Summary: Basic Studies in Science and History

Project READI Technical Report #7

M. Anne Britt, Jennifer Wiley and
Thomas Griffin

PROJECT READi

inquirium

Northern Illinois University

NORTHWESTERN UNIVERSITY

UIC UNIVERSITY OF ILLINOIS AT CHICAGO

WestEd

Citation for this Report: Britt, M.A., Wiley, J., & Griffin, T. (2016). *Summary: Basic studies in science and history*. READI Technical Report #7. Retrieved from URL: projectreadi.org

The authors would like to acknowledge the contributions of the following members of the Project READI basic studies team for their work on this report: Dylan Blaum, Srikanth Dandotkar, Sarah Davis, Peter Hastings, Karyn Higgs, Simon Hughes, Allison Jaeger, Kris Kopp, Mike Mensink, Kathryn Rupp, Carolos Salas, Brent Steffens, Andrew Taylor, and Patty Wallace

Please send us comments, questions, etc.: info.projectreadi@gmail.com

Project READI was supported by the *Reading for Understanding (RFU)* initiative of the Institute for Education Sciences, U. S. Department of Education through Grant R305F100007 to the University of Illinois at Chicago from July 1, 2010 – June 30, 2016. The opinions expressed are those of the authors and do not represent views of the Institute or the U. S. Department of Education.

Project READI operated as a multi-institution collaboration among the Learning Sciences Research Institute, University of Illinois at Chicago; Northern Illinois University; Northwestern University; WestEd's Strategic Literacy Initiative; and Inquirium, LLC. Project READI developed and researched interventions in collaboration with classroom teachers that were designed to improve reading comprehension through argumentation from multiple sources in literature, history, and the sciences appropriate for adolescent learners. Curriculum materials in the READI modules were developed based on enacted instruction and are intended as case examples of the READI approach to deep and meaningful disciplinary literacy and learning.

Overview

The basic studies in science and history defined evidence-based argument as the generation of causal models for natural phenomena or historical events based on comprehension of text sets that consisted of multiple documents. Much of the experimental work was aimed at elucidating characteristics of tasks/task instructions and text sets that were associated with more versus less complete causal models in each discipline. Tasks and task instructions were explored through variations in the prompts provided for reading and writing tasks. Text sets varied in terms of types of documents and modes of information. In all cases, the text sets were constructed so that information relevant to a complete and coherent causal model was distributed across the set of documents, requiring analysis and synthesis within and across multiple as well as individual texts. In addition, to explore the relationship between performance on the comprehension and production of causal models from multiple texts, epistemic cognition surveys in science and history were developed based on existing instruments. These were expanded to tap into beliefs about the value and importance of multiple texts in constructing knowledge in history and science.

First Reporting Period (Year ONE Annual Report, August 2010 - February 2011)

During the initial reporting period, document sets were developed for basic studies on three topics: the causes of Global Temperature Change, the Scopes Trial, and the Panama Revolution. In all three cases, the document sets were designed to require integration across documents in order to achieve an understanding of the topic. The document sets were designed to include information necessary to construct a coherent causal model of the target event or phenomenon. Each document contributed only part of the overall causal model with some overlap between the sources. At least one document presented information visually such as in a map, graph, or political cartoon, and students were also provided with necessary background information in a document. Materials were piloted asking students to read in order to answer a "how or why" question in order to write an argument about the causes of a historical event or scientific phenomenon.

Second Reporting Period (Year TWO Annual Report, March 2011- February 2012)

During the second reporting period, materials continued to be developed including the document sets, the reading/writing prompts that students were given, and the outcome measures that were used to assess comprehension in relation to the coverage and understanding of the causal model underlying the event or phenomena that was the focus of each document set.

