American Educational Research Journal
http://aerj.aera.net

Source Evaluation, Comprehension, and Learning in Internet Science Inquiry Tasks

Jennifer Wiley, Susan R. Goldman, Arthur C. Graesser, Christopher A. Sanchez, Ivan K. Ash and Joshua A. Hemmerich

Am Educ Res J 2009 46: 1060 originally published online 27 March 2009
DOI: 10.3102/0002831209333183

The online version of this article can be found at:
http://aer.sagepub.com/content/46/4/1060

Published on behalf of

American Educational Research Association

and

http://www.sagepublications.com

American Educational Research Journal
December 2009, Vol. 46, No. 4, pp. 1060–1106
DOI: 10.3102/0002831209333183

Source Evaluation, Comprehension, and Learning in Internet Science Inquiry Tasks

Jennifer Wiley
Susan R. Goldman
University of Illinois at Chicago
Arthur C. Graesser
University of Memphis
Christopher A. Sanchez
Arizona State University
Ivan K. Ash
Old Dominion University
Joshua A. Hemmerich
University of Chicago

In two experiments, undergraduates' evaluation and use of multiple Internet sources during a science inquiry task were examined. In Experiment 1, undergraduates had the task of explaining what caused the eruption of Mt. St. Helens using the results of an Internet search. Multiple regression analyses indicated that source evaluation significantly predicted learning outcomes, with more successful learners better able to discriminate scientifically reliable from unreliable information. In Experiment 2, an instructional unit (SEEK) taught undergraduates how to evaluate the reliability of information sources. Undergraduates who used SEEK while working on an inquiry task about the Atkins low-carbohydrate diet displayed greater differentiation in their reliability judgments of information sources than a comparison group. Both groups then participated in the Mt. St. Helens task. Undergraduates in the SEEK conditions demonstrated better learning from the volcano task. The current studies indicate that the evaluation of information sources is critical to successful learning from Internet-based inquiry and amenable to improvement through instruction.

KEYWORDS: comprehension, discourse processes, science education

Students are increasingly turning to the Internet to conduct their research projects, regardless of whether the assignments are intended as Internet research projects or not (Jones, 2002). Internet searches are problematic in that they return multiple sources and sites that may or may not be relevant

or reliable. The use of the Internet for research purposes increases the need for students to critically evaluate information sources for their reliability, credibility, and trustworthiness (Britt & Aglinskas, 2002; Rouet, 2006; Wallace, Kupperman, Krajcik, & Soloway, 2000). Understanding how students engage in the processes of search, selection, evaluation, comparison, and integration of ideas from multiple sources of information is becoming an increasingly important area of research in discourse processing and comprehension (Brem, Russell, & Weems, 2001; Graesser et al., 2007; Rouet, 2006; Stadtler & Bromme, 2007) and in the learning sciences more generally (Linn, Davis, & Bell, 2004; Wallace et al., 2000).

In both history and science, experts routinely engage in selection, analysis, and synthesis within and across multiple sources of evidence (Chinn & Malhotra, 2002; Wineburg, 1991). For example, when scientists read scholarly publications, they rely on information about the scientists, the journals in which the publications appear, and the reputations of the institutions or research groups with which the scientists are affiliated (Bazerman, 1985; Berkencotter & Huckin, 1995). When scientists read

---

JENNIFER WILEY is an associate professor in the Department of Psychology at the University of Illinois at Chicago, 1007 W. Harrison Street (MC 285), Chicago, IL 60607; e-mail: *jwiley@uic.edu*. Her research explores contexts that promote successful comprehension and problem solving.

SUSAN R. GOLDMAN is Distinguished Professor of Liberal Arts and Sciences and Education and co-director of the Learning Sciences Research Institute at the University of Illinois at Chicago; Learning Sciences Research Institute (MC 057), 1007 W. Harrison Street, Room 2048, University of Illinois, Chicago, IL 60607-7137; e-mail: *sgoldman@uic.edu*. She investigates discourse processing and subject matter learning (literacy, science, mathematics, history), instruction, assessment, and roles for technology in supporting these.

ARTHUR C. GRAESSER is a professor in the Department of Psychology and Institute for Intelligent Systems, 365 Innovation Drive, University of Memphis, Memphis, TN 38152; e-mail: *a-graesser@memphis.edu*. He investigates discourse processing, cognitive science, and learning technologies, with particular interests in inferences, tutoring, question asking and answering, and emotions.

CHRISTOPHER A. SANCHEZ is an assistant professor in the Department of Applied Psychology at Arizona State University, Santa Catalina Hall, 7271 E. Sonoran Arroyo Mall, Mesa, AZ 85212; e-mail: *c.sanchez@asu.edu*. His research interests include individual differences in cognitive ability, science learning, and learning through technology.

IVAN K. ASH is an assistant professor in the Psychology Department at Old Dominion University, Norfolk, VA 23529-0267; e-mail: *iash@odu.edu*. He investigates problem solving, reasoning, and decision making.

JOUSHUA A. HEMMERICH is a research project manager in the Department of Medicine, Section of Geriatrics, University of Chicago, Chicago, IL, 60637; e-mail: *jhemmeri@medicine.bsd.uchicago.edu*. His current research concerns medical reasoning and decision making.

research reports within their field, they evaluate the strength of the argumentation and the answers to such questions as, Does the evidence support the claims? Is the evidence reliable? and Does the claim sufficiently explain existing as well as new evidence? (Chinn & Malhotra, 2002; Duschl, Schweingruber, & Shouse, 2007; Goldman, Duschl, Ellenbogen, Williams, & Tzou, 2003). Finally, new results and new explanatory models are framed against the extant literature (Yore, Bisanz, & Hand, 2003). Evaluation, explanation, integration, and corroboration of information across sources are all central processes in the disciplinary expertise of practicing scientists. Thus, from both a general discourse processing and comprehension perspective, as well as from a disciplinary perspective, it is important to understand how learners engage with multiple sources of information.

In the present research, we examined the comprehension processes, information evaluation processes, and learning outcomes of students who engaged in an Internet-based science inquiry task. In this task, students were asked to address the question "Why did Mt. St. Helens erupt?" from the first seven results of an Internet search. The task environment was designed so that students had to selectively integrate information from several sources to be able to complete the inquiry task accurately. First, the search results contained both reliable and unreliable sites. Second, none of the sources contained "the answer" to the question of why Mt. St. Helens erupted. Although this might seem a fairly obvious requirement for inquiry (namely, that the materials not hand students the answer on a silver platter), classroom instructors may incorrectly assume that an inquiry task is merely an activity in which students search through a textbook chapter to find an answer to the question. This form of "pseudo-inquiry" does not require the complex comprehension and integrative processing that a true inquiry task does (Chinn & Malhotra, 2002; Guthrie, 1988; Mosenthal, 1996). Thus, these two characteristics of this inquiry activity were intended to necessitate the use of evaluation and explanation-based processing during comprehension and also to simulate the likely characteristics of real Internet inquiry search results.

