

Annual Report for Period:01/2003 - 01/2004**Submitted on:** 01/05/2004**Principal Investigator:** Wiley, Jennifer .**Award ID:** 0126265**Organization:** U of Illinois Chicago**Title:**

ROLE Proposal: Understanding in Science: Eyetracking Studies

Project Participants

Senior Personnel

Name: Wiley, Jennifer**Worked for more than 160 Hours:** Yes**Contribution to Project:**

Dr. Jennifer Wiley is directing the Dual-Purkinje eyetracking studies, and also led the development and design of the web sites on volcanic eruptions that are the content for the experiments.

Name: Graesser, Arthur**Worked for more than 160 Hours:** Yes**Contribution to Project:**

Dr. Arthur Graesser is directing the combined think-aloud/eyetracking condition of the project in Memphis. He will also be in charge of developing a tutor to support critical reading skills that are identified in the first strand of the project.

Name: Goldman, Susan**Worked for more than 160 Hours:** Yes**Contribution to Project:**

Dr. Susan R. Goldman is directing the UIC think-aloud studies on the project, and developing coding systems for the think-aloud data that is collected.

Post-doc

Name: Moreno, Kris**Worked for more than 160 Hours:** No**Contribution to Project:**

She supervised the components of the experiment other than eye tracking at Memphis. This includes the materials and procedure. She was partially funded on the grant, but not quite 160 hours. She has a background in social cognition.

Graduate Student

Name: Ash, Ivan**Worked for more than 160 Hours:** Yes**Contribution to Project:**

Ivan was the lead graduate student on the Strand 1 Dual Purkinje Eyetracker studies. He is fully trained on eyetracking procedures and analysis, and is currently developing C programs to automatize some elements of data reduction. He also supervised the remainder of the student personnel during 2002-3 under the direction of Dr. Jennifer Wiley.

Name: Sanchez, Christopher**Worked for more than 160 Hours:** No**Contribution to Project:**

Chris helped design the web site during summer 2002. His skills are in Web page design and evaluation of different media formats on learning. He will also be primarily responsible for running the eyetracking evaluation study of the agent once it is developed in 2004.

Name: Colflesh, Greg**Worked for more than 160 Hours:** Yes**Contribution to Project:**

Greg helped design the web pages and helped with the running of eyetracking experiments in 2002.

Name: Hemmerich, Joshua

Worked for more than 160 Hours: Yes

Contribution to Project:

Josh helped with literature searches, materials development, and running some pilot think-aloud studies in Spring and Summer of 2002.

Name: Braasch, Jason

Worked for more than 160 Hours: Yes

Contribution to Project:

Jason is in charge of running and coding the think-aloud protocols under the direction of Dr. Susan Goldman.

Name: Cauty, Reality

Worked for more than 160 Hours: No

Contribution to Project:

Reality helped with the design and content of the web site, as well assisting with running pilot Think-Aloud subjects and coding the protocols. As an undergraduate, Reality was funded by the UIC Summer Research Opportunities Program for Minorities, which provides undergraduate research internships in UIC laboratories each summer.

As a graduate student, he is funded by a diversity fellowship.

Name: Whitten, Shannon

Worked for more than 160 Hours: Yes

Contribution to Project:

Shannon supervised the course of data collection and statistical analysis of the data at Memphis. Her background is cognitive psychology and discourse processing.

Name: Mitchell, Heather

Worked for more than 160 Hours: No

Contribution to Project:

Heather is the new supervisor of all phases of the project at Memphis, including eye tracking, material, procedure, and statistical analyses of the data. She is funded on the grant, but less than 160 hours.

Name: Rokadiya, Tyrun

Worked for more than 160 Hours: No

Contribution to Project:

He is writing programs to perform data mining of the eye tracking data. He is funded by the grant, but less than 160hours.

Name: Joon, Moonjee

Worked for more than 160 Hours: Yes

Contribution to Project:

Part of Memphis Team

Undergraduate Student

Name: Brodowinska, Kamila

Worked for more than 160 Hours: Yes

Contribution to Project:

Kamila is assisting Jason with running think-aloud participants under the direction of Dr. Susan Goldman.

Name: Bland, Melissa

Worked for more than 160 Hours: No

Contribution to Project:

Missy assisted with data collection, entry and organization on the eyetracking portion of the study as an undergraduate during the 2002-2003 school year. She is now a graduate student in School Psychology at Ball State University.

Name: Cooper, Elisa

Worked for more than 160 Hours: Yes

Contribution to Project:

She collected eye tracking data and assisted in the plate tectonics study at Memphis.

Name: Edwards, Michelle

Worked for more than 160 Hours: Yes

Contribution to Project:

She collected eye tracking data and conducted statistical analyses at Memphis.

Name: Haynes, Bryan

Worked for more than 160 Hours: Yes

Contribution to Project:

He programmed the computer for collecting eye tracking data and organizing data files at Memphis.

Name: Petschonek, Sarah

Worked for more than 160 Hours: Yes

Contribution to Project:

She collected eye tracking data and conducted statistical analyses at Memphis.

Name: Treier, Nycole

Worked for more than 160 Hours: Yes

Contribution to Project:

She collected eye tracking data and conducted statistical analyses at Memphis.