In the Spring of 2011, data samples were primarily from students in 9th through 12th grades. In the Fall of 2011, data samples were primarily from students in 6th, 7th and 8th grades. Initial studies indicated that late middle school students could engage at least minimally in these multiple-document comprehension activities while 6th grade students struggled. The initial studies also suggested that students struggled to differentiate between the reading to recall tasks with which they were familiar and the kind of task they were being asked to do in these studies. In short, performance indicated that there was much room for improvement at all grade levels along several dimensions, including what the task demands were (e.g., integrating across documents, constructing a response rather than finding "the answer" in a verbatim sentence or phrase in the text. Initial data also suggested that students might benefit from directions that discouraged creating a simple narrative, listing, or opinion essay; and encouraged writing an interpretation using ideas and evidence from the documents to support their reasoning. Students also needed to be encouraged to think about relations between ideas, and chains of factors, networks, or systems as being responsible for phenomena or events. Based in these initial observations, follow-up studies were run over the next several years of the grant exploring variations in the task instructions, i.e., in

the reading/writing prompts, text sets, and the use of pre-writing scaffolding activities, to understand how different instructional supports might support more effective multiple source comprehension.

Variations in the reading/writing prompts and documents included in the global temperature change activity were explored in follow-up studies run in later years, using the addition of a policy-related prompt (encouraging students to think about what can be done about climate change), and manipulating the presence of policy-related documents (such as one based on Gore's suggestions). The results suggested that including policy-related documents can lead to poorer learning outcomes, with evidence that students focused upon policy in lieu of, rather than in relation to, a causal understanding of the causes of climate change. This work, which includes some data collected during the second reporting period, is reported in Blaum, Wiley, Britt and Griffin (in press).

Variations in the reading/writing prompts were also explored for the Panama Revolution activity. Performance was better when students were prompted to write an essay "explaining the factors that caused the Panamanian Revolution of 1903" than when they were asked to write an "argument about the extent to which U.S. President Theodore Roosevelt and his administration were responsible for bringing about the Panamanian Revolution." Essays also improved when students were explicitly prompted to include evidence to support their claims. In addition, in follow-up studies we manipulated whether students engaged in a pre-writing activity of constructing a timeline to support their understanding and integration across documents. This work, which includes data collected during the second reporting period, is reported in Wiley, Steffens, Britt and Griffin (2014).

Variations in the reading/writing prompts were also explored for the Scopes Trial activity in a series of follow-up studies run across several years of the grant. In addition, we manipulated whether students engaged in a pre-writing notetaker activity to support their understanding of the possible factors and integration of information across documents. Students were asked to fill in a chart of "what was happening" at the time of the Scopes Trial which prompted them to consider a range of social, cultural, economic, and demographic changes. The notetaker activity supported more complete causal models and understanding. This work, which includes data collected in the second reporting period, is reported in Wiley, Griffin, Taylor, Jaeger, and Britt, (2014), currently a working paper as a manuscript reporting this work is in preparation.

Also during this year, a parallel study was attempted for the Global Temperature Change activity. Along with the set of documents, students were given an unlabeled schematic of the Global Climate System, a chart to fill out to document the relationships seen in the schematic, and where the information for each explanation came from in the documents. One sample of seventh graders got the activity on the first day (N=40) while a second sample did not (N=19). All students got a writing prompt that instructed them to write an essay about the causes of changes in global temperatures, using evidence from the documents to support their ideas, emphasizing that a response would have to be constructed from information across documents. The diagram activity led to more sourcing (35% vs. 11%) but did not change the coverage of concepts in the essays (3.58 vs. 3.62). This result was used to inform the design of more effective activities that were developed and tested in later reporting periods.

In an attempt to capture which students may come to these tasks with a better appreciation of the value of using multiple sources of evidence, or a better understanding of how to use multiple sources to understand a topic, we began development of instruments to assess intellectual values and epistemological beliefs about the use of multiple documents, as individual differences measures that may be critical for successful engagement in multiple-document comprehension tasks. This strand of activity was based on prior work (Braten, Britt, Stromso & Rouet, 2010; Griffin, 2008; Lee & Ashby, 1996; Voss & Wiley, 1998) as well as the formulation during Year Two of READI learning goals that included demonstration of an understanding of the nature of science or history and how science knowledge or historical claims are generated. The learning goals emerged as critical in focusing the design of the interventions and thus the development of surveys of epistemological beliefs was an important connection between the basic studies and the design of the classroom interventions. One individual differences measure (CLEAR thinking) was developed to measure students' appreciation of evidence. Performance on this scale was found to be significantly related to learning outcomes on the science unit. These results are reported in Griffin, Wiley, Britt and Salas (2012). Initial versions of epistemological belief surveys were piloted during the second reporting period (with parallel items in history and science). Relations with learning outcomes were found for some items, and the data suggested a general increase (greater sophistication) in the epistemology scores from middle school to high school grades. These data were used to refine later versions of the scales piloted in subsequent years.

