each of the documents contained information that contributed to the creation of this causal model of global warming. No single document contained all the necessary information. Further, none of the documents directly addressed the inquiry question. Each addressed a different specific issue (e.g., CO_2 trapping heat; or human contributions to CO_2) that can be made relevant to the inquiry question only when repurposed and combined with information from other documents by the reader.

After reading the documents and writing their essays (with documents present), the documents were collected and students completed an inference verification test (IVT). The IVT was intended to assess the mental model of the causes of global warming that students constructed while engaging in the multiple document inquiry activity. This learning outcome assessment was based on techniques developed by Royer and his colleagues (Royer, Carlo, Dufrense, & Mestre, 1996). The test contained 18 statements that represented potential conclusions, connections, or inferences that could (or could not) be made based on the information in the document set and were consistent with the *a priori* model shown in Figure 1. In this test, students needed to verify whether propositions followed (or did not follow) from the information contained in the documents. Some example items are "In the past 100 years, both fossil fuel use and CO_2 levels have increased" and "Increases in fossil fuel use increase the amount of heat that escapes into space." The first sentence is an example of a conclusion that is supported by the documents, but was not explicitly stated and requires the reader to make connections across documents to verify. The second is an example of a conclusion that was the opposite of a relation that could be inferred based on the documents. An overall proportion

correct score was computed for the task, and higher levels of performance indicated better understanding of the conclusions that could be drawn by integrating ideas across the documents. Previous work has shown that performance on inference verification tasks reliably correlates with other measures of understanding including the quality of students' written explanations (Griffin et al., 2012; Jaeger & Wiley, 2015; Sanchez & Wiley, 2006; Wiley & Voss, 1999; Wiley et al., 2009).

Hand Scoring of Explanation Essays

The explanation essays derived from this inquiry activity were hand-coded using two different systems. The first system coded the explanation essays from the perspective of the *a priori* causal model in Figure 1 and gave students credit whenever they made causal connections between the nodes in Figure 1. (For similar work using this approach see Griffin et al., 2012). The second system was focussed upon the argument structure present within each student's essay and scored the essays for the presence of causal chains of ideas that culminated in recent changes in global temperatures. (For similar work using this approach see Hastings et al., 2014). Since both of these approaches are based on the *a priori* causal model created by the document set, only ideas and concepts present in the documents received credit for each approach. Both approaches used two independent raters to code the explanation essays. Raters first scored a small subset of the essays on their own (typically about 12-15 examples) and compared their responses. Following discussion, the remainder of the essays was independently scored. Interrater reliabilities computed with Cohen's Kappa were above .80 for all coded measures. Skewness and kurtosis for all measures reported below are less than 1.

Using the *a priori* causal model, humans evaluated the explanations to identify which causal concepts were present (nodes in Figure 1) and which were explicitly linked to each other.

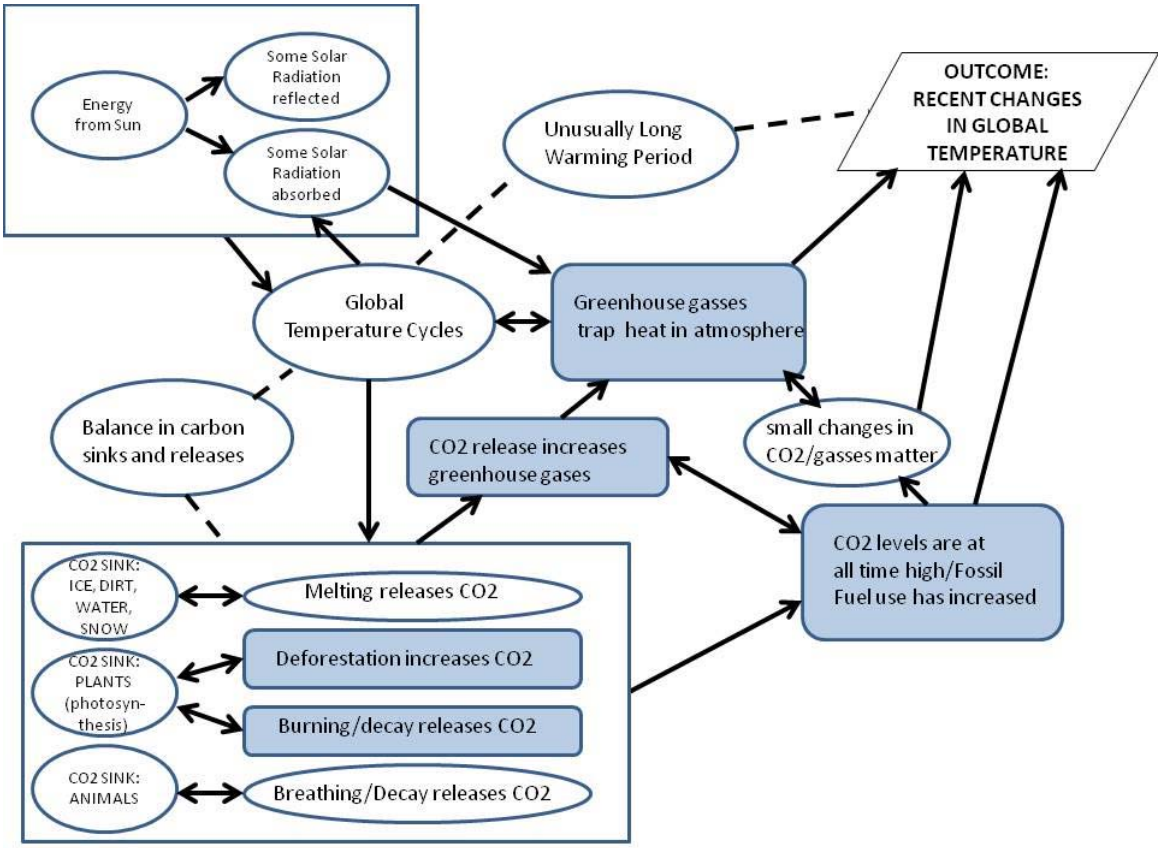

Figure 1. Concepts available in the text related to causes of recent changes in global temperatures.

Of the core concepts included in the *a priori* causal model used to create the document set (MODELCONC), students generally mentioned fewer than 5 in their explanations ($M = 4.37$, $SD = 2.48$). A subscore (TARGCONC) was computed based on the presence of just the 5 critical target concepts that most directly related to recent changes in global temperature as highlighted in Figure 1. Of these 5 target concepts, students mentioned only an average of 1.35 ($SD = 1.21$) in their explanations. The number of explicit connections that students made among the concepts was also coded (MODELCONN). On average, students made 1.70 ($SD = 1.71$) explicit connections between concepts.

The second coding scheme coded the written explanations for the argument structure present in each essay by exploring the number and length of causal chains that students developed against the idealized causal chain shown in Figure 2. The target outcome is represented by a parallelogram and there are several paths that can be connected to explain this outcome. Students' explanation essays were scored for the length and number of explanatory chains. We first scored the overall number of propositions or elements (PROPS) that were included in their explanation essay (rectangles in Figure 2). On average about one-third of these elements were mentioned ($M = 6.58$, $SD = 2.91$). Then we counted the number of connections along causal chains (LINKS) that were connected to the target outcome ($M = 2.61$, $SD = 2.38$), and the number of chains that

included intervening connections between initiating factors and the target outcome (CHAINS) (*M* = 0.70, *SD* = .71). The one exception to chain coding was that one chain that was explicitly described in the text (the fossil fuel chain represented by 0-3-50 below) was coded separately (EASYCHAIN).