Name: Wagner, Wende

Worked for more than 160 Hours: Yes

Contribution to Project:

She collected eye tracking data and conducted statistical analyses at Memphis.

Name: Wallace, James

Worked for more than 160 Hours: Yes

Contribution to Project:

She collected eye tracking data and conducted statistical analyses at Memphis.

Name: Moher, Jeff

Worked for more than 160 Hours: Yes

Contribution to Project:

Worked with Jennifer Wiley during Summer 2003 on eyetracking data coding and cleaning.

Name: McDaniel, Bethany

Worked for more than 160 Hours: No

Contribution to Project:

Part of Memphis Team

Technician, Programmer

Other Participant

Name: Jolly, Cara

Worked for more than 160 Hours: Yes

Contribution to Project:

Cara helped with running pilot experiments, as well as transcription, data coding, entry and analysis for pilot and eyetracking studies under the direction of Dr. Jennifer Wiley from January 2002 to August 2003. She was a recent graduate of UIC with a BS in Psychology who is now at NYU.

Name: Jensen, Melinda

Worked for more than 160 Hours: Yes

Contribution to Project:

Melinda Jensen is a recent graduate with an Honors BS in Psychology from Carleton College. She is gaining research experience before applying to graduate school. She joined the Wiley Lab in October 2003 is currently in charge of running and analyzing the ROLE project.

Name: Gepstein, Rona

Worked for more than 160 Hours: Yes

Contribution to Project:

Assists Dr. Susan Goldman

Name: O'Reilly, Tenaha

Worked for more than 160 Hours: Yes

Contribution to Project:

Tenaha is the head of the research team working in the Graesser Lab.

Research Experience for Undergraduates

Organizational Partners

Other Collaborators or Contacts

We have arranged a symposium at AERA on 'Science literacy:

The Centrality of Taking a Critical Stance' in which we will discuss our research with Clark Chinn, Sarah Brem, Rick Duschl, and Nancy Songer. We look forward to making contact with these researchers who are doing work most closely aligned to our project.

The tutor is also likely to share some similarity with 'Sourcer's Apprentice' a scaffold developed by Anne Britt (along with Gareth Gabrys and Chuck Perfetti). Anne has been a valuable contact in terms of thinking about formats for the tutor.

The project also seems to have much in common with Roger Azevedo's work on explanation and tutoring, and it was very interesting to meet and exchange ideas with him at the PI Meeting.

Activities and Findings

Research and Education Activities:

ROLE Understanding in Science (Wiley, Goldman, Graesser)

Year 2 Research Activity Report

The overall goals of this project are to understand how and when students adopt a critical stance toward scientific information, and how such a stance can be supported. To this end, we proposed two strands of research that focus on how young adult learners interact with science content obtained from the internet. Our goal in the first strand is to examine when and how readers spontaneously adopt a critical stance toward scientific information, and in particular, the relationship between learning and the adoption of a critical stance. In the second strand of research, we will investigate how the use of pre-training exercises and an intelligent conversational agent can assist learners in gleaning understanding of scientific information from internet sites.

The goals for the first year of work included the development of the scientific content on internet sites, learning assessments, and standardized experimental procedures, which involved several pilot studies.

Initial Web Sites/Learning Assessment Development Project

Developing the web sites for this project required extensive internet and literature searches, as well as the re-analysis of some previously collected data from UIC undergraduates. Originally 10 sites were selected for the first pilot studies. These sites were selected from the output of a Google search using the key words 'causes volcanic eruptions' with the constraint in mind that the set of sites needed to reflect a range of reliable information about volcanoes. Based on a literature review and analysis of data previously collected on UIC undergraduates' understanding of volcanoes, we identified various categories of conceptual models and misconceptions related to volcanic activity. A correct causal model was also outlined. Web sites were selected so that at least one included a reference to each of the important steps in a correct

causal model of volcanic activity. Several web sites that made reference to inaccurate 'causes' were also included so that the sources were varied in their accuracy as well as their reliability. The reliable sites contained accurate information and provided converging evidence for each other (i.e. the USGS site, the NASA site and the PBS sites all provided information that could be integrated into the same coherent causal model of volcanic activity). The unreliable, inaccurate sites all provided unique causal information that could not be integrated into the model suggested by the reliable sites. True/False inference questions were developed for a pre-test and post-test based on the correct causal model as well as common misconceptions.

Reliability Pilot Study

A pilot study in which students were asked to judge the reliability of 10 internet sites revealed that students in general do not seem to have a common or stable understanding of what makes some sources more 'reliable' than others. Although they seemed to recognize the 'right' reasons on the rating task, there was a large discrepancy between explicit ratings of their basis for reliability judgments and what they talked about during the think-aloud phase. There was also variability across students.

In practice, the interestingness or usefulness of the site seemed to be the basis for most evaluative activity. A second reliability study run this year showed that high knowledge students are not any better than low content knowledge students at justifying reliability ratings. This suggests that one potential area that we should address is helping students develop a more scientifically-based understanding of reliability, and supporting the use of reliability judgments as a basis for deciding which sites to attend to. This will likely be part of a pre-training module that will instruct students on the concept of reliability.