Third Reporting Period (Year THREE Annual Report, March 2012- February 2013)
In the Spring of 2012 and Fall 2012, data samples were only obtained from students in 6th, 7th and 8th grades. Students at the 6th grade again struggled to engage in the multiple document comprehension activities, as did special education students.

During the third reporting period, some data were collected for follow-up studies exploring variations in the reading/writing prompts for the Panama Revolution activity that are reported in Wiley, Steffens, Britt and Griffin (2014), and for follow-up studies on notetakers for the Scopes activity that are reported in Wiley, Griffin, Taylor, Jaeger, and Britt, (2014).

In addition, during the third reporting period, two studies tested notetakers for the Panama Revolution activity. Neither attempt was successful. In one study with 6th graders, one group of students completed a structured chart with information about actors, motives, and actions, as well as which source document mentioned that information and when it occurred. Compared to a condition that simply prompted students to engage in close reading of the documents, no benefits were seen, and generally performance was minimal. In a second study, we modified the pre-writing chart to focus only on the key motives and actions of agents. Using a sample of 7th and 8th graders, we found no differences in essay quality due to which pre-reading activity they were given.

During Year Three, we began a new strand of studies that attempted to address two consistent shortcomings in students' task performance: absence of integration of information across documents and lack of complexity the causal models. For the global temperature science unit, we created a short (10 min) lesson using a causal process familiar to students (digestion) yet unrelated to the inquiry domain (climate change) to illustrate the concept of causal chains (i.e. distal causes and mediation) as well as qualifying or contextual factors that can influence a process (i.e., moderators) and thus are part of the causal explanation. The results showed positive effects of the causal chain instruction on learning outcomes. This initial study is reported in Wiley, Griffin,

Taylor, Jaeger, and Britt, (2014), currently a working paper as a manuscript reporting this work is in preparation.

In two studies, we also tested the effectiveness of a causal chain instruction in a computer-based tutorial format. This tutorial included both explicit instruction and practice on the reading of single document passages providing explanations of natural or man-made processes (Rupp, Blaum, Wallace, & Britt, 2014; Rupp, Wallace, Blaum, & Britt, 2015). College students within an intact class were randomly assigned to either use the tutorial as homework or read for class (control). We found that students given the causal chain tutorial recalled significantly more elements of the explanation than students in the control condition. They also identified more elements in explanations when the texts were available. This pair of studies suggests that part of the difficulty students have in comprehending scientific explanations is due to not having appropriate goals for representing explanations or strategies for achieving those goals (i.e., an inappropriate Task Model). This appears to be true regardless of whether they are reading from single texts or multiple texts or constructing models from memory or with the text available. This finding aligns with the earlier findings from the studies with adolescent students that showed the benefits of providing reading/writing prompts that were explicit about what the task required.

A third study was begun on the Scopes Unit. The causality tutorial was modified to be about the multiple interacting causes of the American Revolution (instead of digestion). Although a small amount of data was collected in mid-February 2012, most was collected during the fourth reporting period, so it is discussed in more detail in that section.

In addition, during Spring 2012 Burkett, Goldman and Britt (2014) conducted a study to investigate middle school and high school students' recognition of contradictions between multiple representations (text and graph) of predictions about scientific phenomena. Specifically, we examined whether students recognized when data presented in a graph contradicted or corroborated a prediction that could be made from information presented in an accompanying text. The text either explicitly stated the prediction or the prediction had to be inferred from presented information. The main finding was that older students were more likely to discriminate between consistent (agree) and inconsistent (disagree) cases, especially when the relationship was explicitly stated. The finding that 8th graders were less likely to indicate use of the explicit statement of the relationship was somewhat surprising since the presence of the statement improved their performance. It appears that while this explicit statement has an impact on performance for all groups, only the older groups are sensitive to the role it played. This work is reported in Burkett, Goldman, and Britt (2014).