Science Inquiry

Science inquiry on the Internet can be seen as a discipline-specific case of learning from multiple sources. It is a case of what Palincsar and Magnusson (2001) referred to as secondhand investigation: inquiry that occurs through the exploration of materials (texts, diagrams, animations) created by others. Science inquiry on the Internet provides a special opportunity to investigate the more general processes of evaluation, analysis, synthesis, and integration of multiple sources of information for the purposes of producing an explanation of some phenomenon in the physical world. Particularly critical in this endeavor is understanding possible sources of bias in science investigations and how conflicting experimental reports can be evaluated (American Association for the Advancement of Science, 1993).

The inquiry process ordinarily begins with a dilemma, puzzle, or mystery that requires explanation and motivates student research (e.g., Why did dinosaurs become extinct? How can bacteria become resistant to antibiotics? Why are there so many earthquakes in California compared with the northeastern United States? Why did the population of Ireland change so drastically in the 1840s?). Generally speaking, inquiry involves five essential features: (a) engaging students in scientifically oriented questions, (b) using evidence to respond to questions, (c) formulating explanations on the basis of evidence, (d) connecting explanations to scientific knowledge, and (e) communicating and justifying explanations (Bransford, Brown, & Cocking, 2000). Skill in argumentation and the epistemological stance of a learner can influence the effectiveness of these activities (Bell & Linn, 2000; Chinn & Malhotra, 2002; Hofer, 2004; Sandoval & Reiser, 2004). For example, some students are not open to the possibility that there can be contradictions between scientific explanations or that more than one explanation may be plausible. Furthermore, effective inquiry learning requires a host of self-regulatory and metacognitive processes (Azevedo & Cromley, 2004; Brown & Campione, 1996; Graesser et al., 2007; Hacker, Dunlosky, & Graesser, 1998; Quintana, Zhang, & Krajcik, 2005; Winne, 2001). These include (a) planning (e.g., identifying goals, plans, and subgoals), (b) monitoring (e.g., identifying and detecting prior knowledge about the topic, evaluating the content of what is read, monitoring the progress toward the goals and subgoals, judging how much has been learned), and (c) implementing strategies (e.g., coordinating different information sources, drawing, taking notes, rereading, goal-directed searching of specific information, formulating inferences, reflecting, summarizing chunks of material).

Given the potential complexity of successful scientific inquiry, it should not be surprising that adolescents and college students frequently struggle with inquiry tasks, especially when these tasks involve learning through research articles (Janick-Buckner, 1997; Yarden, Brill, & Falk, 2001). Particularly germane to Internet-based science inquiry tasks are the skills and processes associated with searching, evaluating, and understanding information sources. Research indicates that high school and college students have difficulty differentiating claims from evidence, and evidence from conclusions, and tend to pay relatively little attention to source information (Azevedo & Cromley, 2004; Brem et al., 2001; Chinn & Malhotra, 2002; Korpan, Bisanz, Bisanz, & Henderson, 1997; Norris, Phillips, & Korpan, 2003; Stadtler & Bromme, 2007). This is problematic because of the centrality of understanding the claim-plus-evidence structure of scientific arguments and explanations of natural phenomena (Duschl et al., 2007). Thus, many Internet-based inquiry learning environments have found it necessary to include supports for inquiry learning through prompts and questions designed to help students focus on specific information, make critical contrasts and connections, distinguish claims from evidence, evaluate arguments, and monitor their own learning and understanding (Sandoval & Reiser, 2004; Slotta & Linn, 2000; White & Frederiksen, 2005).

Comprehension and Learning From Single and Multiple Sources

There are a number of similarities between the processes and difficulties learners face during Internet-based science inquiry tasks and the comprehension of multiple texts. From a discourse processing perspective, comprehension is a process in which learners attempt to construct mental models, also called situation models, of the subject matter that capture important concepts from the text and their relationships (Graesser, Singer & Trabasso, 1994; Kintsch, 1998; Otero, Leon, & Graesser, 2002). This level of representation can be contrasted with surface-level comprehension, in which there is a more transitory representation of text that preserves the exact words and explicit ideas. A substantial body of research supports the claim that the construction of mental models, particularly the recognition of causal relationships, underlies the comprehension process, as texts with clear causal connections are read faster, remembered better, and support more robust inference generation. (Trabasso & van den Broek, 1985; Wiley & Myers, 2003).

Learning From Single Texts

Research on the comprehension of expository material, namely, texts from which readers are supposed to learn content, indicates that there are several consistent characteristics of more successful readers and learners. Successful readers achieve deep comprehension by connecting ideas within a text to one another, connecting the ideas in the text with relevant prior knowledge, explaining these ideas and connections, and actively engaging with the text to construct coherent situation model representations (Chi, de Leeuw, Chiu, & Lavancher, 1994; Coté, Goldman, & Saul, 1998; Graesser & Bertus, 1998; McNamara, Kintsch, Songer, & Kintsch, 1996; Voss & Silfies, 1996). This type of deep understanding is more likely to occur if learners are prompted to question the material, make inferences or connections between pieces of information, or generate explanations (Chi, 2000; Craig, Sullins, Witherspoon, & Gholson, 2006; Graesser & Person, 1994; King, 1999; McNamara, 2004; Rosenshine, Meister, & Chapman, 1996; Wiley & Voss, 1999).

In contrast, less successful comprehension is reflected in fragmentary representations of information that are likely the result of the failure to make inferences, generate predictions, or draw conclusions about subject matter content (Coté et al., 1998; Garner & Alexander, 1994; Pressley & Ghatala, 1990). When less successful readers do make connections among ideas, they tend to make surface-level connections. Rather than explaining ideas in a text, they tend to paraphrase or restate verbatim the information presented in the text (Coté et al., 1998; Magliano & Millis, 2003; O'Reilly & McNamara, 2007) or focus on details in the text that are not related to developing an understanding of the phenomena (Sanchez & Wiley, 2006; Thiede, Griffin, Wiley, & Anderson, in press). In essence, successful comprehension from expository text involves understanding the underlying explanatory principles and causal mechanisms of a phenomenon, not just the "what" and "where" (Graesser & Olde, 2003; Kintsch, 1998; Wiley, Griffin & Thiede, 2005).