Think Aloud Pilot Study

A think-aloud pilot study indicated that most students would prefer longer than a half hour for the research task, while an hour seems like a natural length of time for most participants. Early results with the evaluation task suggested that it was not yielding research behaviors similar to the argument and description writing tasks, so it was eliminated from the proposed design of Strand 1. The reading time, ranking, and learning results replicated and extended what we found in the reliability pilot study. Students ranked the 'reliable' sources as more reliable than the unreliable sources. They also tended to spend more time reading reliable sources, and that related to better learning. These are exactly the patterns we were hoping to be able to investigate further in Strand 1, using think-aloud and eyetracking methodology to assess when and how students adopt a critical stance toward scientific information they encounter on websites.

Strand 1 Proposed Studies

These experiments were begun in the Fall 2002 semester and completed in Spring/Summer 2003. Using the web sites, materials and procedures developed through piloting, we began our three-part approach to investigating when and how students develop a critical stance toward scientific information. Three parallel studies were run using the same exact materials and design. Study 1 measured reading behavior as students performed their research with a Dual-Purkinje eyetracker (N=29), Study 2 asked students to think-aloud as they performed their research (N=40), and Study 3 combined both methodologies using a head-mounted eyetracker as well as think-aloud instructions (N=28).

All participants were told that their task was to figure out what caused the eruption of Mt. St. Helens from the output of a GOOGLE search using the keywords 'causes volcanic eruptions.' They were given the first page of those results, with 7 links. Of the 7 sources, 3 were meant to be reliable, 3 were intended as unreliable and 1 was reliable but ambiguous. Students were told that their purpose for reading is to either write a Descriptive Report or an Argument of what caused the eruption of Mt. St. Helens. The order of the sites was counterbalanced across two versions of the Google page, yielding a 2x2 design for these studies. They were given about an hour for their research, and about a half hour to write their essays. After essay writing, all students completed learning posttests, ranking/rating reliability surveys for each web site, evaluated a peer essay with an inaccurate causal model, and completed demographic and computer use/familiarity surveys. All students were given a pretest as part of mass testing.

In the second year, data collection and analysis was completed for the first strand of studies as proposed, as well as several additional studies. The coding and analysis of all open-ended tasks: the essays, peer essay evaluations, ranking justifications and think-aloud protocols all required the development of extensive coding schemes. Eyetracking analyses required the agreement on Regions of Interest (ROIs) between the two eyetracking labs, and first pass and total time analyses were run for all readers. Coding schemes were developed for navigation patterns as well. As a result, data analysis was highly time intensive, but yielded rich data of what better and poorer learners did in this learning task.

Supplemental Studies

In addition to our proposed studies and the pilot studies that were performed to refine the materials and procedures, several additional studies have been performed that will inform our future work on this grant. The first study (Evaluation Pilot) concerns a new 'Evaluation' condition to try to prompt students into a critical stance. A second study ran High Knowledge students through the learning tasks to investigate whether prior knowledge was related to better reliability judgments. Two additional think-aloud studies have been run, one on non-native English speakers (Non-native Think aloud), and one on high knowledge students (High knowledge Think Aloud). The fifth study (Agent Interface Pilot) is relevant to developing an effective tutor interface in the later phases of our research.

Evaluation Pilot

The purpose of this study was to test if expecting to evaluate another student's essay will have an effect on learning. Students were run under the same conditions as the Strand 1 studies, expecting to write either an argument or a description. However in addition to the writing instruction, half the participants in each condition were warned that they would need to evaluate another student's essay based on their research. No think-aloud or eyetracking data were collected. 24 undergraduates at UIC participated in this study (6 in each condition). Analyses on learning gains and reliability rankings suggest that the evaluation warning seems to make everyone in the Argument condition perform uniformly better than students in the Descriptive condition (i.e. the warning tightens up the variance in the Argument condition). An explicit warning that students may need to evaluate a peer essay AFTER they write an argument may help students be more critical/selective as they read, and may be an instruction to consider as part of the intervention in Strand 2. Based on preliminary results, an explicit prompt to evaluate another essay appears to be more effective than simply the instruction to write an evaluation (as proposed).

High Knowledge Learning Study

24 Subjects who scored above 19 on the pretest to test if prior knowledge is needed for better judgments of reliability. Reliability judgments, and amounts of time spent reading reliable and unreliable information were similar for this group and the low knowledge groups run in Strand 1, suggesting that prior knowledge is not strongly related to reliability judgments. This suggests that pre-training on reliability may be effective and that prior knowledge may not be necessary or responsible for adopting a critical stance.

Non-native English Speaker Think-Aloud Study

16 non-native English speakers were run through the same condition as the Strand 1 think-aloud study to investigate how these students learn science from web-based sources. These protocols are currently being transcribed and the other dependent measures analyzed.

High Knowledge Think-Aloud Study

The purpose of this study was to ask high knowledge students to identify and describe the strategies, advice, and guidance they would give to less knowledgeable students; in order to generate potential ideas for dialogue moves that could be made by the conversational agent in Strand 2. Suggestions included: spend time on pages that explain the process of volcanoes; look for explicit, expert authorship and valid sources; doubt the credibility of information when an author has an agenda; look for pages that scientists or teachers would agree with; and look for connections across sites that seem reliable. These kinds of advice are potential prompts that could be provided through the agent.