Finally, the samples used for basic studies were also used to develop and refine the epistemology measures. While data from initial versions of these scales suggested a lack of reliability and validity, scores on both the second and third iterations of the epistemology inventories showed some convergent validity with teacher rated student reading skill and other thinking dispositions such as a general tendency to value evidence, and some predictive validity with learning. Scores still generally improved with grade level. However, several items were still problematic. A few of the existing items were removed and revised, and some additional items were added for further testing in the following year. Because the sample during this reporting period included only middle school students, another sample of high school students was also needed before the tests could be finalized and validated.

Fourth Reporting Period (Year FOUR Annual Report, March 2013- February 2014)
In the Spring of 2013, basic studies were run using data samples from both grade bands: 9th, through 11th grades, and 6th, 7th and 8th grades. In addition, epistemology scales were collected from classrooms of a subset of teachers who participated in the Chicago Teacher network.

During this reporting period, some data were collected for follow-up studies exploring variations in the reading/writing prompts and documents included in the global temperature change activity The results suggested that including policy-related documents can lead to poorer learning outcomes, with evidence that students focused upon policy in lieu of, rather than in relation to, a causal understanding of the causes of climate change. This work, which includes some data collected during earlier reporting periods, is reported in Blaum, Wiley, Britt and Griffin (in press).

Additionally, a study to test the effectiveness of a causality tutorial on the complexity of students' causal models in history was conducted during this period. However, there were a number of pragmatic complications that undercut definitive results. Nonetheless, the results suggested that the causal tutorial was a promising support for students. More specifically, in the Spring of 2013, we developed a short lesson on causality in history using the example of the American Revolution that was delivered before students engaged in the Scopes trial activity (basically, explain how and why the Scopes trial occurred when it did). In one high school (School A), there were exceptionally low consent rates, preventing analysis of much of the collected data. At the second school (School B), one teacher's four 11th grade history classes participated but they could not be randomly assigned to conditions. Unfortunately for condition comparisons, two-thirds of the students in the control classrooms were above average students compared to only one-third in the treatment classrooms. Importantly, a multivariate analysis of variance (MANOVA) including all 3 learning measures (Essay quality, Inference-Verification-Test, and Multiple-choice) revealed that expected course performance was a strong predictor of learning ($F(3, 58) = 6.45$, $p < .01$), accounting for about 15% of the variance. The disproportionate number of better-performing students in the control classrooms worked against findings hypothesized benefits of the causal chain instruction treatment. Indeed, a second multivariate analysis of variance including all 3 learning measures showed no learning differences between treatment and control conditions ($F(3, 58) = 2.05$, $p > .10$). Due to the non-random assignment of classrooms, control classrooms would have been expected to learn more in the Scopes unit. The fact that the causal-chain classrooms performed as well as the control classrooms suggests that the treatment instruction may have in fact been beneficial in helping those students learn more than they otherwise would have been predicted to. Although publishing the null results of this study alone would be difficult, and follow-up work would be needed, these findings do indicate the causal chain instruction is promising for improving learning in history just as it was in science for the Global Temperature Change unit.

Additional data were also collected from 6th grade samples. Because all our previous work with sixth graders had suggested that they had great difficulty engaging in multiple document inquiry tasks, several attempts were made to try to understand what aspects of multiple document inquiry tasks might be most difficult for middle school students. First, we tested whether the multiple document context itself was partially responsible for low performance. We varied whether sixth-graders received the Global Temperature Change activity as a single document or as multiple documents (simplified to contain only 4 texts but the same 3 graphs as our standard set). The results suggested that learning was neither harmed nor helped by presenting information in a single versus multiple document format. Second, we investigated whether individual differences in

spatial ability might affect learning from this activity. We added a new graph comprehension activity, and added a short assessment of spatial ability at the end. The graph comprehension worksheet instructed students to identify important features of the climate change graphs included with the text. Performance on the graph comprehension worksheet significantly predicted learning, but spatial ability did not. Third, we attempted to support learning with an analogy for the greenhouse effect. The analogy did not improve learning, but when students received the analogy they gave inflated judgments of learning and overestimated their understanding of the text. This work is reported in Jaeger and Wiley (2015).