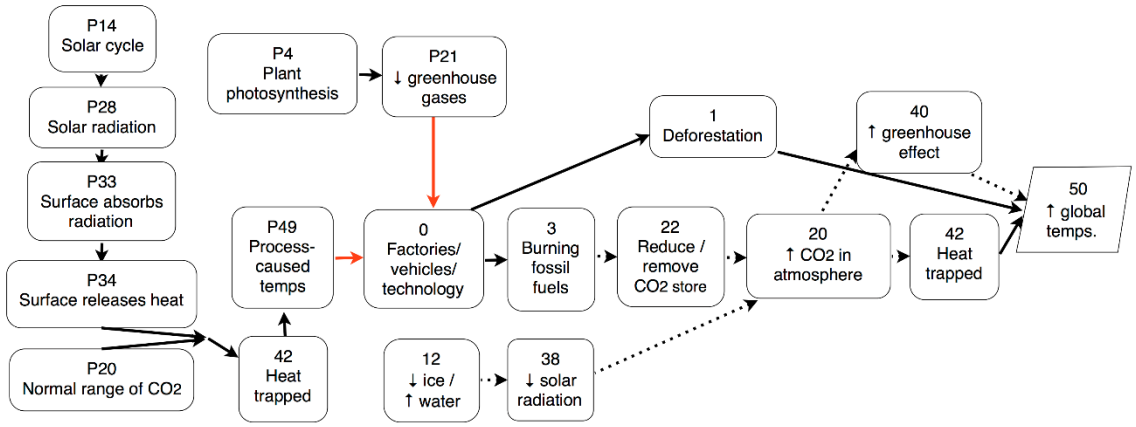

Figure 2. Example Argument Structure in Student Essay linked to Outcome

Finally, a holistic code was developed to categorize the explanations into five hierarchical levels of quality (EXPQUAL): (1) *No core content* (did not include any elements) (N = 8), (2) *No causal chains* (included core elements but did not connect to the outcome) (N = 32), (3) *No intervening factors* (connected at least one element to the outcome directly, but had no chains of connections to the outcome) (N = 38), (4) *Simple intervening factor* (EASYCHAIN: the only argument chain was the one that was stated explicitly in the text: from increased factories/vehicles/technology to increased use of fossil fuels) (N = 14), (5) *Advanced intervening factor* (at least one causal chain with at least one intervening element other than the "easy" chain mentioned above) (N = 86). The *No core content* responses failed to identify any important information that could be connected to the outcome. The *No causal chain* responses included elements that could be part of the explanation, but did not make it clear that the element was leading to the outcome by articulating an explicit connection. Almost 23% of the students did not create a minimally connected explanation. *No intervening factors* responses were also very common. In these explanations, at least one element was directly connected to the outcome. Often 2 or more distinct chains were present, but these chains were not connected to each other. A common example of the argument structure present in this type of response was the student asserting that increased fossil fuel use leads to increased temperatures, that increased CO_2 in the atmosphere leads to increased temperatures, and that increased temperatures are due to more heat being trapped, but these three relations were stated separately not linked together by the student. The final two levels represented explanations that included at least one chain with an initiating cause connected to an intervening factor that was then connected to the outcome. One chain (from increasing vehicles, factories, and technology to the increase in fossil fuel use to global warming) was treated separately because the links between these were included in the documents, and therefore the student did not need to construct this argument. Because we were interested in

identifying transformation, we separated out those explanations that did not require much text-based transformation because they included only this chain, from those that involved more active transformation of the sources. A similar hierarchical approach has been used in a study on another scientific topic, coral bleaching (Hastings et al., 2016). In that study, the researchers found that the four quality categories used were associated with learning using scores on a multiple choice test as comparison. The lowest quality group had significantly lower learning (32%) than the middle two quality categories (47%, 52%) which were each lower than the highest quality group (63%). These results show that the quality categorization scoring has utility as a measure of learning.

RESULTS

Basic Descriptive Measures

On average, the explanations that students wrote were around four paragraphs ($M = 4.12$, $SD = 1.75$), or 21.12 ($SD = 8.32$) sentences and 313.33 ($SD = 108.44$) words long. The three measures of response length (words, sentences and paragraphs) were all related (rs > .50, ps <. 01). To avoid multicollinearity issues, only the smallest grain size for length (words) is included in tables and analyses (LENGTH). Only a third of the students included at least one reference to the documents to cite the source of their evidence or ideas in their responses (N = 60/178, SOURCE). A simple macrostructure code categorized whether the student answer directly responded to the prompt (ANSWERQ). This code indicated whether students attempted to write an explanation about how and why recent patterns in global temperature are different from what has been observed in the past (1), or if they did something else (0) such as write their opinion about global warming, or just simply list ideas from the text without using ideas to try to answer the question directly. The majority of students were coded as attempting to address the essay prompt in their explanations (N = 142/178).

Table 1. Correlations among descriptive measures, hand-scored measures of explanation quality, and learning outcomes (IVT)

	EXPQUAL	LENGTH	ANSWERQ	SOURCE	MODEL CONC	TARG CONC	MODEL CONN	PROPS	LINKS	CHAINS	EASY CHAIN
IVT	.34**	.23**	.18*	-.07	.37**	.42**	.55**	.33**	.38**	.40**	-.20**
LENGTH	.18*		.12	.23**	.50**	.26**	.35**	.47**	.25**	.21**	-.09
ANSWERQ	.58**			-.03	.34**	.37**	.39**	.41**	.46**	.39**	-.01
SOURCE	-.09				-.01	-.17*	-.09	-.02	.-.02	-.10	.10
MODELCONC	.38**					.64**	.55*	.53**	.34**	.38**	-.17*
TARGCONC	.47**						.56**	.41**	.47**	.51**	-.15*
MODELCONN	.55**							.49**	.56**	.60**	-,18*
PROPS	.58**								.61**	.57**	.04
LINKS	.79**									.84**	.01
CHAINS	.70**										-.22**
EASYCHAIN	.05										

Note. N=178, * $p<.05$, ** $p<.01$, IVT = comprehension test scores, EXPQUAL = Explanation Quality, LENGTH = Number of words, ANSWERQ = whether response was structured to answer the question, SOURCE = reference or citation to source, MODELCONC = concepts present from causal model, TARGCONC = target concepts from causal model, MODELCONN = connections among concepts, PROPS = propositions in written argument, LINKS = argument elements connected to outcome, CHAINS = number of paths of explanation for outcome including an intervening factor, EASYCHAIN=argument about fossil fuels that was present in the text,.

Predicting Explanation Quality from Hand Scoring Approaches

One main purpose of this study was to test which of the various approaches to coding the explanations might best predict explanation quality and performance on tests of student understanding. The simple correlations among the basic descriptive and hand-scored metrics and their ability to predict the holistic explanation quality scores (EXPQUAL) and performance on the test of understanding (the IVT) are shown in Table 1. Table 1 shows that the number of words (LENGTH) in each explanation was a significant predictor of both explanation quality (EXPQUAL) and comprehension test scores (IVT).

The presence of references to the documents or citations of the sources of the documents as part of the essay (SOURCE) was not associated with explanation essay quality or test scores. If anything, readers who were more likely to refer to the documents when writing about this science topic were less likely to focus on the most important information (TARGCONC). While the presence of citations when writing from multiple documents in history is usually related to better quality essays (e.g., Britt & Aglinskas, 2002), this may be due to the important role of a source in evaluating historical documents (e.g., perspective, bias, time, culture; see Rouet et al., 1996); or when a document set includes opposing theories or discrepancies (Bråten, Strømsø, & Britt, 2009; Rouet et al., 1996), which was not the case in this particular activity.

The overall macrostructure of the response (whether it attempted to answer the question by providing an explanation; ANSWERQ) was also a significant predictor of both explanation quality and test scores.

In the simple correlations shown in Table 1, all three measures derived from the coding based in the *a priori* causal model (MODELCONC, TARGCONC, MODELCONN) predicted both explanation quality and comprehension test scores. When measures listed in Table 2 were entered into a simultaneous regression to test for unique predictors of explanation quality, the overall model provided a good fit, $F(4, 173) = 39.05$, $MSE = 0.96$, $p < .001$, and accounted for 47% of the variance in explanation quality. Attempting to answer the question (ANSWERQ), including target concepts (TARGCONC), and making connections (MODELCONN), all accounted for unique variance. (MODELCONC was not in the model to avoid multicollinearity due to its high correlation with TARGCONC). Essay length (LENGTH) was not a significant unique predictor of explanation quality when included in a model with these other measures.

Table 2. Holistic Explanation Quality Scores as predicted by codes for a priori causal model

Variable	Unstandardized Beta (*B*)	Std. Error	Standardized Beta (β)	*t*-value	*p*-value
(Constant)	2.13	.26		.188	<.001
LENGTH	.00	.00	-.02	-.26	.80
ANSWERQ	1.33	.20	.40	6.58	<.001
TARGCONC	.17	.08	.16	2.28	<.03
MODELCONN	.24	.06	.31	4.43	<.001

Note: LENGTH = Number of Words, ANSWERQ = whether response was structured to answer the question, TARGCONC = target concepts from causal model, MODELCONN = connections among concepts

As shown in Table 1, significant correlations were also seen between holistic explanation quality scores and the measures derived from the analysis of the arguments present in the student explanation essays (PROPS, LINKS, EASYCHAIN, CHAINS). When the measures in Table 3 were entered into a simultaneous regression, the overall model provided a good fit of the data, $F(4, 173) = 97.68$, $MSE = .56$, $p < .001$, and accounted for 69% of the variance in explanation quality scores. All measures were unique predictors of explanation quality. The strong prediction of these codes for explanation quality would be expected because the holistic explanation quality score was computed based on these measures.

Table 3. Holistic Explanation Quality Scores as predicted by codes for essay argument structure

Variable	Unstandardized Beta (*B*)	Std. Error	Standardized Beta (β)	*t*-value	*p*-value
(Constant)	2.38	.14		16.78	<.001
PROPS	.05	.03	.11	2.03	<.05
LINKS	.17	.05	.31	3.57	<.001
CHAINS	.90	.16	.15	3.26	<.001
EASYCHAIN	.75	.23	.49	5.63	<.001

Note: PROPS = propositions in written argument, LINKS = argument elements connected to outcome, CHAINS = number of paths of explanation for outcome including an intervening factor, EASYCHAIN=argument about fossil fuels that was present in the text.

Predicting Student Understanding from Hand-Scoring Approaches

The next question was to what extent the two sets of hand-coded measures would uniquely predict student understanding. When the measures in Table 2 were entered into a simultaneous regression to test for unique predictors of comprehension scores (IVT), the best fitting model included only the number of target concepts (TARGCONC) and number of connections between concepts (MODELCONN). This model, shown in Table 4, provided a good fit for the data, ($F(2, 179) = 43.19$, $MSE = .01$, $p < .001$), and accounted for 33% of the variance in test scores.