Effective metacognition also plays a role in the processes and outcomes of comprehension. Successful comprehenders better monitor the adequacy of their text representation and use a range of strategies in response to failures to understand what they are reading (Duke & Pearson, 2002; Garner, 1987; Griffin, Wiley, & Thiede, 2008; Hacker et al., 1998; McNamara, 2004; Palincsar & Brown, 1984; Pressley, 2002; Thiede, Griffin, Wiley, & Redford, in press). When students fail to monitor their understanding accurately, such as by making inaccurate judgments of learning, information quality, and the relevance of information to goals, they make poor study decisions and fail to reread misunderstood information (Thiede, Anderson, & Therriault, 2003). The result is little or no improvements in comprehension and ultimately poor overall learning outcomes (Wiley et al., 2005; Winne, 2001).

Learning From Multiple Texts

Research on multiple source comprehension and learning is just emerging, although some initial models that build on single text models have been proposed (Perfetti, Rouet, & Britt, 1999; Rouet, 2006). The few studies that have explored comprehension from multiple texts have concentrated on history (Britt & Aglinskas, 2002; Rouet, 2006; Wiley & Ash, 2005; Wiley & Voss, 1999; Wolfe & Goldman, 2005). Just as it is for single texts, the role of explanation-based processing is also important for developing an understanding of the subject matter from multiple texts (Graesser et al., 2007; Perfetti et al., 1999). For example, in one multiple-text study conducted with 11- and 12-year-old students, integration across sources and explanations that built causal connections among concepts were significant predictors of more complex understanding of a historical event (Wolfe & Goldman, 2005).

Prior research has suggested that focusing students on explanations and information integration across sources during Internet inquiry tasks with multiple sources leads to improved learning outcomes (Wiley & Voss, 1999). In Wiley and Voss (1999), the writing task was manipulated to focus students on the task of integrating evidence across sources. Students who were asked to write arguments from multiple sources wrote essays with more causal connections and better integration of ideas than students who were asked to write narratives or descriptive essays. Students also showed better performance on several learning outcome measures. The present research included a similar writing task manipulation to test whether an argument instruction might also prompt more evaluative processing than the instruction to write a descriptive essay. At the same time, there is little evidence in either the single-text or multiple-text comprehension literature that learners spontaneously engage in evaluations of the quality of sources or the information in the sources (Britt & Aglinskas, 2002; Rouet, 2006; Rouet, Favart, Britt, & Perfetti, 1997; Wineburg, 1991). Indeed, in other studies on history, high school students tended to approach the different texts uncritically (Greene, 2004; Rouet, Britt, Mason, & Perfetti, 1996), as opposed to the way in which expert historians process and represent information from historical

sources (Wineburg, 1994). On the basis of these results, Perfetti et al. (1999) proposed a new *theory of documents representation* to capture the additional elements that arise from the simultaneous comprehension of multiple sources. The new framework consisted of the individual representations (situation models) for each of the sources, the situations model that reflects the overall understanding of the event or phenomenon from integrating across a set of sources (or multiple situations), and the intertext model, which contains representation of metainformation about individual texts (such as the authors, attributions about the sources of the texts, and evaluations of text reliability or quality) in document nodes. It is also the intertext model that contains information about the relations between texts, such as instances of converging or corroborating evidence or contradictions across sources.

The elements of the intertext model are what are generally missing in the representation process of novice readers (Britt & Aglinskas, 2002; Rouet et al., 1997; Voss & Wiley, 2006; Wineburg, 1991). They also seem to be critical features for the comprehension of multiple sources during Internet-based science inquiry tasks, although this has not yet been tested. Because much of this information is on a metalevel, requiring reflection, evaluation, and monitoring on the part of the student, comprehension from multiple Internet sources may be even more reliant on effective metacognition than comprehension of single texts (Quintana et al., 2005; Stadtler & Bromme, 2007). Thus, these processes may need particular support during Internet inquiry tasks.

The Present Research

The source evaluation and comprehension processes of students who were engaged in an Internet-based science inquiry task were examined using a *multiple-source comprehension framework*. In the process of comprehension, learners were expected to attempt to construct coherent causal mental models of volcanic eruptions as they read multiple sources. The current experiments manipulated learning conditions that were expected to encourage the creation of intertext models, or improve their quality, to explore their role in multiple source comprehension.

In the first experiment, we attempted to manipulate the quality of intertext models through a writing task manipulation. The main finding of Experiment 1 was that more successful learners did indeed create stronger intertext models than less successful learners. Accordingly, in Experiment 2, an intervention was designed to support learners in the construction of intertext models by providing instruction on how to determine the reliability of Web sites in the context of a unit about the Atkins low-carbohydrate diet. The effects of this intervention were then tested in relation to learning outcomes during a subsequent inquiry task on the Mt. St. Helens eruption.

Experiment 1: Comprehension and Learning During an Internet Inquiry Task

In Experiment 1, we tested whether a writing task manipulation would affect the development of intertext models during the completion of an Internet-based science inquiry activity involving multiple information sources, and more generally how the creation of intertext models would relate to learning. Students were instructed to write essays explaining why the Mt. St. Helens volcano erupted. They were provided with a set of Web sites with which they were to conduct research on the topic of the essay. Structuring the task as an inquiry task rather than a memorization or recognition task was intended to encourage active exploration, engagement, and a critical, explanation-seeking stance toward the information sources on the Web sites. Traces of exploration behavior, including what sites were accessed, how often, and for how long and how this time was distributed across information within each site, were all collected. These processing profiles were used to understand the learners' approaches to finding information and constructing meaning across multiple sources of information about volcanoes. In addition, because of this emphasis on process, two other sources of processing data were also collected for subsets of the sample. For some of the participants, eye movement data were collected to obtain more fine-grained measures of the processing of the texts. For some participants, think-aloud protocols were obtained. These data were collected to provide greater insight into readers' selection, evaluation, and comprehension processes, the details of which are reported elsewhere.

Two main measures of learning were used to test comprehension: a student essay answering the question "What caused the eruption of Mt. St. Helens?" and a volcano concept recognition test that tested elements of the causal model. Furthermore, several measures of evaluative processing served as assessments of how extensively students developed intertext models of the sources during the inquiry task. Students' evaluation skills were assessed after the inquiry task in three main ways: (a) judgments of the reliability of the sources that were read, (b) the frequency of references to the quality or nature of each source that were expressed in justifications of reliability judgments, and (c) judgments of the quality of an essay purportedly written by a peer.