In addition to the data collected from Project READI middle school and high school populations, convenience samples of undergraduates participated in several studies exploring the promise of several types of interventions. One study was part of a Masters project exploring the effects of "taking a side" vs. "understanding the perspectives" prompts on learning in history (Steffens, 2013). The results suggested that "take-a-side" prompts are not more engaging and do not lead to better learning from a multiple-document comprehension activity than other types of prompts. Students given the "take-a-side" prompt included the same number of key motives and actions as students given the "understand the perspectives" prompt. Unlike students given the "take-a-side" prompt however, the students given the "understand the perspectives" prompt included more distal causes than proximal causes in their essay and had better memory for distal key motives and actions than proximal motives and actions. These findings suggest that the "understand the perspectives" prompt was leading students to focus on critical contextual factors of the event, whereas the "take-a-side" prompt did not. This research is reported in the Masters thesis (Steffens, 2013), and a manuscript is in preparation.

A similar study tested whether asking students to make a judgment on a historical controversy (i.e. explain the extent to which the use of children during the Birmingham Children's Crusade was justified) would be more engaging and therefore lead to better learning compared to asking students to read to understand the causes of a historical event. In addition, some students were also asked to write down how each document contributed to their response (contribution task). Participants read 6 short texts including a background document and several other types of sources (e.g., newspaper article, textbook, letter from a Civil Rights leader) about the Birmingham Children's Crusade of 1963. (This study also served as pilot data for development of an evidence-based assessment task and text set for potential use in a Year 5 randomized control trial.) A significant interaction for the reading prompt and contribution task was seen on memory for content, $F(1, 96) = 6.34$, $MSE = 342.28$, $p = .013$, $\eta^2 = .011$. When the contribution task was absent, participants reading to make a judgment ($M = 5.12$, $SD = 2.34$) provided more correct responses than participants reading to understand the causes ($M = 3.83$, $SD = 1.43$). When the contribution task was present, no differences were found between the judgment and ($M = 3.58$, $SD = 1.86$) and causal prompts ($M = 3.83$, $SD = 1.43$). However, participants given the causal prompt ($M = 2.76$, $SD = 1.15$) included more causes in their essays than participants given the judgment prompt ($M = 2.26$, $SD = 1.19$). The results suggest that a judgment reading task may improve memory for the textual content, but at the expense of learning the causal model. Portions of this work were presented in 2014 at the Society for Text & Discourse (Steffens, Britt, & Manderino, 2014). We have recently collected additional data. We are currently analyzing the data and will write a manuscript in the coming year.

A third study was run as part of a dissertation (Kopp, 2014). The goal of this study was to investigate the amount of instruction that might be needed to improve source-evaluation skills in

science. We manipulated the amount of instruction provided to students before they engaged in a multiple document argumentation task about global warming. In the *Argument Tutor* condition, participants were given a short instructional intervention that provided examples, instructions, definitions, and opportunities to practice with feedback regarding how to assess source information. In the *Argument Definition* condition students were only given the definitions part of the original tutorial. Finally, a third condition (*Control*) did not provide any instruction but provided baseline performance. In addition, half of the documents that students received were from sources that were clearly "flawed" while half were from credible sources (i.e., experts) and published in credible outlets. Students in the Argument Tutorial condition selected significantly more reliable documents to read than the Argument Definition and Control conditions. In addition, only participants in the Tutorial condition incorporated significantly more facts from reliable documents compared to unreliable documents. These results suggest that students have impoverished source evaluation skills. This research is reported in Kopp, Britt, Millis and Rouet, 2014, and a manuscript based on this work is in preparation (Kopp, Britt, Millis and Rouet, in preparation).

In addition to the studies of tasks and text sets, we continued work on the epistemology scales. Data from the sample of classrooms drawn from teachers in the Chicago Teacher Network (pretest) were used to inform the design of the epistemology scale. These data suggested that two subscales capture unique variance from each other, have sound psychometric properties, and show convergent and divergent validity as evidenced by their differing relations to the various types of learning measures and other individual difference measures provided by teacher ratings. However, inadvertently, the scales were altered from 6 – point to 5 – point scales due to a misunderstanding. These made it necessary to collect additional data for purposes of demonstrating the reliability of the survey using 6-point scales.