Table 4. Comprehension Test Scores (IVT) as predicted by codes for a priori causal model

Variable	Unstandardized Beta (*B*)	Std. Error	Standardized Beta (β)	*t*-value	*p*-value
(Constant)	.62	.01		46.62	<.001
TARGCONC	.02	.01	.16	2.15	<.03
MODELCONN	.04	.01	.47	6.30	<.001

Note: TARGCONC = target concepts from causal model, MODELCONN = connections among concepts

The measures derived from the analysis of the arguments present in the student explanation essays (Table 3) were entered into a simultaneous regression to test for unique predictors of comprehension test scores (IVT), the best fitting model included the number of propositions and number of links. The model shown in Table 5 was a good fit of the data, $F(3,$

174) = 14.70, *MSE* = .02, $p < .001$, and accounted for 20% of the variance in the test scores. The number of links and the presence of the chain that was present in the text were significant unique predictors of IVT scores, while the number of propositions in the written argument was marginal.

Table 5. Comprehension Test Scores (IVT) as predicted by codes for essay argument structure

Variable	Unstandardized Beta (*B*)	Std. Error	Standardized Beta (β)	*t*-value	*p*-value
(Constant)	.62	.02		26.73	<.001
PROPS	.01	.00	.16	1.92	<.06
LINKS	.02	.01	.28	3.31	<.001
EASYCHAIN	-.10	.03	-.20	-2.99	<.01

Note: PROPS = propositions in written argument, LINKS = argument elements connected to outcome, EASYCHAIN=argument about fossil fuels that was present in the text

These analyses suggest that measures of coverage of ideas in essays, and the extent to which ideas are connected or integrated, are critical features for predicting understanding.

Automated Scoring of Explanation Features using LSA/Coh-Metrix

The second phase of analyses attempted to automatically capture these critical features of the student explanations that emerged from hand-coding. In a first pass at automated scoring, we attempted to leverage existing technologies including using LSA to assess the similarity of the explanations with an idealized peer explanation to generate a coverage score; LSA to assess similarity of the explanations to the source material to create a plagiarism or copying score; and Coh-Metrix to assess cohesion, causality, and lexical diversity of the explanations.

Idealized Peer Explanation Similarity Scores. One LSA approach compared student explanations to an idealized peer explanation (i.e. an explanation constructed by the researchers from the best student responses). The idealized student explanation is included in the Appendix. The approach of using an idealized student explanation essay rather than explanation essays written by experts for comparison was based on Ventura et al. (2004) who reported better prediction from peer-based examples. The idealized explanation was assembled from the two best student essays such that it would score highly on all of the hand-coded features that predicted learning outcomes. We verified that the LSA similarity scores with the idealized essay correlated well with the hand-coded measures (.39 with TARGET CONCEPTS, .45 with MODEL CONNECTIONS, .67 with PROPOSITIONS, .41 with LINKS). LSA was used to compare all student explanations for similarity to the idealized peer explanation essay using the whole-essay-to-whole-essay comparison tool at lsa.colorado.edu. As shown in Table 6, similarity to the idealized explanation essay (IDEAL) predicted both holistic explanation quality scores (EXPQUAL) and the learning outcome measure (IVT).

Table 6. Relation among LSA-derived COPY and IDEAL scores, Coh-Metrix measures, Machine Learning measures, Learning Outcomes (IVT) and Holistic Explanation Quality Scores (EXPQUAL)

	IVT	COPY	IDEAL	CAUSAL	COH	LEXDIV	MLCODES	MLCONN	MLEASY	MLCHAINS
EXPQUAL	.34**	-.29**	.47**	.10	.22**	-.22**	.56**	.32**	.08	.23*
COPY	.02		.27**	-.03	-.01	.06	-.01	-.15*	-.05	-.05
IDEAL	.40*			.12	.19	-.31*	.59**	.26**	.02	.19*
CAUSAL	.21**				.12	-.06	.13	.22	.09	.07
COH	.26**					-.40**	.09	.22**	-.02	.26*
LEXDIV	-.08						-.28**	-.36**	-.00	-.35
MLCODES	.32**							.44**	.15*	.23**
MLCONN	.14								.27**	.55**
MLEASY	-.02									-.06
MLCHAINS	.15*									

Note. N=178, * $p<.05$, ** $p<.01$, IVT = Comprehension Test Score, EXPQUAL = Explanation Quality, COPY = LSA estimate of similarity to original sources, IDEAL = LSA estimate of similarity to idealized peer explanation, CAUSAL = Coh-Metrix index of causal terms, COH = Coh-Metrix similarity between paragraphs, LEXDIV = Coh-Metrix Lexical Diversity, MLCODES = Machine Learning estimated number of propositions, MLCONN = Machine Learning estimate for number of connections, MLEASY = Machine learning detection of easy chain from text, MLCHAIN = Machine Learning detection of chains to outcome.

Plagiarism Scores. Based on previous work (Britt et al., 2004)**,** LSA similarity scores were computed between student explanations and the original sources to estimate how much of each student's essay was copied directly from the source documents (and therefore not transformed). Using lsa.colorado.edu's TASA "General_Reading_up_to_12th_Grade (300 factors)" document space comparison, we computed the cosine between each student sentence and each of the source document sentences. When cosines were above .75, we considered that sentence as copied from the original source (Britt et al., 2004). On average, 32% (*SD* = .20) of sentences in student explanation essays appeared to be copied from the sources. Table 6 shows the simple relations between the LSA-based COPY score, explanation quality and student learning, including a significant negative correlation with explanation essay quality. This suggests that argument quality suffers as students fail to transform information as they write. However, in this case, no overall negative effect of copying (COPY) was seen on the learning outcome measure (IVT).

As shown in Table 7, entering the similarity to the idealized explanation essay score (IDEAL) along with the plagiarism score (COPY) predicted 41% of the variance in the explanation quality scores (EXPQUAL), $F(2, 175) = 61.20$, $MSE = 1.06$, $p < .001$. When entered together in this simultaneous regression, similarity to the idealized explanation essay (IDEAL) positively predicted explanation quality (EXPQUAL), while similarity to source documents (COPY) negatively predicted explanation quality. For the learning outcome measure, adding similarity to source documents (COPY) to a model with the IDEAL scores did not improve the fit. IDEAL scores predicted learning at r=.40 as shown in Table 6, meaning that they accounted for 16% of the variance in comprehension test scores.

Table 7. Holistic Explanation Quality Scores as predicted by similarity to source documents and idealized essay

Variable	Unstandardized Beta (*B*)	Std. Error	Standardized Beta (β)	*t*-value	*p*-value
(Constant)	-.55	.53		-1.04	.30
IDEAL	7.65	.77	.59	9.84	<.001
COPY	-3.00	.40	-.45	-7.48	<.001

Note: COPY = LSA estimate of similarity to original sources, IDEAL = LSA estimate of similarity to idealized peer explanation

Interestingly, even though the plagiarism scores (COPY) and the similarity to the idealized explanation essay scores (IDEAL) predicted explanation quality in opposite directions, the two were found to be positively related to each other (Table 6). This positive relation suggests that students who were copying individual sentences were generally selecting relevant content to transcribe into their explanations, which may explain why copying did not have a negative relation with learning. Although actively transforming information may be the best strategy for understanding, selecting and copying relevant information seems likely to be better than writing irrelevant information or failing to engage with the text at all. Also, since none of the documents specifically provided an answer for the essay question, even copying isolated sentences entailed some level of repurposing of information.

Coh-Metrix Indices for Casuality, Cohesion and Lexical Diversity. The explanations were also submitted to Coh-Metrix as an automated approach to scoring the extent to which the essays integrated and transformed information into a coherent essay. In particular, we used SMCAUSvp (the incidence causal verbs and causal particles) as a measure of causality (reported as CAUSAL in the table). This was motivated by earlier work that found that students who demonstrate better understanding of the material on comprehension tests tend to write essays that have more connected ideas, and more causal connections (Britt & Aglinskas, 2002; Voss & Wiley, 1997; 2000; Wiley, 2001; Wiley & Voss, 1999). For this measure, a higher score means higher incidence of causal terms in the essays (Results are similar for SMCAUSv and SMCAUSp. We selected SMCAUSvp because hand-scoring used both verbs and particles. The other SMCAUS measures did not predict learning outcomes). As a measure of cohesion, we used standardized scores for LSAPP1 (LSA similarity among adjacent paragraphs, reported as COH in the table). This was supplemented by standardized scores for LSASS1 for the 21 essays that were only one paragraph long (instead of using 0 scores for LSAPP1 as assigned by Coh-metrix). For this LSA measure, a higher score means more similarity across parts of the response, representing more cohesion. Finally, we also explored the lexical diversity of all words using LDTTRa (LEXDIV in the table), as another potential measure of cohesion or focus. A higher LEXDIV score means that the response contained a broader range of vocabulary, while a lower LEXDIV score means that a more restricted range of words were used. The CAUSAL scores for student responses ranged from 19.87 to 138.89 (M = 52.71, SD = 16.18). The LSAPP1 scores that were used to compute COH ranged from .09 to .70 (M =.33, SD =.12). LEXDIV scores ranged from .30 to .75 (M =.49, SD =.07). The relations among the measures are shown in Table 6. Consistent with the notion that too much lexical diversity can be a sign of a lack of focus or coherence in explanation essays, there was a significant negative relation between cohesion indices (COH) and lexical diversity (LEXDEV).