The main experimental variable was the writing condition that was presented to students. Half of the students were asked to write arguments about why Mt. St. Helens erupted, whereas the other half were asked to write descriptive essays. Previous studies have found that argumentation instructions lead to the generation of more integrated and causal essays, as well as improvements in a number of measures of learning when learning from multiple sources (Wiley, 2001; Wiley & Voss, 1999). Thus, an argument writing task appears to have analogous effects to other "explanation" instructions in terms of prompting students to integrate information across units of text and with prior knowledge (Chi, 2000; McNamara, 2004; Wiley & Voss, 1999) and should facilitate the construction of single-source situation

models. It could also lead to more coherent integration into a situations model. In addition to prompting integration and explanation, argumentation tasks are thought to improve learning by focusing students on the relations between claims and evidence and prompting them to engage in evaluative behaviors (Andriessen, Baker, & Suthers, 2003; Bell & Linn, 2000; Kuhn, 1993). Thus, an argument writing task was predicted to lead to better development of intertext models because students should be more likely to note sources of evidence, identify instances of corroboration across sources, and possibly pay attention to the quality of the sources. In summary, our hypothesis was that an argument writing condition should support the construction of better document models during an Internet inquiry task with multiple information sources.

Method

Participants

This research was conducted with undergraduates who participated as part of an introductory psychology subject pool at two state universities. The introductory psychology subject pool at one institution contains approximately 700 participants each semester. The average age of participants is 19 years, and most are in their first or second year of college and have not yet declared majors. The pool is 60% female and 40% male. Participation is generally 20% Hispanic, 10% African American, 30% Asian, and 40% White. These match the demographics of the university as a whole. The minimum American College Test (ACT) score required for the institution is 18, with an average ACT score for incoming classes of around 24.

The introductory psychology subject pool at the second institution contains approximately 950 participants each semester. The average age and the proportion of men to women are similar to those of the first pool. The ethnic makeup is generally 40% African American and 55% White, with about 2.5% Hispanic and Asian. The minimum ACT score required for the institution is 18, with an average ACT score for incoming classes of around 22.

For the main study, 110 undergraduates engaged in an Internet research activity in which they were tasked with understanding the causes of volcanic eruptions. These participants were run in four different methodology conditions: 28 performed the Internet research task on a Dual-Purkinje eye tracker, 34 engaged in thinking aloud as they performed the Internet research task, 24 performed the task on a head-mounted eye tracker while thinking aloud, and 24 performed the task with neither eye tracking nor thinking aloud. Half of each methodology condition was assigned the argument essay instruction, and half was assigned the descriptive essay instruction.

During the semester that this main inquiry study was run, a pretest was administered to the entire subject pool during mass testing at the beginning of the term. Only participants who scored at the mean or less (19 correct out of 30) were allowed to participate in the main study. This ensured that students did not already possess the target knowledge to be learned in the

inquiry activity. The average pretest score for the inquiry sample was 16.79 (*SD* = 1.94). There was no significant difference in pretest scores between the two writing conditions. However, there were significant differences across methodology conditions, $F(3, 102) = 3.54$, $MSE = 3.55$, $p < .02$, $\eta^2 = .09$. Importantly, the interaction was not significant. Pretest scores are included as a covariate in the multivariate analysis of covariance (MANCOVA) for this study.

The mean age of participants in the main inquiry study was 19.14 years (*SD* = 1.67 years; 1 missing data point). In this sample, 42 students were male and 67 were female (1 missing). By class, 50 participants were freshmen, 38 were sophomores, 15 were juniors, and 7 were seniors (2 missing). The average number of college science courses taken was 1.05 (*SD* = 1.24; 1 missing), including courses in which they were currently. The average number of courses taken specifically in earth science was 0.38 (*SD* = 0.49). On a scale ranging from 1 to 10, with 10 meaning *a lot* and 1 meaning *not much*, students rated their prior understanding of plate tectonics and volcanoes at a mean of 3.50 (*SD* = 2.15; 1 missing). They also rated their familiarity with the reading material from 0% to 100%. Average familiarity with the reading material was 41.47% (*SD* = 24.98; 1 missing). There were no significant differences among either writing or methodology conditions on these background variables.

Another 90 students participated in a no-reading comparison group, writing essays on the causes of volcanic eruptions without having engaged in the inquiry activity. They did not take the pretest or background survey.

Procedure

Students in the inquiry study were given 1 hour to read the Internet sources provided through a browser window for the goal of writing a report on what caused the eruption of Mt. St. Helens. Half the participants were asked to write "descriptions" of what caused the eruption, and the other half were told to write "arguments" of what caused the eruption. As soon as a participant indicated that he or she was finished with the research, the browser was closed. At that point, the participant was instructed to write the essay. Following reading and writing, all participants received the volcano concept recognition task. Following the learning assessments, students were asked to evaluate the quality of each Web site they read and to justify their evaluations. Next, they were asked to evaluate the quality of a "peer" essay. At the end of the study, participants were asked to fill out a short survey with questions about their age and educational backgrounds.

A second sample of participants (the no-reading comparison group) was asked to respond to the open-ended questions "What caused the eruption of Mt. St. Helens?" and "What is a volcano and why do they erupt?" without having read the Internet sources. This was the only task they completed.

Materials and Coding

Understanding Volcanic Eruptions

Volcanic eruptions were selected as the inquiry topic because the time scale of plate movement makes secondhand investigation particularly appropriate. Furthermore, volcanic eruption is a complex phenomenon that results from many interacting causal factors.

An explanation of the eruption of Mt. St. Helens requires that the learners understand several basic concepts about earth structure dynamics: the principles of plate tectonics, the crustal cycle, and their role in a causal model underlying volcanic eruptions. An adequate understanding of these concepts involves the integration of scientific information from several subdisciplines of science, including information on the physics and chemistry of the earth's crust, interpretation of data about earthquake and volcanic occurrences, and analysis of descriptive information on different types of earthquake and volcanic events (Wiley, 2001). An understanding of plate tectonics therefore requires more than the memorization of facts and procedures. It requires a representation that reflects the integration of multiple causal factors and sources of evidence.