Agent Interface Pilot Study

Year 3 of the grant will be developing a conversational agent that assists the student in taking a critical stance while reading the web pages. The question arises as to the type of agent that we should use. Animated conversational agents have facial features synchronized with speech and in some cases appropriate gestures (Graesser et al., 2001; Johnson, Rickel, & Lester, 2000). A pilot study demonstrated that an animated agent is clearly a strong magnet of attention, and this effect does not decrease over time. This result raises important questions about the sort of tutor we should have in assisting the student in to take a critical stance. Perhaps voice alone would be better than a talking head. Alternatively, an animated conversational agent may occur at the beginning of the session while instructions are given, whereas voice alone will be used when college students interact with the websites.

Findings: (See PDF version submitted by PI at the end of the report)

The major findings from our Strand 1 studies are included in the attached file.

The conclusions that can be reached from our first strand of studies are that better learning seems to be related to the ability to discriminate between reliable and unreliable sources, the selection of reliable reading material for a high percentage of research time, the detection and rejection of unreliable sites soon after they are selected, and an approach to learning that emphasizes explanatory processes. At the same time, even the best learners in our Strand 1 studies seemed to have a fragile understanding of what reliability is. Very few could verbalize it or use it to justify their evaluations.

In the next phase of our research program, we will be developing supports for learners that will build on these observations. Using both pre-task training in the concept of reliability, as well as an on-line interactive conversational agent that will prompt readers to evaluate sources, we will endeavor to support more critical and explanation-based processing among all learners in the second phase of this study.

Training and Development:

The most notable training opportunity on this project is that it involves two highly skill-intensive methodologies: eyetracking and think-aloud interviews. In order to start collecting data on this project, several graduate students and a dozen undergraduates have been trained in eyetracking and/or think-aloud methodology and analyses. Both methods require attention to detail so that

calibration and prompting are standardized across subjects and across sites. Hence, rigorous training in these procedures is imperative.

The combined think-aloud/eyetracking conditions conducted at the Memphis site presents a major opportunity for learning how to coordinate these two methodologies. The project staff at both sites have needed to take a proactive approach to figuring out how these methodologies can be best combined.

For undergraduates and new research assistants, additional training was given in the practical aspects of running experimental sessions. This includes the rudiments of experimental design (importance of counterbalancing, random assignment, coding blind, etc.), data entry, data coding, and file manipulation, as well as in getting basic distributions and statistical analyses in SPSS. All project staff were also required to attend an IRB training for new investigators.

Graduate students are being trained in these skills themselves, but are also responsible for playing a role in the training of undergraduates. The development of the web sources at the start of this project required the training of several graduate and undergraduate students in web page design and formatting; while the development of the learning assessment required extensive research into prior studies on earth science instruction, some studies examining learning from scientific text, and identification of the misconceptions that students hold about volcanoes. Finally, the need for data reduction from navigation and eyetracking logs has required several students to acquire programming skills, as they develop programs to streamline the re-formatting of data so that it can be analyzed more immediately.

To complete the analysis of the Strand 1 studies, students received detailed training in coding open-ended data such as protocols and essays, using rubrics developed in several reiterations.

Outreach Activities:

The eye tracking equipment and AutoTutor is being featured as a major image for the new FedEx Technology Institute. The Institute is now housed in a new building on the University of Memphis campus. A picture has been taken with someone using the eye tracking equipment and looking at a page with an animated agent. This picture will occur in the Memphis International Airport and other locations in the city.

There was a presentation by Art Graesser to the public in the city of Memphis on the eye tracking research. This occurred in the Fogelman Executive Center on October 25, 2002.

There was an interview with Dr. Art Graesser by the *Newsday's Sunday Journal* in New York State (December 8, 2002). The interview addressed the importance of taking a critical stance while students read web pages because there is a large volume of incorrect information on the web.

Journal Publications

Graesser, A.C., Person, N., & Hu, X., "Improving comprehension through discourse processes.", *New Directions in Teaching and Learning*, p. 33-44, vol. 89, (2002). Published

Graesser, A.C., & Olde, B.A., "How does one know whether a person understands a device? The quality of the questions the person asks when the device breaks down.", *Journal of Educational Psychology*, p. 524-536, vol. 95, (2003). Published

Graesser, A.C., Hu, X., Person, P., Jackson, T., and Toth, J., "Modules and information retrieval facilities of the Human Use Regulatory Affairs Advisor (HURAA).", *International Journal on eLearning*, p. , vol. , (). Accepted

Books or Other One-time Publications

Hemmerich, J. & Wiley, J., "Do argumentation tasks promote conceptual change about volcanoes?", (2002). Conference Proceedings, Published

Editor(s): W. G. Gray and C.D. Schunn (Eds.)

Collection: Proceedings of the Twenty-Fourth Annual Conference of the Cognitive Science Society.

Bibliography: Hillsdale, NJ: Erlbaum.