The correlations in Table 6 also showed that scores of explanation quality (EXPQUAL) and comprehension test scores (IVT) both increased with cohesion (COH). Comprehension test scores (IVT) increased with the number of causal expressions (CAUSAL), while lexical diversity (LEXDIV) was found to be a negatively related to the holistic scores of explanation quality (EXP QUAL). Although the significant relation between causality (CAUSAL) and learning (IVT) replicates prior work using hand coding for causal expressions (Britt & Aglinskas, 2002; Voss & Wiley, 1997; 2000; Wiley, 2001; Wiley & Voss, 1999), the magnitudes are modest compared to the relations seen with hand-coding for connections in Table 1 (c.f. MODELCONN and LINKS). One possible reason for this difference is that the set of verbs and particles that Coh-Metrix uses to code for causality might be limited to a set of generic causal terms (change, cause, enable, and make) while the terms used to code as causal connections during hand-coding (MODELCONN and LINKS) included more context-specific causal terms, for example “increases”, “helps”, “intensifies”, “traps”, “heats” and “melts”. Also, Coh-Metrix counts causal connectives that are completely redundant with others, as well as causal terms that are used in expressions unrelated to expressing causal relations (e.g., “We need to *change* our way of life.”, “This *makes* sense.”), which could weaken the relation. The predictions from cohesion scores (COH) and lexical diversity scores (LEXDIV) were also modest in magnitude (rs beween -.22 and .26).

Table 8. Holistic Explanation Quality Scores as predicted by Coh-Metrix indices

Variable	Unstandardized Beta (*B*)	Std. Error	Standardized Beta (β)	*t*-value	*p*-value
(Constant)	5.46	.76		6.22	<.001
CAUSAL	.01	.01	.07	.94	.35
COH	.18	.11	.14	1.69	.09
LEXDIV	-3.14	1.49	-.17	-2.11	<.04

Note: CAUSAL = Coh-Metrix index of causal terms, COH = Coh-Metrix similarity between paragraphs, LEXDIV = Coh-Metrix Lexical Diversity

When the Coh-Metrix measures were submitted to a simultaneous regression predicting explanation quality, only lexical diversity was found to be a unique predictor as shown in Table 8. This model was significant, $F(3, 173) = 4.67$, $MSE = 1.68$, $p < .01$, but predicted only 8% of the variance in explanation quality. On the other hand, causality and cohesion were significant predictors of comprehension test scores as shown in Table 9. Again this model was significant, $F(3, 177) = 6.44$, $MSE = .02$, $p < .001$,but predicted only 10% of the variance in test scores.

Table 9. Comprehension Test Scores (IVT) as predicted by Coh-Metrix indices

Variable	Unstandardized Beta (*B*)	Std. Error	Standardized Beta (β)	*t*-value	*p*-value
(Constant)	.62	.08		8.11	<.001
CAUSAL	.00	.00	.17	2.42	<.02
COH	.03	.01	.24	3.04	<.01
LEXDIV	.01	.14	.01	.07	.94

Note: CAUSAL = Coh-Metrix index of causal terms, COH = Coh-Metrix similarity between paragraphs, LEXDIV = Coh-Metrix Lexical Diversity

Best Fitting Model from Out-of-the-Box Approaches

The simple correlations among the indices derived from the three approaches are shown in Table 6. The IDEAL scores positively related to the cohesion measure derived from Coh-Metrix and negatively related to the lexical diversity measure. Submitting the idealized peer explanation to Coh-Metrix showed that it had above average cohesion, COH=.57, and below average lexical diversity, LEXDIV = .38, compared to the corpus. Both of these features could reflect the focus of the idealized essay on explaining a particular topic, which results in restricted vocabulary usage and overlap among sentences. The idealized explanation essay also had a CAUSAL score of 59 which was slightly above average. This is consistent with prior work showing that causal connections in explanation essays are generally a positive predictor of student understanding.

When the LSA measures (COPY and IDEAL) and the three Coh-Metrix measures were included in a simultaneous regression to predict explanation quality scores, none of the Coh-Metrix measures captured any unique variance. The best fitting model was the model shown in Table 7 which predicted 41% of the variance in the explanation quality scores. In contrast, when the measures derived from both LSA and Coh-Metrix were included in a simultaneous regression to predict student understanding, all measures except the COPY score were found to be significant unique predictors, as shown in Table 10. The model provided a good fit for the data,

$F(5, 171) = 10.25$, $MSE = .02$, $p < .001$, and predicted 23% of the variance in comprehension test scores.

Table 10. Comprehension Test Scores (IVT) as predicted by LSA/Coh-Metrix indices

Variable	Unstandardized Beta (*B*)	Std. Error	Standardized Beta (β)	*t*-value	*p*-value
(Constant)	.15	.11		1.39	.17
COPY	-.06	.05	-.09	-1.25	.21
IDEAL	.54	.10	.41	5.45	<.001
CAUSAL	.00	.00	.13	1.95	.05
COH	.03	.01	.22	2.94	<.01
LEXDIV	.28	.15	.15	1.93	.05

Note. COPY = LSA estimate of similarity to original sources, IDEAL = LSA estimate of similarity to idealized peer explanation, CAUSAL = Coh-Metrix index of causal terms, COH = Coh-Metrix similarity between paragraphs, LEXDIV = Coh-Metrix Lexical Diversity

As in the previous analyses, similarity to the idealized peer explanation (IDEAL), incidence of causal terms (CAUSAL), and cohesion (COH) were all positive predictors of test scores. However, in this combined analysis, lexical diversity was now also a positive predictor of test scores. Follow-up analyses indicated this was due to the addition of the similarity to the idealized explanation scores (IDEAL) to the model. In the presence of this measure (along with measures of cohesion and causality), the use of a broader range of vocabulary emerged as a positive predictor.

In sum, attempts to use out-of-the-box tools were to some extent successful as the models based in metrics derived from automated LSA and Coh-Metrix scores predicted a significant amount of the variance for each outcome measure. However, the fit and amount of variance explained for each outcome was clearly inferior in magnitude to the best fit from the models based on measures derived from human coding in the previous section.

AUTOMATED SCORING USING MACHINE LEARNING

Although LSA and Coh-Metrix did provide some useful indices of writing quality, our next step was to explore a more specific content-based automated scoring approach (Graesser & McNamara, 2012; Magliano & Graesser, 2012). To provide an example of such an approach, prior work on the dialog-based ITS mentioned above, MetaTutor, has also included a task where students wrote a paragraph describing their existing knowledge on a topic (Lintean, Rus & Azevedo, 2011). When researchers compared three methods for identifying students' mental models of a topic from the paragraphs: content-based measures derived from LSA, cohesion-based measures from Coh-Metrix, and word-weighting features specially derived from their corpus of paragraphs; they found that the word-weighting features outperformed the other approaches. To provide another example, researchers have found that neither general indicators of reading strategies nor indicators of textual complexity were effective at predicting 3-5th graders comprehension of stories, but a machine learning approach using a combination of some of these features was effective (Dascalu et al., 2015).

While many Automated Essay Scoring (AES) systems have been developed to provide a more efficient evaluation of student writing (e.g. Larkey & Croft, 2003; Shermis & Hamner, 2012), these systems use a wide variety of generic features -- lexical, syntactic, and semantic (e.g. Deane, 2013; Roscoe, Crossley, Snow, Varner & McNamara, 2014) -- to effectively provide holistic, summative evaluations of student essays. Yet they have been criticized for failing to accurately judge the relevance and appropriateness of student responses (Dikli, 2006) and for their lack of construct validity (Condon, 2013; Roscoe et al., 2014). Some recent research has addressed the construct validity issue for students' persuasive essays by detecting statements of opinion to provide a holistic measure of persuasive essay quality (Farra, 2015). Other recent research has focused on providing formative assessment, using logs of keyboard activity while students were writing (Zhang & Deane, 2015). However, to our knowledge, no one else has tried to use a machine learning algorithm to assess the causal structure of student arguments or explanations in order to serve as an assessment of student understanding.

Developing a system that can identify the causal structure of any text is very difficult. Working with newspaper texts, Rink et al. (2010) tried to develop a system that could identify causal relations between events using a wide range of linguistic resources and techniques, including part-of-speech tagging, syntactic parsing, WordNet (Miller, 1995), VerbOcean for semantic links between verbs (Chklovski & Pantel, 2004), dependency parsing, word sense disambiguation (Mihalcea & Csomai, 2005), and a semantic parser for identifying the semantic frame (Bejan & Hathaway, 2007). With all of these techniques combined, their system achieved Precision[1]= 0.33, Recall = 0.61, and F1 = 0.43. By including manual annotation of the temporal relations between events in the text, they increased performance to F1 = 0.58. Thus, creating a system that may be able to detect causal structure present in student essays in order to measure student understanding represents a central challenge and goal for the present work.

In this section we explore the utility of using a machine learning (ML) approach for assessing the quality of student explanation essays, and using metrics produced by the ML approach to predict student understanding. An overview of the three main steps involved in this process is that we trained ML models on the annotated explanation essays (annotated with the codes in Figure 2) to identify each individual concept code. In the second phase, the identified concepts were used to train models which identified the existence of components of causal connections, and of the specific causal connections between pairs of concepts. Finally, we used the same rule-based process that the human coders did to calculate each essay's holistic explanation quality score. The subsections below describe the first two processes.

Concept Detection

For concept detection, we treated the problem similarly to a tagging problem like part-of-speech tagging. We preprocessed the explanation essays by doing spelling correction and stemming.