An accurate understanding of the causal model underlying the eruption of Mt. St. Helens entails the 13 core concepts described in Appendix A. The foundation of this model was information available from the U.S. Geological Survey's "This Dynamic Earth" Web site (http://pubs.usgs.gov/publications/text/dynamic.html), a NASA Classroom of the Future module on volcanoes (http://www.cotf.edu/ete/modules/volcanoes/volcano.html), and an introductory geology college textbook (Marschak, 2001). This model represents current scientific understanding of volcanic eruptions, which of course is subject to revision on the basis of advances in the area. The emphasis here is not on learning this particular model per se but rather on learning the model that is best supported by current scientific thinking and evidence.

This causal model was used to guide our selection and adaptation of the content of the information sources and to design the assessments and scoring systems that would be used to determine what students knew prior to the research and what they learned in the course of their inquiry during the research study.

Information Sources

Students in inquiry conditions were given seven "Internet sources" on volcanic eruptions. Students were told that we did a search on Google using the phrase "causes volcanic eruptions" and that they were being presented with the top seven hits (the first page of results) from this search. The sources were presented on what appeared to be a Google search results page, with the page titles listed as hotlinks, with original uniform resource locators, and with short descriptions of the contents of the pages (shown in Figure 1). All pages were presented through a browser but were actually

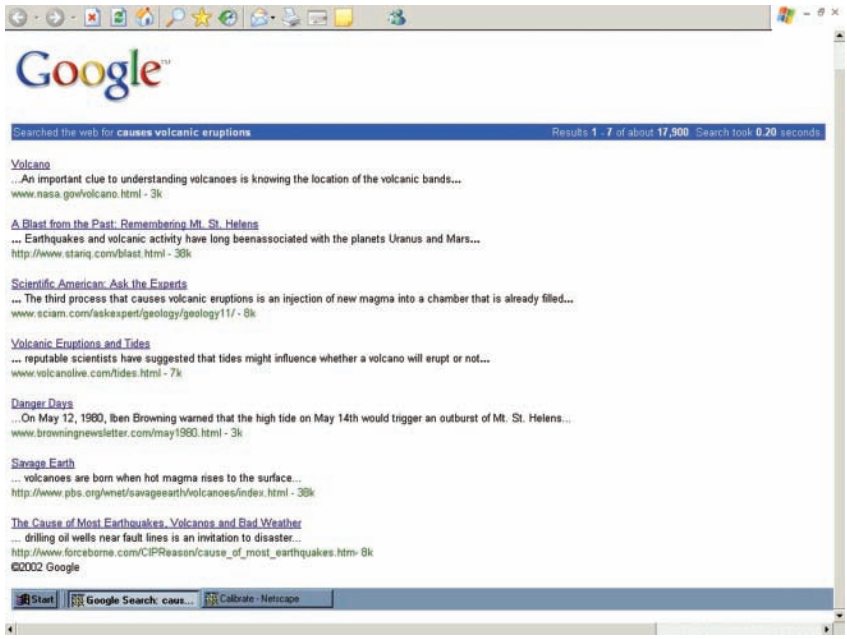

Figure 1. **Google search output page.**

stored locally. There are several advantages to providing students with a finite set of preselected materials that appear to be sources found via an Internet search. It is easier to keep track of what information is available, and what specific areas of text are important for understanding, while preserving a naturalistic feel to the Internet inquiry task. Most pages contained between 300 and 500 words of text. In addition, most pages included diagrams, maps, or photographs to illustrate the point of the text. All pages contained headers, notes, logos, or symbols to identify their sources.

There were two versions of the Google search results page that differed in the order in which sites were listed. Half the participants viewed one order, and the other half saw the other. No differences in any dependent measures due to order were found. The sources were modified to fit in discrete pages for eye tracking (i.e., scrolling pages were separated into multiple pages that were each one screen long). In some cases, the language was simplified for readability and consistency across sites, whereas in other cases, the sources were pruned to present only part of the whole site. In all other respects, they looked like the original sources.

The seven sources were selected to represent a range of reliability. Three reliable sources were hosted by known reputable organizations: NASA, *Scientific American*, and PBS. These three Web sites contained

accurate information and provided converging evidence for one another (i.e., the reliable sites all provided information that could be integrated into the same coherent causal model of volcanic activity shown in Figure A1). The pages were edited so that each contributed to the creation of a complete causal model of eruptions. No single site contained all the necessary information. It is important to point out that there was some overlap between pages, so that different parts of the causal account were corroborated across sources. The reliable sites contained a total of 5,939 words.

Three other sources were included to offer incomplete, and unreliable, accounts of seismic and volcanic activity so that across the seven sources, there was variance in accuracy and reliability. The unreliable sites were an astrology site (StarIQ.com) that attributed the Mt. St. Helens eruption to the location of the planets and stars; an inventor's site (Forceborne.com) that was promoting an engine that did not run on fossil fuels and claimed that oil drilling caused volcanoes to erupt; and a site (the Browning Newsletter; http://www.browningnewsletter.com) written by a corporate forecaster (Iben Browning) who claimed that tidal fluctuations allowed him to predict earthquakes and volcanic eruptions. All sites except the Browning Newsletter site were portions of real sites found via a Google search. Iben Browning was in fact a real corporate forecaster who claimed to have predicted the eruption of Mt. St. Helens and who produced newsletters in the 1980s. His printed newsletters were put in Web format for this set of studies. In general, the unreliable sites were similar in format to the reliable sites. These sites all provided evidence for their positions but also offered unique causal information that could not be integrated into the model suggested by the reliable sites and could not be corroborated with any other source. A total of five erroneous causes for volcanic eruptions were contained in the unreliable sites. The unreliable sites contained a total of 5,090 words.

The seventh site was Volcano Live (http://www.volcanolive.com), a commercial educational site that had 683 words. The information on this page is reliable, as it cites U.S. Geological Survey as the source of its information, and the evidence fits into the causal model supported by the other reliable sites. However, its status as a commercial site (with a .com top-level domain) rather than an educational site (with a .edu top-level domain) should make the reliability of the site more difficult for most people to assess.

Learning Outcome Assessments

Our most important indicators of student comprehension were the essays that students wrote in response to the prompt "What caused the eruption of Mt. St. Helens?" as the product of the inquiry task. These were coded for the number of core concepts mentioned from the underlying causal model using a detailed coding rubric that was developed on the basis of theoretical assumptions from prior studies and data from a subset of the responses. Concepts that were expressed that did not fit into the accurate

causal model of eruptions were coded as erroneous causes. All essays were coded by two raters. The Pearson correlation between the total number of causes in each essay as identified by two independent raters was .99. Discrepancies were resolved by a third rater. The detailed guidelines provided to our raters afforded generally strong agreement.