Wiley, J. & Bailey, J., "Effects of Collaboration and Argumentation on Learning from Web Pages.", (). Book Chapter, Accepted
 Editor(s): A. O'Donnell & C. Hmelo
 Collection: Collaborative Learning, Reasoning, And Technology
 Bibliography: Hillsdale, NJ: Erlbaum.

Wiley, J., Goldman, S.R., & Graesser, A.C., "Promoting critical inquiry from web sources.", (2002). Conference Proceedings, Published
 Editor(s): W. G. Gray and C.D. Schunn (Eds.)
 Collection: Proceedings of the 24rd Annual Conference of the Cognitive Science Society
 Bibliography: Mahwah, NJ: Erlbaum.

Otero, J., Leon, J.A., & Graesser, A.C., "The psychology of science text comprehension", (2002). Book, Published
 Bibliography: Mahwah, NJ: Erlbaum.

Graesser, A.C., Gernsbacher, M.A., & Goldman, S., "Handbook of Discourse Processes", (2003). Book, Published
 Bibliography: Mahwah, NJ: Erlbaum.

Graesser, A.C., McNamara, D.S., & Louwse, M.M., "What do readers need to learn in order to process coherence relations in narrative and expository text.", (2003). Book Chapter, Published
 Editor(s): A.P. Sweet and C.E. Snow
 Collection: Rethinking reading comprehension
 Bibliography: New York: Guilford Publications

DiPaolo, R.E., Graesser, A.C., Hacker, D.J., White, H.A., & TRG, "Hints in human and computer tutoring.", (). Book Chapter, Accepted
 Editor(s): M. Rabinowitz
 Collection: The impact of media on technology of instruction.
 Bibliography: Mahwah, NJ: Erlbaum

Goldman, S. R. & Wiley, J., "Discourse Analysis: Written Text.", (). Book Chapter, Accepted
 Editor(s): N. Duke & M. Malette
 Collection: Literacy Research Methods
 Bibliography: New York: Guilford.

Graesser, A.C., Gernsbacher, M.A., & Goldman, S.R., "Introduction to the Handbook of Discourse Processes.", (2003). Book Chapter, Published
 Editor(s): A.C. Graesser, M.A. Gernsbacher, and S.R. Goldman
 Collection: Handbook of discourse processes
 Bibliography: Mahwah, NJ: Erlbaum

Graesser, A.C., Hu, X., & McNamara, D.S., "Computerized learning environments that incorporate research in discourse psychology, cognitive science, and computational linguistics.", (). Book Chapter, Accepted
 Editor(s): A.F. Healy
 Collection: Experimental Cognitive Psychology and its Applications: Festschrift in Honor of Lyle Bourne, Walter Kintsch, and Thomas Landauer.
 Bibliography: Washington, D.C.: American Psychological Association.

Graesser, A.C., Leon, J.A., & Otero, J.C., "Introduction to the psychology of science text comprehension.", (2002). Book Chapter, Published
 Editor(s): J. Otero, J.A. Leon, & A.C. Graesser
 Collection: The psychology of science text comprehension
 Bibliography: Mahwah, NJ: Erlbaum.

Web/Internet Site

Other Specific Products**Contributions****Contributions within Discipline:**

One of the major contributions of this line of research is the development of research techniques to examine both verbal and eyemovement data while participants read web sites. We are exploring and perfecting a methodology for collection and analysis of simultaneous think aloud and eye tracking protocols.

A second major contribution is the description of student research behaviors as they attempt to learn from multiple sources. The vast majority of research on text processing has examined learning from single texts. The complexities of how students integrate across multiple texts must be examined in order to understand how learning from inquiry tasks may be best supported.

Further, our work has uncovered an important and crucial weakness in students' information literacy skills. Students suffer from a general lack of understanding about how to evaluate the reliability of sources that they encounter as a result of internet searches. Most research on student use of the World Wide Web has focussed in the information seeking process. Our research suggests that there is a critical need for skills in information evaluation that may need to be taught before effective learning can take place in activities requiring internet use.

Contributions to Other Disciplines:**Contributions to Human Resource Development:****Contributions to Resources for Research and Education:****Contributions Beyond Science and Engineering:****Special Requirements**

Special reporting requirements: None

Change in Objectives or Scope: None

Unobligated funds: less than 20 percent of current funds

Animal, Human Subjects, Biohazards: None

Categories for which nothing is reported:

Organizational Partners

Any Web/Internet Site

Any Product

Contributions: To Any Other Disciplines

Contributions: To Any Human Resource Development

Contributions: To Any Resources for Research and Education

Contributions: To Any Beyond Science and Engineering

Findings from Wiley, Goldman, Graesser ROLE Strand 1 Studies:

Learning Gains:

There was a significant, but small positive gain of 1.5 points from pretest to posttest for all subjects ($t(96)=4.87$, $p<.0001$).