[1] Precision is defined as Hits / (Hits + False Alarms) or True Positives / (True Positive + False Positives) where Hits or True Positives are defined as a match between automated and manual coding, misses are defined as when the automated system missed a feature that manual coding marked, and false alarms are when the automated system detected a feature that was not present in manual coding. Recall is defined as Hits / (Hits + Misses). Misses are also known as False Negatives. The F score combines the two. F1 balances them evenly. F1 = 2*P*R/(P+R)

We did not remove stop words. We did replace unique words (words which appeared only once in the entire corpus, most often because they were badly misspelled) with a special UNKNOWN token. Because these words occurred only once, the system could not learn useful information from them anyway. The UNKNOWN token occurs frequently enough that it does not carry strong semantic content for the system, similar to words like "a" or "the". Then we applied our machine learning approach with a 7-word sliding window across the text to identify concepts within that window (Hughes et al., 2015). The fixed-size sliding window approach allows us to avoid the difficulties for machine learning from variable-length input, but the size of the window ensures that the words of a concept will almost always fall entirely within one of the windows (Hughes et al., 2015). For each of the concept codes (the nodes in Figure 2), we trained a logistic regression classifier in which the features were the words and the bigrams within the window, as well as their relative positions within the window.

For example, consider the student sentence, "Factories began to burn large amounts of fossil fuels to create energy." The word "factories" was coded as Concept 0 by the human annotators. For the classifier which predicts Concept 0, the first sliding window across this sentence would include 13 features. The first 7 would be word-based features signifying that the stem of the target word was "factory", the following word stems were "begin", "to", and "burn", and the words before the target word that came before the start of the sentence. There would also be 6 bigram features, including "START-factory-1", "factory-begin+0", "begin-to+1", etc., where the numbers represent the relative position within the window. This set of features would be a positive example of the target class (Concept 0). The next window would have "began" as its central target word, and would be a negative example of Concept 0.

We trained and evaluated the classifiers for the word-level tagging task using 5-fold cross-validation, using 80% of the explanation essays for training and the remaining 20% as the test set (repeated for each of 5 test sets). This gave a classification for each concept code. The results showed that the classification is quite reliable. For the entire set of explanation essays[2], the macro-averaged Precision was 0.77, Recall was 0.71, and F_1 was 0.74.

Detecting connections

In contrast with Hastings et al. (2014) which used hand-coded essays to determine concepts directly, in this study we used the output of the concept detection from the automatic inference mechanism described above as input for automated detection of causal connections.

Because causal connections between concept codes generally occur over a wider span of text, we cannot use the same type of sliding window method to automatically identify them. Instead we trained a higher-level classifier which used as inputs the results of the concept tagging along with three other tags that were learned by the window-based tagger. One was for connectors (e.g. "because of", "as a result") that the coders had annotated in the text. The other two were for concept codes that had been marked as causers and results. The second-level

[2] A total of 222 essays were annotated and used for machine learning. Only 178 are included in the analyses predicting student understanding because the remainder did not have data on that measure. For comparability, the same subset of essays are used for analyses predicting explanation quality scores

classifiers also used these other features based on the results of the first-level classifiers: the minimum and maximum probabilities for each predicted label (code, causer tag, result tag, cause-effect tag), the binary yes/no prediction for each label, and a binary combination prediction for each pair of codes that was identified in the sentence. These were used to train a logistic regression classifier for each causal connection between two concepts that occurred in the training essays.

Again we assessed this method using 5-fold cross-validation. On this task, the classifiers had macro-averaged predictions with Precision = 0.64, Recall = 0.40, and $F_1 = 0.49$. Although this level of prediction accuracy is not as strong as that for concepts alone, this is unsurprising because it relies on the outputs of the first-level concept predictions. The task is also considerably more complex. Instead of "just" trying to predict the 19 codes from Figure 2, we had to potentially distinguish between 19*19 connections (although only 34 combinations actually appeared in the responses).

Using the output from the machine learning of concept coding (MLCODES), connections coding (MLCONN), detection of the easy chain (MLEASY) and detections of other chains to the outcome (MLCHAINS), predictions were then computed for the overall quality of the explanations (MLQUAL). To assign explanation essays to appropriate categories, we employed the same criteria to compute the holistic explanation quality scores as with the hand-coding.

Predicting Explanation Quality and Test Scores with Machine Learning

The ML approach used explanation essays that had been annotated with structure codes as its input. The predictions derived from the ML approach correlated very highly with hand-scoring for number of propositions or elements in the arguments ($r = .71$, $p > .001$), and moderately for the number of links ($r = .35$, $p < .001$), the easy chain ($r = .33$) and other chains ($r = .25$).

Table 11 shows the results of a simultaneous regression predicting explanation quality (EXPQUAL) using metrics derived from all automated approaches (ML, LSA and Coh-Metrix). This model predicted 49% of the variance in the hand-coded holistic explanation quality scores (EXPQUAL), $F(9, 167) = 18.11$, $MSE = .95$, $p < .001$. Only the number of codes that were detected by the ML approach (MLCODES), the two LSA scores (COPY and IDEAL), and the Coh-Metrix cohesion score (COH) were significant unique predictors.

The Coh-Metrix cohesion score (COH) seems to have come closer to capturing connections between sentences better than the connection measure derived from the ML approach (MLCONN) which failed to predict explanation quality. The lack of prediction by the number of connections derived from the ML approach (MLCONN) suggests there is still more work that needs to be done to automatically detect and identify relations within student arguments. In many cases, students used vague anaphoric references across sentences and explicit marking of rhetorical structure in earlier sentences. The humans were able to use both of these features to determine structural relationships in the hand-coding, but currently the ML approach is not able to use either of these sources of information to classify the structure and connections that may be present in student arguments. This also affected the ability to detect chains in the explanation essays.

Table 11.Hand-coded Holistic Explanation Quality Scores as predicted by all automated approaches

Variable	Unstandardized Beta (*B*)	Std. Error	Standardized Beta (β)	*t*-value	*p*-value
(Constant)	-.21	.91		-.23	.82
MLCODES	.28	.06	.34	4.54	<.001
MLCONN	-.02	.08	-.02	.30	.76
MLEASY	.04	.47	.01	.08	.93
MLCHAINS	.09	.09	.07	1.00	.32
COPY	-2.59	.40	-.39	-6.52	<.001
IDEAL	4.75	.97	.37	4.90	<.001
CAUSAL	-.00	.01	-.01	-.18	.86
COH	.17	.08	.13	2.09	<.04
LEXDIV	1.43	1.23	.08	1.17	.24

Note: MLCODES = Machine Learning estimated number of propositions, MLCONN = Machine Learning estimate for number of connections, MLEASY = Machine Learning detection of easy chain from text, MLCHAINS = Machine Language detection of chains to outcome, COPY = LSA estimate of similarity to original sources, IDEAL = LSA estimate of similarity to idealized peer explanation, CAUSAL = Coh-Metrix index of causal terms, COH = Coh-Metrix similarity between paragraphs, LEXDIV = Coh-Metrix Lexical Diversity

Table 12 shows the results of a simultaneous regression predicting student understanding using metrics derived from all automated approaches (ML, LSA and Coh-Metrix). The model shown in Table 12 was a good fit for the data, $F(8, 168) = 6.80$, $MSE = .02$, $p < .001$, and it predicted 25% of the variance in test scores. Of the new ML measures, the estimated number of propositions (MLCODES) was found to contribute unique variance.

Table 13 provides a summary of all the various approaches used to predict holistic essay quality scores (EXPQUAL) and student understanding as assessed by comprehension test scores (IVT). Although the measures derived from hand-coding of the students' arguments are the best predictors of the explanation quality score, and the measures derived from hand-coding of the *a priori* causal model are the best predictors of comprehension test scores, a combination of several automated scores provides relatively good prediction for both dimensions of student performance. The addition of ML metrics explains additional variance beyond LSA and Coh-Metrix measures for both outcomes.

Table 12. Comprehension Test Scores (IVT) as predicted by all automated approaches

Variable	Unstandardized Beta (*B*)	Std. Error	Standardized Beta (β)	*t*-value	*p*-value
(Constant)	.19	.11		.68	.50
MLCODES	.02	.01	.17	1.97	.05
MLCONN	-.01	.01	-.08	-.80	.42
MLEASY	-.02	.06	-.03	-.42	.68
MLCHAINS	.01	.01	.09	1.00	.32
IDEAL	.38	.11	.29	3.37	<.001
CAUSAL	.00	.00	.14	2.01	<.05
COH	.03	.01	.23	2.98	<.01
LEXDIV	.31	.15	.16	2.01	<.05

Note: MLCODES = Machine Learning estimated number of propositions, MLCONN = Machine Learning estimate for number of connections, MLEASY= Machine Learning detection of the easy chain, MLCHAINS= Machine Learning detection of other chains to outcome, IDEAL = LSA estimate of similarity to idealized peer explanation, CAUSAL = Coh-Metrix index of causal terms, COH = Coh-Metrix similarity between paragraphs, LEXDIV = Coh-Metrix Lexical Diversity

Table 13. Summary of variance explained (R^2) in explanation quality scores (EXPQUAL) and comprehension test scores (IVT) as predicted by different methods

	EXPQUAL	IVT
Causal Model Coding	.47	.33
Argument Structure Coding	.69	.20
LSA	.41	.16
Coh-Metrix	.08	.10
LSA and Coh-Metrix	.41	.23
Machine Learning, LSA, and Coh-Metrix	.49	.25

DISCUSSION

The results show promise in combining multiple automated methods as part of an attempt to approximate the success of hand-coding approaches in assessing the quality of student understanding from written explanations. This study focused on understanding of a single topic, using a single explanation essay to serve as an assessment of each student's understanding at a single time point, and using a comprehension test as the criterion measure. Using this approach, both hand-coding approaches achieved high inter-rater reliability and yielded scores on multiple dimensions that were predictive of performance on the comprehension test. The hand-coding method that was based on the *a priori* casual model of the phenomenon of global warming was the best single predictor of comprehension test performance. This makes sense when considering that the comprehension test itself was directly dependent on whether students constructed the mental models they needed to verify potential causal relations, while being less relevant for predicting whether students could or would be inclined to write essays in which they were sure to explicate how every causal relation ultimately impacts the target phenomenon. The coding approach based in the *a priori* model may be more sensitive to variation in understanding of causal relationships within various parts of the model, but it is also less sensitive to how well students can explicate in writing how all those relationships fit together, and articulate causal chains that ultimately lead to the outcome to be explained.