A main area of focus during this year was developing various ways of coding the quality of the corpus of essays collected as part of the basic studies (Jaeger, Griffin, Wiley, Britt & Blaum, 2015), This was in conjunction with developing automated scoring algorithms for the essays (Hastings, Hughes, Britt, Wallace & Blaum, 2014; 2015). The set of essays collected from the Global Temperature Change unit were fully hand-coded for their coverage of the concepts contained in the a priori causal model. In several passes, we identified causal language, connections among concepts, and relational markers by hand, and also used LSA and LIWC to assess aspects of essay coherence and cohesion (Jaeger, Griffin, Wiley, Britt & Blaum, 2015). We also hand-coded explicit connections being made between concepts to identify the length and number of causal chains being constructed by the reader as part of their explanations for the target outcome (e.g., increased global temperature, Hastings, Hughes, Britt, Wallace & Blaum, 2014; 2015). Inter-rater reliability was high (Kappa = 0.85), and the scoring was useful in discriminating essays based on both completeness as well as on coherence. The coding indicated that 38% of the concepts and 66% of other supporting statements were not explicitly connected within essays, showing that many student essays lacked explicit coherence. Many essays consisted of a listing of key sentences taken from the documents with very little transformation or integration. Further, even when students made explicit connections among ideas in their essays, they often constructed only one causal chain. Several participants who did list more than one causal chain, listed the initial and intervening factors each as individually causing the to-be- explained outcome without considering inter-relations or connections among the intermediate factors (e.g., One reason for increases in global temperatures is increased fossil fuel use. Another reason is an increase in greenhouse gases in the atmosphere.)

Additional computational work was devoted to developing machine learning and natural language processing techniques for evaluating students' argumentative essays. To infer the causal structure of student essays, we used a standard parser to derive grammatical dependencies in the essays and converted them to logic statements (Hastings, Hughes, Britt, Blaum, & Wallace, 2014). Then a simple inference mechanism was used to identify concepts linked to syntactic connectors by these dependencies. The results suggest that we will soon be able to provide explicit feedback that enables teachers and students to improve comprehension. Toward that end, to improve machine detection accuracy, we tried using a two-phase machine learning approach for detecting causal relations (Hughes, Hastings, Britt, Wallace, & Blaum, 2015). For each core essay concept, we initially trained a window-based tagging model to predict which individual words belonged to that concept. Using the predictions from this first set of models, we then trained a second stacked model on all the predicted word tags present in a sentence to predict inferences between essay concepts. The results were promising and indicate the potential for such a system to provide explicit feedback to students to improve reasoning and essay writing skills.

In addition, we are examining methods of detecting categories of essay quality (Hughes, et al., 2015). Approximately 1,000 high-school students completed two reading-to-write activities that had the purpose of learning from multiple documents about the causes of two scientific phenomena (before and after an intervention). Humans evaluated the essays for four hierarchical levels of quality: (1) No core content (irrelevant or vague), (2) No causal chains (mentioned elements without connections), (3) Causal chain(s) without intervening (elements directly linked to the outcome), (4) Chain(s) with intervening (successfully included intervening elements). Inter-rater reliability for human scoring was high (Kappas of .87 and .93). Automatic scoring used window-based tagging models trained to label each word with an element code, one model per code. The maximum and minimum probabilities assigned by each model per word were computed across each entire sentence, and fed into another logistic regression model using a form of 'stacked generalization'. This model was trained to predict the elements and causal relations at the sentence level. Heuristics were used to compute metrics from these predictions to assess the quality. Human and automatic scoring for quality were highly correlated for both topics (r = .64, r = .59) for pretest performance. We are currently evaluating automated techniques to detect change within an individual that can then be used to provide feedback to students (e.g., attend to directional modifiers, encourage chaining). We are currently writing a working paper that will be submitted to AIED in early April and this will serve as the working paper in addition to the publications.