Writing Task Manipulation: Argument vs. Description

No significant differences were found in learning outcomes (either post-tests or essay measures) due to the writing task instruction ($t_s < 1$). As opposed to earlier studies (Wiley & Voss, 1998; Wiley 1999), the instruction to write an argument did not lead to better learning outcomes than the instruction to write a descriptive report. Two major differences between these studies and earlier ones are that 1) the texts provided in these studies included unreliable information and 2) the think-aloud and eyetracking procedures were both run in individual settings, whereas previous studies were run in groups. One possibility is that an argument writing task might not be conducive to better understanding when readers are given a mix of reliable and unreliable sources. However, a group study run simultaneously with these materials replicated previous learning advantages (around 3.5 points on the post-test) for the argument writing condition (versus less than one point (.83) in the descriptive report condition, $t(22)=1.88$, $p<.07$). The small overall learning gains in the present studies suggest that the advantage of argument writing on learning may have been reduced by the individualized procedures of eyetracking and think-aloud protocol collection.

Recognition of Reliability, Reading Behaviors and Learning

Although the writing instruction manipulation had no effect on learning, and in general there was not a large positive learning gain for all subjects in this study, there was still a great deal of variability in the sample in terms of learning outcomes. To investigate how readers learn best from an internet research task such as this one, we more closely examined the relationship of learning to reliability judgments and reading behaviors using the entire sample of 96.

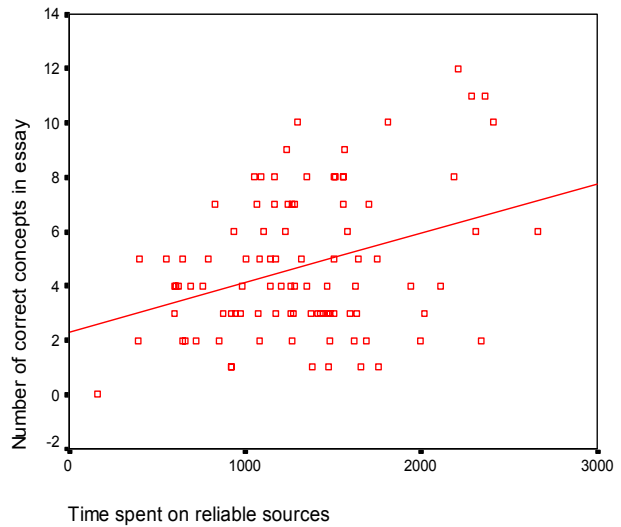
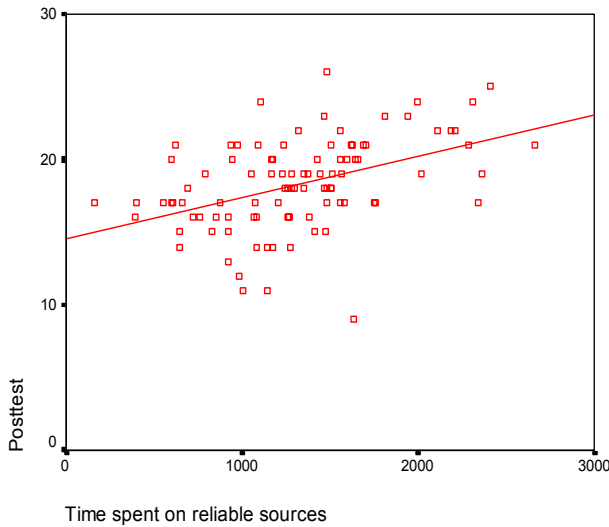
In general, students did make accurate evaluations of reliability. They ranked the reliable sites as more reliable ($M=4.75$) than the unreliable sites ($M=3.25$), $t(96)=8.4$, $p<.01$. (Reliability was assessed on a 1-7 scale with higher numbers meaning more reliable).

Overall, students also spent more time reading the 3 reliable sites (21.9 minutes) than the 3 unreliable sites (12.9 minutes), $t(96)=7.93$, $p<.0001$. (Similar data were seen for number of visits to sites and depth of navigation into sites).

As presented in the correlation table below, there was a three-way relationship between the amount of time spent on reliable sources, the reliability rankings for sources, and learning outcomes.

The more time students spent on reliable sources, the higher they ranked those sources in reliability ($r=.20$), the better they did on the post-test ($r=.40$), and the more they included

correct concepts in their essays ($r=.34$), and fewer misconceptions ($r=-.23$). (All correlations are significant at the $p<.05$ level.)



N=96	Reliability Rankings	Posttest Scores (pretest residualized)	Correct Conceptions in Essays	Misconceptions In Essays
Reading Time on Reliable Sites	.20*	.40**	.34**	-.23*
Reliability Rankings		.27**	.39**	-.25*
Posttest Scores			.29**	-.20*
Correct Conceptions				-.15

* $p<.05$
 ** $p<.01$

In general, better learning and development of a correct causal model of volcanic eruptions seemed to accompany the ability to discriminate between reliable and unreliable sources, as well as the selection of reliable reading material for a larger proportion of the research time.

Comparison of Better and Poorer Learners

To further investigate the behaviors that led to the best learning in this task, each lab examined its best and poorest learners in more detail. The good learners had gains on correct volcano concepts with no increase in misconceptions. The poor learners had

either no gain or a loss in correct volcano concepts with either no decrease or an increase in misconceptions.