In addition, several measures were less important than expected, or than they seemed from simple correlations. Essay length did not predict essay quality or comprehension scores once more direct measures of coverage and connectedness were taken into account. Similarly, responsiveness to the essay prompt also no longer predicted performance after coverage and connectedness were included in regression models. These results suggest that both length and responsiveness to a prompt may in some cases serve as proxies for coverage or thematic focus within an explanation, but that actually scoring for the content and structure is a more powerful approach. That is, longer essays will not always be better explanations, just as longer summaries may be less focused and can contain irrelevant details (Wade-Stein & Kinstch, 2004).

Interestingly, the presence of citations or references to particular documents as part of student explanations was if anything a negative feature. Readers who were more likely to refer to the documents when writing about this science topic were less likely to focus on the most important information. This suggests that these students may have been engaging in a knowledge-telling approach of simply relating information from each source, as opposed to a knowledge-transforming approach in which they selected the most important information in an attempt to integrate it. While the presence of citations when writing from multiple documents in history is usually related to better quality essays (e.g., Britt & Aglinskas, 2002), this may be due to the important role of a source in evaluating historical documents which may be particularly needed when a document set includes opposing theories or discrepancies (Bråten, Strømsø, & Britt, 2009; Rouet et al., 1996). However, this was not the case in this particular activity. Under other circumstances, such as when contradictions are present between sources and need to be reconciled, sourcing may emerge as a more positive feature.

In sum, the results from the hand-coding approaches demonstrated that measures representing the coverage of key ideas in the essays, and the extent to which ideas are connected or integrated, were critical features for predicting understanding. As existing technologies, both LSA and Coh-Metrix were worth exploring next to determine how well they might be able to capture the quality of student understanding before investing substantial effort in machine learning approaches.

Our attempts to use out-of-the-box tools were to some extent successful, and metrics derived from LSA and Coh-Metrix were found to predict a significant amount of the variance for each outcome measure. As found in previous work (Ventura et al., 2014), computing similarity to an idealized peer essay with LSA provided a useful metric that predicted both hand coding and student understanding. The LSA-based plagiarism score was also useful. Copying scores were negatively correlated with overall explanation quality, and this relation became even stronger when similarity to the idealized peer essay was also included the regression model. The idealized peer explanation itself had a modest copying score, and the similarity-to-idealized-essay scores were positively correlated with copying scores in simple correlations. The fact that the relation between copying scores and explanation quality became negative once similarity-to-idealized-essay scores were added to the regression model suggests that the simple correlation might represent the tendency for students to copy the most task-relevant sentences from the texts. Thus, the negative correlation for copying scores that emerged in the regression that already included similarity-to-idealized-essay scores represents copying of less relevant sentences from the original sources. Although actively transforming information may be the best strategy for understanding, selecting and copying relevant information may be better than writing irrelevant information or failing to engage with the text at all. Also, since none of the documents specifically provided an answer for the essay question, even copying isolated sentences entailed some level of repurposing of information. In addition to these reasons, the low proportion of copied sentences in these essays (the average was only around 30%) may explain why the plagiarism scores did not have a negative relation with understanding in this study.

In contrast, the metrics derived from Coh-Metrix were the poorest predictors of comprehension test performance, and were only related weakly to the measures derived from hand-scoring. The causal metric also showed no relationship to how similar each essay was to the idealized peer explanation essay. It is notable that the idealized peer explanation essay was only about average in its causal Coh-Metrix score, which might be due to that metric giving credit for redundant or task-irrelevant causal terms in many essays. In addition, the standard generic causal terms used by Coh-Metrix may be unable to recognize topic-specific terms that reflected causal relations in this particular context (e.g., "CO_2 *traps* heat"). These issues may have obscured relations that may have been seen with a more topic-specific measure of causality.

Yet, even though the relations between the metrics derived from Coh-Metrix and student understanding were modest, they were still interesting to the extent that they provided a contrast to prior work that has used Coh-Metrix to explore the compositional quality of persuasive student essays. While prior work has found negative relations between cohesion and causal expressions and expert ratings of composition quality, in the present work these features were positive predictors of student understanding. Similarly, while prior work has suggested that lexical

diversity may be a positive predictor of expert ratings of composition quality in persuasive essays, in general it was a negative predictor of explanation quality in this study. Only once the coverage and connectedness of student explanations were taken into account in regression models did lexical diversity emerge as a positive predictor. One interpretation of these results is that lexical diversity may sometimes represent a third variable (such as student ability or verbal intelligence), and it may predict expert ratings of essay quality because more-able students may generally produce better essays for a wide variety of reasons. In the context of an explanation essay, however, using a more diverse set of words to describe a particular phenomenon may be a sign of a lack of focus on developing an integrated causal model of that phenomenon.

Finally, the best prediction of explanation quality from the automated measures was from a model that included machine-learning scores in addition to LSA and Coh-Metrix indices. The new machine-learning approach did a reasonable job of learning and applying the coding rules employed in the more structure-sensitive hand-coding system, and the machine learning scores added 8% to the total variance explained over and above the contributions of LSA and Coh-Metrix. The machine learning scores also improved the prediction of comprehension test performance over other automatic methods for detecting structure. These results provide for the key conclusions of this study, and point to the utility of using this machine learning approach in combination with LSA for detecting similarity of a response to an idealized response and similarity to original sources, and Coh-Metrix for detecting similarity, focus, and causality within a response. They suggest promise for hybrid methods combining those that are good for detecting content (similarity to sources, similarity to an idealized essay, ML concepts) and those that are good for detecting structure (similarity with idealized essay, cohesion, causality). It is possible that these methods may eventually be able to be applied within AES systems as well.

One reason why the benefits from the machine-learning approach were so modest in the present study may be because of the complexity of this particular global warming document set. In other studies, we have begun using simpler document sets for inquiry tasks on coral bleaching ("explain how and why coral bleaching rates vary at different times") and skin cancer ("explain how and why rates of skin cancer differ around the globe"). The inquiry prompts for these activities still require inferences across multiple documents, but both document sets are less complex than the global warming set. They have fewer and shorter documents, fewer initiating causes, and fewer and simpler elements. The causal model for the global warming text set could be viewed as very complex on all dimensions, while the two newer text sets are only moderately complex in that there are only 2 initiating causes and 10 key elements across 5 documents. For both topics, human and machine-learning scoring for explanation quality were found to be highly correlated (Hughes et al., 2015), and the machine learning approach was better able to predict student understanding of the coral bleaching and skin cancer units from student essays.

The current study focused on detecting student understanding of a single topic, by using a single explanation essay to serve as an assessment of each student's understanding at a single time point. However, in most cases, developing a coherent understanding of topic will require working through ideas, building and revising explanatory models, and constructing understanding iteratively over time. Such a process requires revision, and providing real-time, tailored feedback to students can facilitate and enable this process. The long-term goal for this

work is to enable near instantaneous calculation of what is included in student explanatory essays and what is missing, which would represent the basis for an intelligent tutor that could help students to improve the quality of their written explanations as well as their understanding of the subject matter. Because achieving this future goal requires the ability to provide detailed feedback about the quality of the reasoning that is present in explanations (such as whether they include explicit connections to the target outcome or to other initiating causes), an assessment of the structure of students' explanations is needed, which is what the present machine language approach attempted to capture by sorting essays into quality categories.

A recent study used a similar set of quality categories to give college students feedback on initial drafts of explanation essays written as part of the simpler coral bleaching inquiry unit (Kopp et al., 2016). After writing initial explanations, students were randomly assigned to either receive targeted feedback (in relation to the completeness or coherence of their essays as indicated by the quality categories) or no feedback about their drafts (students were simply asked to revise). The targeted feedback prompted students to create longer chains and to give more complete answers, and was intended to benefit those students who failed to include intervening elements or multiple initial causes. Overall, students included significantly more connected concepts in their explanations after revision. However, the targeted feedback condition particularly helped those whose initial essays were of poor quality. Receiving appropriate feedback helped them significantly improve their explanations and learn more from the activity.

These results are promising and such a multidisciplinary approach to providing feedback may eventually have utility in a classroom setting. With current calls in science education for students to learn about explanation and argumentation in science classes, and the increasing appreciation for writing-to-learn activities, an intelligent tutoring system that can give immediate feedback based on student understanding will be helpful to teachers in the classroom. Other areas of research have shown the importance of feedback and revision for student progress. For example, a meta-analysis on the effectiveness of feedback on quality of student compositions has shown moderate effect sizes (e.g., 0.77, Graham & Perin, 2007). Similarly, revision is essential to improving the quality of written compositions (Hayes & Flower, 1981). At present, however, much more work needs to be done to extend these findings from demonstrating the effectiveness of feedback and revision in a learning-to-write context to demonstrating the effectiveness of feedback and revision in a writing-to-learn context (i.e. as part of subject-matter learning).