Essays were also categorized in terms of the models of volcanic eruptions expressed in the essays (Hemmerich & Wiley, 2002, based on Gobert, 2000). The different levels of the scoring typology are described below. Model categorization was derived in large part from the causal coding. The only additional coding required was whether each essay that contained multiple causes of heat, pressure, and movement made an explicit causal connection between them. Agreement on the presence of this causal connection as coded by two raters was 100%.

Type 0: Incorrect, superficial models. Models of Type 0 contained explanations of the cause of volcanoes that were related to irrelevant surface features of the earth. Examples of Type 0 explanations are that volcanic eruptions are caused by surface conditions, such as wind, avalanches, landslides, mountains, weather, being near the equator, tropical climate, the sun, the orbit of planets, tides, faults, time, old age, dormancy or too much lava, or a specific kind of rock, as well as nonexplanations. Essays that did not include any of the major causal agents identified below were classified as Type 0.

Type 1: Local models. Models of Type 1 mentioned one (and only one) of three local causes: heat, movement, or pressure. Explanations were coded as being of Type 1 if they mentioned hot, melting, or molten magma, the temperature of the magma, or the heat of the earth's core; the movement, shifting, colliding, rubbing, or interacting of plates; or pressure-related ideas (e.g., that volcanoes or the earth were full of gas, that the magma or lava had too much gas, that there was pressure, or that the magma or lava was being kept under force).

Type 2: Mixed models. Models that included sets of factors, including plate movement with heat, pressure, force, or chemical processes, were coded as Type 2. In these models, multiple correct factors were mentioned but not causally related to one another.

Type 3: Integrated models. Only models that causally related heat or pressure and plate movement in either direction (i.e., convection currents cause plate movement, or plate movement causes plates to subduct and melt, forming magma that rises under volcanoes) were coded as Type 3 models. Thus, the highest level of conceptual understanding in this coding scheme required an explanation that involved both the notions of heat or pressure and plate movement and the causal relation between them.

A second learning outcome assessment, the volcano concept recognition task, was used to test whether students could correctly verify statements

related to the scientifically accepted causal model of volcanic eruptions on the basis of their inquiry. The 30-item recognition test (based on Wiley, 2001) contained 10 items that were true according to the causal model of volcanic eruptions (i.e., "Volcanoes occur at plate boundaries") and 10 statements that were false (e.g., "Volcanoes are randomly located around the earth"). There were also an additional 10 items that were not related to the correct causal model. Five items were false but related to misconceptions stated in the Internet sites that were presented to participants in the inquiry conditions (e.g., "Oil drilling causes eruptions"), and 5 additional misconception statements were false and not mentioned in the reading material (e.g., "Sunspots cause eruptions"). A second version of this test, with items in a different order, was administered as a pretest in a mass testing session at the beginning of the term.

Navigation and Reading Behaviors

To capture processing activities during the inquiry tasks, reading times were collected on all visits to each page. Navigation behaviors were also recorded through navigation logs. Navigation patterns were examined with several assumptions in mind. First, as readers become familiar with the information available to them, they should choose to return to the reliable information more often as they attempt to construct complete and coherent causal accounts of volcanic eruptions. This requires recognizing and generating connections across reliable sites. Failure to return to information could indicate a lack of constructive or integrative processing on a reader's part. Returning to unreliable information suggests that readers may be actively constructing an inaccurate model of eruptions. On the basis of these assumptions, the observed navigation patterns were sorted into four categories, listed in order of assumed increasing effectiveness:

1. *Selective rereading of unreliable information:* This involves returning to unreliable sites more than once, and if reliable sites are also reread, returning to unreliable sites at least twice as often.
2. *Single reading:* Each site is read only once, with at most one return to any site.
3. *Nonselective rereading:* This involves making more than one return visit to a site, with visits neither to reliable nor to unreliable sites occurring more than twice as frequently as the other.
4. *Selective rereading of reliable information:* This involves returning to reliable sites more than once and, if unreliable sites are also reread, returning to reliable sites at least twice as often as unreliable sites.

Evaluation Tasks

There were two main tasks that yielded three measures of evaluative activity. For the source reliability evaluation task, students were provided with a printout of the Google search page and asked to rank the seven sites on the basis of how reliable they thought they were (with a rank of 1

meaning *most reliable*). For this task, the evaluation of the Volcano Live site was of particular interest, because it contained information that was corroborated by the reliable sites, but surface features (its .com address) might have led students to think of it as unreliable. It was assumed that source evaluations would constitute part of the intertext model and that better developed models would lead to better discrimination among reliable and unreliable sources.

After making their rankings, students were also asked to explain why they ranked the sites the way they did and what was important in their decisions. These explanations were broken down into eight categories by content. The κ value for the agreement of the two raters who categorized the reliability justifications comments was .92. Discrepancies were resolved by discussion. The categories were comments on the source of the information, comments on the presence of evidence for claims, an evaluation of the information (whether the reader agreed with or believed the information), whether the site provided an explanation, whether the information seemed relevant or contained keywords, whether the student liked the site or found it interesting, comments about the style of the page (including whether it contained images), and other (simple paraphrases or comments too vague to be coded). The first category, explicit comments on the source of the information, was a measure that was of particular interest because it is an indicator of a document node and at least initial formation of an intertext model of the sources.

In the second evaluation task, students were asked to evaluate the quality of an essay "written by another student" about the eruption of Mt. St. Helens. They were told to use what they learned in their research to evaluate the quality of the other student's explanation and provide reasons for their evaluation. All students read the same peer essay, which was based on an actual student response that was edited so that it contained some accurate information about the causes of volcanic eruptions but also an incorrect assertion that oil drilling causes volcanic eruptions.