Results from Eyetracking analyses:

In essence, the eye tracking data were highly compatible with the reading times that were gleaned from navigation logs for the reliable and unreliable sites. For example, among the 8 good learners and 7 poor learners from the 28 participants at the University of Memphis, eye tracking data supported the conclusion that good learners had a bias towards inspecting correct, reliable information whereas the poor learners drifted more towards the misconceptions and unreliable information. Regions of interest (ROI's) were identified for both reliable and unreliable information. The total amount of time on reliable ROI's was significantly higher for good learners than bad learners, 10.8 versus 6.7 minutes, $t(13) = 2.04$, $p < .05$ (one-tailed); there was the opposite trend for unreliable ROI's, 3.1 versus 4.3 minutes, but the difference was not significantly different.

Interestingly, better learners did not avoid unreliable information altogether. However, they did seem to react to it differently than poorer learners. An in-depth analysis of navigation and reading patterns was performed on the UIC eyetracking sample. All readers were likely to read the first site on the Google output list first, but better learners were more likely to deviate from serial order. When students deviated from serial order, they seemed to choose links based on keyword match (looking specifically for the words Mount St. Helens) or topic match (the phrase "Cause of Volcanoes" in the link title). Once taking a link, better learners detected unreliable sites quickly (in 30 seconds or less). Poorer learners who chose unreliable sites took longer to reject it than the good learners who chose the same site. This suggests that prompting the reader to consider the reliability of the source they are reading about 30 seconds into the source could be a useful support for some students.

Results from Think-aloud protocols

Think aloud protocols were analyzed by segmenting the content into 9 different think aloud categories and counting the instances of each category. The verbal protocols of the 30 best and worst learners in the Memphis and UIC Think Aloud studies were coded (15 at each site). In general, students in the UIC Think Aloud study made more comments overall ($F(1,26)=33.5$, $MSE=2294$, $p<.0001$), and better learners tended to make more comments than poorer students ($F(1,26)=5.68$, $MSE=2294$, $p<.03$), but there were no interactions in any category.

Think Aloud Categories

Examples of each category are provided along with the text immediately preceding the TA comment. Text that was read from the website is in *italics* and the coded comment example follows.

Paraphrase- *All volcanoes are born when hot magma rises to the surface, infiltrates a weak spot or opening in the Earth's outer crust, and erupts through.*

"That explains how the magma goes through the earth's crust... And it explains how it starts to erupt."

Self-explanations - *Lava temperature is one of several indicators that volcanologists regularly monitor in hopes of forecasting major eruptions.*

“I would think somehow that getting the pressure underneath rather than the magma... just coming from the writing, I would think that more the pressure would determine whether it’s going to explode rather than the temperature.”

Evaluation - *Let’s carefully examine the edges of colliding plates to see why volcanoes might be found there.* “Okay.. this definitely looks like I need to use this information”

Monitoring - *Who would have thought that the moon had that kind of power, not only to be able to cause the world’s oceans to bulge, but also to squeeze terra firma twice a day? But it does, so it should not come as a complete shock that reputable scientists have suggested that these squeezings may influence whether a volcano will erupt or not.* “I don’t know what they mean by oceans bulging... that doesn’t make sense to me but”

Monitoring/Evaluation - *Outlying cities were covered in flowing mud and a drift of gray ash, and very little was ever found of anyone or anything near ground zero.* “I don’t know... I don’t know if that’s really significant.”

Irrelevant Associations - *In the early morning hours of Sunday, May 18, 1980,* “Hmmm... I wasn’t even born”

Surface Connections - *Volcanoes tend to cluster along mountain belts. The folding and fracturing of plates that creates mountains, breaks up solid rock and creates channels for magma.* “Umm.. that’s the same thing as the second paragraph”

Prediction - *What causes the plates to move?* “There is gonna be an answer to this question”

Navigation – “I am clicking next... Umm ...I am scanning through and I am trying to find more about Mt. St. Helens”

Results of Think-Aloud Coding

The frequency of comments observed in each category are presented in the table below. When good versus poor readers were compared, only the self explanation category yielded a significant difference ($F(1,26)=8.56$, $MSE=272.42$, $p<.01$). Good learners tended to self-explain more than poor learners. The difference in navigation comments also approached significance, with good learners making more comments on their navigation ($F(1,26)=3.11$, $MSE=259.89$, $p<.09$).

More refined analyses of the think-aloud data are still needed. In particular, analyses that look at what part of which sites the comments refer to would indicate whether good learners are processing reliable content in ways that poor learners are not. An obvious next step in the analysis of the TA codes is to look at them for reliable versus unreliable sites. However, these results suggest that better learners are more focused on the task of

constructing an explanation of what causes volcanic eruptions. Supporting this behavior in all students, either through an agent or pre-task instruction, could be useful.

Frequency of Think-Aloud Codes

	Good learners	Poor learners	F(1,26)	p value
Total Codes	163.95	122.19	5.68	0.025*
Repeat/Paraphrase	29.05	23.02	1.23	0.27
Self-Explanation	39.19	25.12	8.59	0.001*
Evaluation	26.5	24.0	0.23	0.64
Monitoring	20.75	16.19	1.10	0.30
Monitoring/Evaluation	0.80	.64	0.16	0.69
Irrelevant Association	2.76	3.64	0.33	0.57
Surface Connection	2.08	1.38	1.29	0.27
Prediction	2.07	1.48	1.35	0.26
Navigation	40.8	30.3	3.11	0.09

Results of Other Open-ended Measures of Evaluation and Judgment

In addition to the above measures, two other open-ended measures were examined: comments on a peer essay and justification of reliability rankings for each site.