Finally, even without the explicit metacognitive emphases of iSTART (McNamara et al., 2007) and MetaTutor (Lintean et al. 2011), it is hoped that a system that provides explicit feedback on specific weaknesses in student explanations will lead to more complete reasoning and better learning from multiple-document inquiry tasks, which in turn might transfer and support better performance in other writing-to-learn tasks (as in Britt, et al. 2004). There are many other types of writing activities that may be employed besides causal explanations or arguments (Braaten & Windschitl, 2011), and we believe our approach can be extended to try to detect student understanding from other types of open-ended responses such as problem-solution or compare-and-contrast essays. Given the differences that have been seen between the features that have predicted the quality of persuasive and explanatory essays, exploring detection of student understanding from different essay types will be an important step for future work.

References

Bejan, C. A., & Hathaway, C. (2007). UTD-SRL: A pipeline architecture for extracting frame semantic structures. *Proceedings of the 4th International Workshop on Semantic Evaluations* (pp. 460-463). Association for Computational Linguistics. Prague, Czech Republic.

Bennington, B.J., (2009). *The Carbon Cycle and Climate Change*. Retrieved from http://www.cengage.com/custom/enrichment_modules.bak/data/Carbon_Cycle_0495738557_LowRes.pdf

Braaten, M., Windschitl, M. (2011). Working toward a stronger conceptualization of scientific explanation for science education. *Science Education, 95*, 639-669.

Bråten, I., Strømsø, H.I., & Britt, M.A. (2009). Trust matters: Examining the role of source evaluation in students' construction of meaning within and across multiple texts. *Reading Research Quarterly, 44*, 6-28.

Britt, M.A., & Aglinskas, C. (2002). Improving student's ability to use source information. *Cognition and Instruction*, *20*, 485-522.

Britt, M.A., & Sommer, J. (2004). Facilitating textual integration with macro-structure focusing task. *Reading Psychology*, *25*, 313-339.

Britt, M.A., Wiemer-Hasting, P., Larson, A., & Perfetti, C.A. (2004). Automated feedback on source citation in essay writing. *International Journal of Artificial Intelligence in Education*, *14*, 359–374.

Chklovski, T., & Pantel, P. (2004). VerbOcean: Mining the web for fine-grained semantic verb relations. *Proceedings of the Conference of Empirical Methods in Natural Language Process* (pp. 33-40). Barcelona, Spain.

Condon, W. (2013). Large-scale assessment, locally-developed measures, and automated scoring of essays: Fishing for red herrings?. Assessing Writing, 18(1), 100-108.

Crossley, S. A., Kyle, K., & McNamara, D. S. (2015). The tool for the automatic analysis of text cohesion (TAACO): Automatic assessment of local, global, and text cohesion. *Behavior Research Methods*, 1-11.

Crossley, S. A., & McNamara, D. S. (2010). Cohesion, coherence, and expert evaluations of writing proficiency. *Proceedings of the 32nd annual conference of the Cognitive Science Society* (pp. 984-989).

Crossley, S. A., & McNamara, D. S. (2011). Text coherence and judgments of essay quality: Models of quality and coherence. *Proceedings of the 29th Annual Conference of the Cognitive Science Society* (pp. 1236-1241).

Dascalu, M., Stavarache, L. L., Dessus, P., Trausan-Matu, S., McNamara, D. S., & Bianco, M. (2015). Predicting comprehension from students' summaries. In *International Conference on Artificial Intelligence in Education*, Madrid (pp. 95-104). Springer International Publishing.

Deane, P. (2013). On the relation between automated essay scoring and modern views of the writing construct. Assessing Writing, 18(1), 7-24.

Dikli, S. (2006). Automated essay scoring. Turkish Online Journal of Distance Education, 7(1), 49-62.

Foltz, P.W., Britt, M. A., & Perfetti, C. A. (1996). Reasoning from multiple texts: An automatic analysis of readers' situation models. In G. W. Cottrell (Ed.) *Proceedings of the 18th Annual Cognitive Science Conference* (pp. 110-115). Mahwah, NJ: Lawrence Erlbaum Associates.

Graesser, A. C., Wiemer-Hastings, P., Wiemer-Hastings, K., Harter, D., Person, N. & the Tutoring Research Group (2000). Using latent semantic analysis to evaluate the contributions of students in AutoTutor. *Interactive Learning Environments*, 8, 129–147.

Graesser, A. C., Hu, X., & McNamara, D. S. (2005). Computerized learning environments that incorporate research in discourse psychology, cognitive science, and computational linguistics. In A. F. Healy (Ed.), *Experimental cognitive psychology and its applications: Festschrift in honor of Lyle Bourne, Walter Kintsch, and Thomas Landauer.* (pp. 183-194). Washington, DC: American Psychological Association.

Graesser, A. C., & McNamara, D. S. (2012). Automated analysis of essays and open-ended verbal responses. In H. Cooper & A.T. Panter (Eds.), *APA handbook of research methods in psychology* (pp. 307-325). Washington, DC: American Psychological Association.

Graesser, A. C., McNamara, D. S., & VanLehn, K. (2005). Scaffolding deep comprehension strategies through AutoTutor, and iSTART. *Educational Psychologist*, *40*, 225-234.

Greene, S. (1994). Students as authors in the study of history. In G. Leinhardt, I. Beck and C. Stainton (Eds.), *Teaching and learning in history* (pp. 137-170). Hillsdale, NJ: Erlbaum.

Griffin, T.D., Wiley, J., Britt, M.A., & Salas, C. (2012). The role of CLEAR thinking in learning science from multiple-document inquiry tasks. *International Electronic Journal of Elementary Education, 5*, 63-78.

Hastings, P., Hughes, S., Magliano, J. P., Goldman, S. R., & Lawless, K. (2012). Assessing the use of multiple sources in student essays. *Behavior Research Methods*, *44*, 622-633.

Hastings, P., Hughes, S., Britt, A., Blaum, D., & Wallace, P. (2014). Toward automatic inference of causal structure in student essays. In S. Trausan-Matu and K. Boyer (Eds.), *Proceedings of Intelligent Tutoring Systems 2014*, Honolulu, HI (pp. 266-271). Berlin: Springer.

Hastings, P. Hughes, S., Blaum, D., Wallace, P., & Britt, M.A. (2016). Stratified learning for reducing training set size. In A. Micarelli, J. Stamper, and K. Panourgia (Eds.). *Proceedings of Intelligent Tutoring Systems 2016*. Paper presented at the 13th International Conference, Zagreb (pp. 341-346). Berlin: Springer..

Hemmerich, J., & Wiley, J. (2002) Do argumentation tasks promote conceptual change about volcanoes? In W.D. Gray & C.D. Schunn (Eds.), *Proceedings of the Twenty-Fourth Annual Conference of the Cognitive Science Society* (pp. 453-458). Hillsdale, NJ: Erlbaum.

Hughes, S., Hastings, P., Britt, M. A., Wallace, P., & Blaum, D. (2015). Machine learning for holistic evaluation of scientific essays. In C. Conati, N. Heffernan, A. Mitrovic, M.F. Verdejo (Eds.), *Proceedings of Artificial Intelligence in Education 2015* (pp. 165-175). Berlin: Springer.

Jaeger, A. J., & Wiley, J. (2015). Reading an analogy can cause the illusion of comprehension. *Discourse Processes, 52*, 376-405.

Landauer, T. K., Foltz, P. W., & Laham, D. (1998). Introduction to latent semantic analysis. *Discourse Processes, 25*, 259-284.

Larkey, L. S., & Croft, W. B. (2003). A text categorization approach to automated essay grading. Automated Essay Scoring: A Cross-discipline Perspective: Mahwah, NJ, Lawrence Erlbaum.

Lintean, M., Rus, V., & Azevedo, R. (2011). Automatic detection of student mental models based on natural language student input during metacognitive skill training. *International Journal of Artificial Intelligence in Education*, *21*, 169-190.

Magliano, J. P., & Graesser, A. C. (2012). Computer-based assessment of student-constructed responses. *Behavior Research Methods*, *44*, 608-621.

McNamara, D. S., Boonthum, C., Levinstein, I. B., & Millis, K. (2007). Evaluating self-explanations in iSTART: Comparing word-based and LSA algorithms. In Landauer, T., D.S. McNamara, S. Dennis, & W. Kintsch (Eds.), *Handbook of latent semantic analysis* (pp. 227-241). Mahwah, NJ: Erlbaum.

McNamara, D. S., Crossley, S. A., Roscoe, R. D., Allen, L. K., & Dai, J. (2015). A hierarchical classification approach to automated essay scoring. *Assessing Writing, 23*, 35-59.

McNamara, D. S., Graesser, A. C., McCarthy, P., & Cai, Z. (2014). *Automated evaluation of text and discourse with Coh-Metrix*. Cambridge: Cambridge University Press.

McNamara, D. S., O'Reilly, T., Rowe, M., Boonthum, C., & Levinstein, I. B. (2007). iSTART: A web-based tutor that teaches self-explanation and metacognitive reading strategies. In D.S. McNamara (Ed.), *Reading comprehension strategies: Theories, interventions, and technologies* (pp. 397-421). Mahwah, NJ: Erlbaum.