The peer essay evaluation task was intended as a test of whether students could apply what they learned to evaluate an account of volcanic eruptions. An evaluation that explicitly identifies and rejects the conception that oil drilling causes eruptions suggests that a student has acquired an accurate understanding of volcanic eruptions and also that the student may have encoded the source of the oil-drilling hypothesis as being from an unreliable source as part of an intertext model. An understanding that oil drilling is a poor explanation is considered the best response on this task. Explicit acceptance of oil drilling as a cause suggests that the student has not learned the scientifically accepted causes of volcanic eruptions, is unwilling to engage in critical evaluation of the oil-drilling account, or has failed to develop an evaluation of each document as part of an intertext model. Thus, this is considered the poorest response on this task. The coding scheme that was used for these responses was divided into five categories (from poor evaluation to best): explicit acceptance of oil drilling as a cause of volcanoes,

Table 1
Proportions of Essays Coded by Model Type in Experiments 1 and 2 (frequencies in parentheses)

	Experiment 1			Experiment 2	
Model	No-Reading Comparison	Argument Condition	Description Condition	SEEK Condition	Comparison Condition
Type 0	27% (24)	5% (3)	11% (6)	0% (0)	17% (5)
Type 1	52% (47)	38% (21)	53% (29)	13% (4)	37% (11)
Type 2	17% (15)	38% (21)	25% (14)	27% (8)	20% (6)
Type 3	4% (4)	18% (10)	11% (6)	60% (18)	27% (8)

Note. SEEK = source, evidence, explanation, knowledge (see text).

implicit acceptance of oil drilling as a cause, neither acceptance nor rejection, implicit rejection of oil drilling as a cause, and explicit rejection of oil drilling as a cause. The κ value for the agreement of the two raters on the peer essay coding into the five categories was .89. Discrepancies were resolved by a third rater.

Results and Discussion

The results of the no-reading comparison group are presented first to provide a baseline for understanding on this topic, followed by learning outcomes for the groups that engaged in the Internet inquiry task. Next, the effects of the writing task manipulation on learning, processing, and evaluation measures are examined, and then relations between the three sets of measures are considered. These quantitative analyses are followed up with a few selective contrasts between more and less successful students to better understand the quality of the intertext models that were developed.

Volcanic Understanding in the No-Reading Comparison Group

A sample of undergraduates was used to determine the typical level of understanding of this topic in the same college population. This was useful for comparison with our instructional conditions. The answers to the open-ended questions were written without this sample being given anything to read about volcanoes. These answers were classified according to the models of volcanic eruptions that were defined above. As shown in Table 1, over a quarter of the undergraduates (27%) in this no-reading sample expressed either simplified, incomplete, or incorrect models of why Mt. St. Helens erupted (e.g., volcanoes are mountains that have or emit lava, eruptions are caused by the weather). Several undergraduates included ideas that did not correspond to a correct understanding of eruptions, such as that volcanic eruptions are caused by too much heat or lava in some kind of closed-container metaphor, the age of the mountain or its dormancy, the nature of

the rock being full of minerals or lava, the location of a mountain in the tropics or Hawaii, the seasons, planetary orbits, and a buildup of dust or dirt. Only 4% of undergraduates in the sample expressed an integrated basic model of volcanic eruptions that combined both elements of plate movement and the buildup of heat or pressure. Statements about the nature of the magma, and its formation from the subduction and melting of plates, were almost entirely absent. They did not convey how eruptions ultimately relate to other concepts, such as rock formation or the crustal cycle.

The results from this no-reading comparison group confirm that volcanic eruptions are a topic with which many undergraduates have difficulty, especially in relation to the dynamics of the earth's crust (Barrow & Haskins, 1996; Marques & Thompson, 1997). Undergraduate students in this population typically possessed an incorrect, a superficial, or a highly simplified understanding of volcanic eruptions.

Learning From Inquiry Task

The models of the no-reading comparison group were compared with the models of students who engaged in the inquiry task, as shown in Table 1. Students in inquiry conditions were more likely to have Type 2 or 3 models and less likely to have Type 0 models than were students in the no-reading condition, $\chi^2(3) = 21.6$, $p < .001$, $\varphi = .33$. Students in inquiry conditions also included more correct causal concepts in their essays ($M = 3.89$, $SD = 2.04$) than students in the no-reading condition ($M = 1.73$, $SD = 1.55$), $F(1, 198) = 68.5$, $MSE = 3.37$, $p < .001$, $\eta^2 = .26$. Interestingly, they also included more erroneous causes in their essays ($M = 0.96$, $SD = 1.20$) than students in the no-reading group ($M = 0.23$, $SD = 0.52$), $F(1, 198) = 28.7$, $MSE = 0.91$, $p < .001$, $\eta^2 = .12$. In terms of performance on the volcano concept recognition test, there was a significant gain of 1.66 items correct out of 30 ($SD = 3.1$) from pretest to posttest for all students in inquiry conditions, $t(109) = 5.55$, $p < .0001$, $\eta^2 = .22$. Thus, outcome measures from both the essays and the concept recognition tests indicated that learning did occur as a function of the inquiry task.

Effects of Writing Condition

To test for the effects of receiving an argument writing task versus a description writing task on learning, processing, and evaluation measures, all continuous variables were entered into a 2 × 4 MANCOVA (Writing Condition × Methodology Condition), with pretest scores entered as a covariate. The results of the MANCOVA are presented in Table 2. Overall, there was a significant main effect for writing condition, $F(9, 93) = 2.89$, $p < .01$. There was also a significant main effect for methodology condition, $F(27, 285) = 3.36$, $p < .01$. Importantly, there was not a significant interaction ($F < 1.3$). The results of the follow-up tests on the main effects for each measure, effect sizes, and the direction of significant effects in terms of conditions are reported in Table 2.

Table 2
Results of Experiment 1 Multivariate Analysis of Covariance and Follow-Up Tests

	Main Effect of Writing	Main Effect of Methodology
Core causes	Arg > Desc (η^2 = .06)	—
Erroneous causes	Arg < Desc (η^2 = .15)	—
Posttest scores	—	—
d′ on posttest	—	—
Time on reliable pages	—	ET, ET/TA < TA, NO (η^2 = .17)
Time on unreliable pages	Arg < Desc (η^2 = .06)	ET, NO < TA (ET/TA in between) (η^2 = .13)
Returns to reliable pages	—	NO < ET (TA, ET/TA in between) (η^2 = .13)
Ranking discrimination	—	—
Volcano Live site ranking	—	—
References to source	Arg > Desc (η^2 = .07)	—

Note. Arg = argument condition; Desc = descriptive essay condition; ET = eye tracking alone; TA = think-aloud alone; ET/TA = eye tracking and think-aloud combined; NO = neither. Only significant effects are reported. Conditions not included in these comparisons were not significantly different from other listed conditions.

Effects of the writing manipulation were seen in learning, as students in the argument condition had more correct causal concepts (M = 4.27, SD = 1.97) and fewer erroneous causal concepts (M = 0.51, SD = 0.81) in their essays than students in the descriptive essay condition (correct: M = 3.51, SD = 2.05; erroneous: M = 1.42, SD = 1.36). No effects were seen in either total correct or *d′* values on the concept recognition posttest. Effects were seen in a few processing and evaluation behaviors, as students in the argument condition spent less time reading unreliable pages (M = 11.62 minutes, SD = 5.30 minutes) than students in the description condition (M = 13.94 minutes, SD = 6.56 minutes) and were more likely to explicitly mention the nature or quality of the sources when justifying their evaluations of the reliability of the Web sites (argument: M = 0.88, SD = 1.27; description: M = 0.44, SD = 0.86).