Peer Essay Evaluation

A peer essay which included a misconception (that oil drilling causes volcanoes) was presented to learners. Learners were asked to comment on the quality of the essay. Reactions to the quality of the peer essays were coded into 5 categories: Explicit acceptance of oil drilling as a cause of volcanoes, implicit acceptance of oil drilling as cause, neither acceptance nor rejection, implicit rejection of oil drilling as a cause, and explicit rejection of oil drilling as a cause. (This analysis is performed only on the UIC Think-aloud and Eyetracking study participants, N=36, 7 subjects transcripts were lost due to camera failure.)

This table presents the frequency of each peer evaluation category by learning group.

	Accept oil	Implicit accept	Neither	Implicit reject	Explicit reject
Poor learners	8	3	2	1	3
Good learners	7	2	2	1	7

Better learners were more likely to explicitly note the weakness of an oil drilling explanation in the peer essay. Further, the evaluation of peer essays correlated well with our other measures. Students who explicitly rejected the oil drilling explanation had larger learning gains ($r=.46, p<.01$), higher reliability rankings for reliable sources ($r=.55, p<.01$), and tended to have fewer misconceptions in their essays ($r=-.29, p<.10$).

These results suggest better learners were somewhat more able to detect misconceptions/unreliable information in a peer essay. However, when their reasons for rejecting the oil account are examined, they seem to be fairly simplistic rejections (“I don’t agree with it”, “it doesn’t explain it well”) and were not justified in terms of reliability.

Justifications of Reliability Rankings

As discussed earlier, results from the entire sample indicated that better learning was associated with the ability to differentiate between reliable and unreliable web sites. In general reliable pages were ranked as more reliable than the unreliable pages, and the difference was greater for better learners.

After making rankings, students were asked to justify their rankings out loud. The transcripts of these justifications were examined in detail for the UIC Eyetracking sample, for those who improved on post-tests (learners, N=15) versus those that did not improve or did worse (nonlearners, N=14). In this whole sample, few students seemed to have a real understanding of reliability. Many students referred to whether the page seemed interesting or useful. Sometimes usefulness was defined as giving explanations about volcanic eruptions, other times people were specifically interested in finding information about Mount St Helens. Based on a surface coding of transcripts, we created 8 categories of contents, presented in the following table. Both good and poor learners used all categories of justifications. The frequency of comments in each category are presented in the following table.

Justification of Rankings	Nonlearners	Learners	
Page Design	.70 (.82)	.15 (.38)	$t(21)=2.1, p<.05$
Authenticity of Content	2.2 (3.5)	.84 (1.0)	
Reliability of Sources	.30(.67)	.38(.97)	
Relevance of Content	3.4 (3.5)	2.2 (1.3)	$p<.2$
Quantity/Interest	2.2 (2.2)	1.0 (1.3)	$p<.12$
Paraphrase of Content	1.7 (.5)	1.6 (1.1)	
Content Explains Why	0.7(1.1)	1.8(1.4)	$p<.2$
Other	2.8(2.0)	1.8 (2.0)	

Poor learners (nonlearners) were significantly more likely to justify their rankings by mentioning the design of the page. Overall, nonlearners were more likely to mention the Authenticity of Content, but this was all due to one nonlearner who made 11 comments on factuality of the information. Nonlearners were more likely to mention the direct

relevance of the content (i.e. Did it mention MSH?) or amount of information or interestingness of the information on page.

Both sets of learners often just paraphrased what content was on the page as a justification for their ratings. The learners tended to rate pages that explained why or how volcanoes erupt as more reliable. (This emphasis on explanation seems similar to what we observed in the High Knowledge advice-giving study and in the think-aloud protocol results for Good learners).

Summary:

On average there was very little use of the concepts that should have been used to justify judgments of reliability: source reliability or authenticity. Good learners attributed reliability to sites that offered explanations of volcanic eruptions, but they also seemed to use relevance (whether the page contained information related to volcanoes) as a basis for their justifications. Poor learners had even more confusion about what reliability should be based on, as they used the design of the page, interestingness of the material, and presence of information specifically about Mount St Helens to justify their ratings. These data suggest that readers need to be instructed in what reliability actually means.

Conclusions from Strand 1:

The conclusions that can be reached from our first strand of studies are that better learning seems to be related to the ability to discriminate between reliable and unreliable sources, the selection of reliable reading material for a high percentage of the research time, the detection and rejection of unreliable sites soon after they are selected, and an approach to learning that emphasizes explanatory processes. At the same time, even the best learners in our Strand 1 studies seemed to have a fragile understanding of what reliability is. Very few could verbalize it or use it to justify their evaluations of reliability.

In the next phase of our research program, we will be developing supports for learners that will build on these observations. Using both pre-task training in the concept of reliability, as well as an on-line interactive conversational agent that will prompt readers to evaluate sources, we will endeavor to support more critical and explanation-based processing among all learners in the second phase of this study.