Fifth Reporting Period (Year FIVE Annual Report, March 2014- February 2015)
During this reporting, there was no funding or support for data collection in basic studies in middle school or high school age bands. Some data was collected using college students as a convenience sample In addition, epistemology scales were collected from 9th grade classrooms involved in the science randomized control trial study. Extensive coding, statistical analysis, and writing were done for several lines of basic studies research which resulted in additional coding techniques to assess the quality of student essays, new approaches in the automated coding systems for the essays, and manuscripts representing data that was collected over several years of the grant (Blaum, Wiley, Britt & Griffin, in press; Hastings, Hughes, Britt, Wallace & Blaum, 2014; 2015; Jaeger & Wiley, 2015; Wiley, Steffens, Britt & Griffin, 2014).

One new line of research, begun in Fall 2014, examined whether prompting learners to think of solutions to a scientific problem, instead of prompting them to understand the causes, is beneficial

to learning the causal mechanisms of the phenomenon. To test this, we manipulated both the causal prompt (present vs. absent) and the solution prompt (present vs. absent). This design gave students one of four possible reading prompts: Causal only ("…explain how and why global climate change is occurring"), Solution only ("…understand what we can do about global climate change"), Causal and Solution ("…explain how and why global climate change is occurring and understand what we can do about it"), and Comprehension ("…comprehend the information in the documents about global climate change"). Solution prompts led to better recall of core causal concepts (M = 16.4%, SD = .15) than causal prompts (M = 11.9%, SD = .12), $F(1, 74) = 3.96$, $p = .050$, $d = .33$. Also, when asked how a particular change (hybrid cars) could be an effective solution to the problem, solution prompts led to higher quality solution responses (M = 45%, SD = .22) than causal prompts (M = 35%, SD = .24), $F(1, 74) = 4.45$, $p = .038$, $d = 0.43$. These data are reported in Blaum, Britt, Platt, Clark, Griffin, and Wiley (2015) and follow-up studies are being pursued as part of a Masters thesis.

In addition, we attempted to finally establish reliability of the epistemology scales using a 6-point Likert scale for all items with data collected as part of pretest data for the RCT studies in Fall 2014. These analyses show that two subscales capture unique variance from each other, have sound psychometric properties, and show convergent and divergent validity as evidenced by their similar relations across disciplines and grade levels, yet differing relations to the various types of learning measures and other individual difference measures. This research is reported in Salas, Griffin, Wiley, Britt, Blaum and Wallace (2015) which serves as a working paper as a manuscript of this work is in preparation.

A main area of focus during this year continued to be developing various ways of coding quality for the corpus of essays collected as part of the basic studies (Jaeger, Griffin, Wiley, Britt & Blaum, 2015), in addition to developing automated scoring algorithms (Hastings, Hughes, Britt, Wallace & Blaum, 2014; 2015). Building on work done in Year 4, analyses indicated several important measures of essay quality including number of sentences, coverage of the model, number of connections among concepts, cohesion among sentences, proportion of sentences borrowed, and proportion of explanation elements mentioned. Improvements were also made in methods of capturing structure (e.g., average number of distinct explanations connected, longest explanation, mentioning at least one intervening cause, and mentioned both unique causes) and assigning essays to categories of explanation quality: (1) No core content, (2) No causal chains, (3) Causal chain with no intervening factors, (4) Chain with intervening. These additional quality codes informed a new approach for the automated scoring systems. The new approach involved using a novel, two-phase machine learning approach for detecting causal relations (Hastings, Hughes, Britt, Wallace & Blaum, 2015). For each core essay concept, we initially trained a window-based tagging model to predict which individual words belonged to that concept. Using the predictions from this first set of models, we then trained a second stacked model on all the predicted word tags present in a sentence to predict inferences between essay concepts. The refinements improved the automated scoring algorithms. These related lines of work providing insights into quality of products that students generate when comprehending multiple documents from both hand-scoring and automated scoring were presented at an NSF-sponsored workshop (Multidisciplinary Advances in Reading and Writing for Science Education) in May 2015, and are currently written up and accepted in combination (Wiley, Hastings, Blaum, Jaeger, Hughes, Wallace, Griffin & Britt, 2016) for publication in a special issue of AI-ED.