Meyer, B. J. F. (1985). Prose analysis: Purposes, procedures, and problems. In B. K. Britton, & J. Black (Eds.), *Analyzing and understanding expository text* (pp. 11-64). Hillsdale, NJ: Erlbaum.

Mihalcea, R., & Csomai, A. (2005). Senselearner: Word sense disambiguation for all words in unrestricted text. *Proceedings of the ACL 2005 on Interactive Poster and Demonstration Sessions* (pp. 53-56). Association for Computational Linguistics. Ann Arbor, Michigan, USA.

Miller, G., (1995). WordNet: A lexical database for English. *Communications of the ACM,* 38, 39-41.

Page, E. B., (1994). Computer grading of student prose, using modern concepts and software. *Journal of Experimental Education, 62*, 127-142.

Pennebaker, J. W. (1993). Putting stress into words: Health, linguistic, and therapeutic implications. *Behaviour Research and Therapy, 31*, 539-548.

Perfetti, C.A., Britt, M.A., & Georgi, M.C. (1995). *Text-based learning and reasoning: Studies in history*. Hillsdale, NJ: Lawrence Erlbaum Associates.

Rink, B., Bejan, C.A., & Harabagiu, S. (2010). Learning textual graph patterns to detect causal event relations. In *Proceedings of the Twenty-Third International Florida Artificial Intelligence Research Society Conference (FLAIRS),* Daytona Beach, FL, USA. (pp. 265-270), Applied Natural Language Processing Track. AAAI Press..

Roscoe, R., Crossley, S. A., Snow, E. L., Varner, L. K., & McNamara, D. (2014). Writing quality, knowledge, and comprehension correlates of human and automated essay scoring. In *Proceedings of the 27th International Florida Artificial Intelligence Research Society Conference, FLAIRS 2014*, Pensacola, FL (pp. 393-398). AAAI Press.

Rouet, J.-F., Britt, M.A., Mason, R.A., & Perfetti, C.A. (1996). Using multiple sources of evidence to reason about history. *Journal of Educational Psychology, 88*, 478-493.

Rouet, J.-F., Favart, M., Britt, M. A., & Perfetti, C. A. (1997). Studying and using multiple documents in history: Effects of discipline expertise. *Cognition and Instruction*, *15*, 85-106.

Royer, J. M., Carlo, M. S., Dufresne, R., & Mestre, J. (1996). The assessment of levels of domain expertise while reading. *Cognition and Instruction, 14*, 373–408.

Rus, V., & Graesser, A. C. (2006). Deeper natural language processing for evaluating student answers in intelligent tutoring systems. In Y. Gil & R.J. Mooney (Eds.), *Proceedings of the American Association of Artificial Intelligence* (pp. 1495-1500). Menlo Park, CA: AAAI Press.

Rus, V., Lintean, M. C., Graesser, A. C., & McNamara, D. S. (2009). Assessing student paraphrases using lexical semantics and word weighting. In V. Dimitrova, R. Mizoguchi, B. du Boulay, & A. C., Graesser (Eds.), *Artificial intelligence in education: Building learning systems that care; From knowledge representation to affective modeling* (pp. 165-172). Amsterdam, The Netherlands: IOS Press.

Sanchez, C., & Wiley, J. (2006). An examination of the seductive details effect in terms of working memory capacity. *Memory & Cognition*, *34*, 344-355.

Sanchez, C., & Wiley, J. (2009). To scroll or not to scroll: Interactions of text presentation and working memory capacity. *Human Factors, 51*, 730-738.

Sanchez, C. A., & Wiley, J. (2010). Sex differences in science learning: Closing the gap through animations. *Learning and Individual Differences, 20,* 271-275.

Sanchez, C. A., & Wiley, J. (2014). The role of dynamic spatial ability in geoscience text comprehension. *Learning & Instruction, 31*, 33-45.

Scardamalia, M., & Bereiter, C. (1987). Knowledge telling and knowledge transforming in written composition. *Advances in Applied Psycholinguistics, 2*, 142-175.

Shermis, M. D., & Hamner, B. (2012, April). Contrasting state-of-the-art automated scoring of essays: Analysis. In Annual national council on measurement in education meeting (pp. 14-16).

Spivey, N. N. (1990). Transforming texts constructive processes in reading and writing. *Written Communication*, *7*, 256-287.

Steinhart, D.J. (2001). *Summary Street: An intelligent tutoring system for improving student writing through the use of latent semantic analysis* (Unpublished doctoral dissertation). University of Colorado, Boulder, CO.

Ventura, M. J., Franchescetti, D. R., Pennumatsa, P., Graesser, A. C., Hu, G. J. X., & Cai, Z. (2004). Combining computational models of short essay grading for conceptual physics problems. In Lester, J.C., Vicari, R.M., Paraguac¸u, F. (Eds.), *Proceedings of the Intelligent Tutoring Systems Conference* (pp. 423-431). Berlin: Springer.

Voss, J. F., & Wiley, J. (1997). Developing understanding while writing essays in history. *International Journal of Educational Research*, *27*, 255-265.

Voss, J. F., & Wiley, J. (2000). A case study of developing historical understanding via instruction: The importance of integrating text components and constructing arguments. In P. Stearns, S. Wineburg, and P. Seixas (Eds.), *Knowing, teaching and learning in history* (pp. 375-389). New York: NYU Press.

Wade-Stein, D., & Kintsch, E., (2004). Summary Street: Interactive computer support for writing. *Cognition and Instruction, 22*, 333-362.

Wiley, J. (2001). Supporting understanding through task and browser design. In J.D. Moore & K. Stenning (Eds.), *Proceedings of the Twenty-third Annual Conference of the Cognitive Science Society* (pp. 1136-1143). Hillsdale, NJ: Erlbaum.

Wiley, J., & Voss, J. F. (1996). The effects of "playing" historian on learning in history. *Applied Cognitive Psychology*, *10*, 63-72.

Wiley, J., & Voss, J. F. (1999). Constructing arguments from multiple sources: Tasks that promote understanding and not just memory for text. *Journal of Educational Psychology, 91,* 301-311.

Wiley, J., Ash, I.K., Sanchez, C.A., & Jaeger, A. (2011). Clarifying readers' goals for learning from expository science texts. In M. McCrudden, J. Magliano, & G. Schraw (Eds.), *Text relevance and learning from text* (pp. 353-374). Greenwich, CT: Information Age Publishing.

Wiley, J., Goldman, S., Graesser, A., Sanchez. C., Ash, I., & Hemmerich, J. (2009). Source evaluation, comprehension, and learning in internet science inquiry tasks. *American Educational Research Journal, 46*, 1060-1106.

Wiley, J., Steffens, B., Britt, M.A., & Griffin, T. D. (2014). Writing to learn from multiple-source inquiry activities in history. In G. Rijlaarsdam (Series Ed.) and P. Klein, P. Boscolo, C. Gelati, & L. Kilpatrick (Volume Eds.), *Studies in Writing, Writing as a Learning Activity*. (pp. 120-148). Leiden/Boston: Brill.

Appendix

IDEALIZED PEER EXPLANATION:
Recent patterns in global temperature are different from what has been observed in the past due to several factors which each influence each other. Although temperatures typically alternate between cooling and warming cycles, global temperatures have remained at the highest levels for much longer than in prior warming periods. The factors causing this are an increase in use of fossil fuels and deforestation, causing greater amounts of CO2 in the atmosphere, which increases the effects of the greenhouse effect, which melts ice caps and results in larger oceans, which in turn causes the earth to retain even more heat as a result. The beginning of this chain of consequences is an increase in the burning of fossil fuels.

In the late 1800s, people began to burn bigger amounts of fossil fuels for energy. CO2 is produced when fossil fuels are burned. Because an increase in burning of fossil fuels means an increase in CO2 production, as a result there is more CO2 in the atmosphere today than there ever has been. Also contributing to the rising levels of CO2 in the atmosphere is the destruction of forests and swamps, which absorb and store much more carbon than farmlands. Over the past hundred years, half of the forests of the world have been destroyed. There is a direct relationship between the destruction of forests and the amount of CO2 in the atmosphere because when forests are destroyed, the carbon still has to go somewhere. The increased amounts of CO2 in the atmosphere intensify the greenhouse effect. Plus, as our population increases and as swamps and forests are converted to cities and animal farms, the increase in people and livestock also causes an increase in CO2 emissions.

All of the energy on earth originally comes from the sun. The greenhouse gases, such as CO2, trap heat in the atmosphere rather than letting it escape back into space. So, the more greenhouse gases there are, the warmer the Earth becomes. The increase in CO2 means more heat will be trapped, resulting in an increase in global temperatures. Warmer temperatures melt the ice caps and frozen ground which releases the CO2 stored there for thousands of years. This creates a cycle that makes the problem get worse and worse. The CO2 from fossil fuels winds up releasing even more CO2 from other sources into the atmosphere. Also, less ice means more area on the surface of the earth that absorbs energy from the sun, and less energy that gets reflected back into space. The absorbed energy is emitted is as heat and remains in our atmosphere due to greenhouse gases.

The recent increase in global temperature is different and greater than the increases observed in the past due to a chain of events beginning with increased dependencies on fossil fuels and ending in a stronger greenhouse effect, resulting in higher temperatures.