Three measures (model type, navigation patterns, and peer essay evaluations) could not be included in the above analysis because of the categorical nature of the data, so nonparametric tests were used to examine the effects of writing condition on these variables. Because of the lack of an interaction in the MANCOVA, analyses examined the distributions in terms of only the main writing manipulation. As shown in Table 1, essays in the argument writing condition were significantly more likely to fall in the upper two model type categories (31 of 55) than essays in the description writing conditions (20 of 55), $\chi^2(1) = 4.42$, $p < .03$, $\varphi = .20$.

As shown in Table 3, when navigation patterns were considered, participants in the argument writing condition were more likely to engage

Table 3
Navigation Patterns by Writing Condition and for More and Less Successful Students in Experiment 1 and SEEK Instruction and Comparison Groups in Experiment 2

	Experiment 1				Experiment 2	
Navigation Patterns	Argument Condition	Description Condition	More Successful Students	Less Successful Students	SEEK Condition	Comparison Condition
Reread bias for unreliable	13% (7)	20% (11)	5% (1)	25% (5)	0% (0)	10% (3)
Read once	42% (23)	51% (28)	40% (8)	60% (12)	47% (14)	43% (13)
Reread nonselective	20% (11)	18% (10)	20% (4)	15% (3)	20% (6)	43% (13)
Reread bias for reliable	25% (14)	11% (6)	35% (7)	0% (0)	33% (10)	3% (1)

Note. SEEK = source, evidence, explanation, knowledge (see text). Data given are shown as percentage of students using each pattern, with frequency in parentheses.

in rereading and selective rereading of reliable sources, whereas participants in the description writing condition were more likely to read once or engage in selective rereading of unreliable sources. However, these trends were not significant, $\chi^2(3) = 5.00$, $p > .16$, $\varphi = .21$.

Students who wrote arguments were significantly more likely to make more critical evaluations of the peer essay and explicitly reject the oil drilling account (13 of 50; data missing for 5 participants) than students who wrote descriptions (4 of 54; data missing for 1 participant). Conversely, students who wrote descriptions were more likely to accept the oil-drilling account (31 of 54) than students who wrote arguments (14 of 50). This led to a significant difference in the distributions of reactions to peer essays due to writing condition, $\chi^2(4) = 13.1$, $p < .01$, $\varphi = .36$.

Summary of Writing Condition Effects

As seen in Table 2, the follow-up tests revealed that there were a number of medium to large effects (Cohen, 1992) of the writing condition on some learning outcomes, as well as on a few processing and evaluation measures. Specifically, the argument condition produced better performance on those learning measures based on essays but had little impact on the concept recognition posttest measures. There was some evidence that the argument condition encouraged participants to favor reliable over unreliable sites and to engage in some source evaluation and more critical evaluation of a peer essay. Although the effect sizes of the writing manipulation on the peer evaluation measure were in the medium to large range, effects of this size were not significant and robust across all processing and evaluation measures.

Relationships Among Learning, Processing, and Evaluation Measures

Given that the writing manipulation had an impact on a few of the dependent measures, the relationships among learning outcomes, processing measures, and evaluation measures were further examined to better understand what behaviors were related to better learning on this task. Correlations among variables are shown in Table 4. Most of the correlations were Pearson's *r* values, whereas model type, navigation patterns, and peer essay evaluations were analyzed using Spearman's ρ. One important observation is that each of our three groups of measures showed internal coherence. The five learning measures showed relationships with one another. Total correct on the recognition posttest was positively related to *d'* values on the posttest. These measures were also positively related to the number of correct causal concepts mentioned in essays and to coding of the essays in terms of eruption models. All four of these measures were negatively related to the number of erroneous causes included in essays. Similarly, the three processing measures, relative time spent on reliable versus unreliable pages, returns to reliable sites, and navigation patterns, all showed significant positive correlations with one another. Furthermore, the three evaluation measures were also related to one another. Better discrimination between reliable and unreliable sources in the rankings (a larger difference between the average ranking of reliable and unreliable sites) was positively related to the number of times that an evaluation of the source was explicitly mentioned in justification of the rankings and positively related to the explicit rejection of an oil-drilling account of volcanic eruptions. No significant correlation was found between the ranking of the Volcano Live site with any other evaluation, processing or learning measure (all *r* values < .11). This last evaluation measure is not included in the table or in the composite factor scores given below.

Given the patterns of relations among the sets of correlated measures, a composite score was created for each by entering each set into a separate factor analysis and extracting a single composite factor score prior to rotation. Each derived composite factor had an eigenvalue > 1. As shown in Table 5, pretest scores, the evaluation composite score, and the processing composite score were all correlated with the learning composite score.

To determine if evaluation and processing behaviors had unique effects on learning that were also independent from prior knowledge as assessed by the pretest, we computed two hierarchical regressions, entering pretest scores in the first step and varying the orders of evaluation and processing in later steps. Pretest scores alone accounted for 14% of the variance in learning and the model was significant, $F(1, 101) = 16.8$, $MSE = 0.85$, $p < .0001$. When processing scores were entered in the next step, they accounted for a significant amount of additional variance in model fit, R^2 change = .09, $F(1, 100) = 12.1$, $p < .0001$. When evaluation scores were entered in a final step, they added yet another increment of significant additional variance, R^2 change = .10, $F(1, 99) = 15.2$, $p < .0001$. When evaluation scores were

Table 4

Relations Between Learning, Processing, and Evaluation Measures in Experiment 1 ($n = 110$)

Variable	Correct Causes in Essay	Erroneous Causes in Essay	Model of Eruptions	Time on Reliable/ Unreliable	Returns to Reliable Pages	Navigation Patterns	Ranking Difference Score	References to Source in Justification ($n = 104$)	Evaluation of Peer Essay ($n = 104$)
Posttest score	.211*	−.244*	.267**	.383**	.037	.172	.269**	.159	.416**
d' on posttest	.270**	−.224*	.305**	.422**	.133	.109	.244**	.149	.384**
Correct causes		−.276**	.723**	.214*	.045	.214*	.291**	.280**	.279**
Erroneous causes			−.207*	−.372**	−.028	−.358**	−.207*	−.100	−.233*
Model of eruptions				.274**	−.020	.161	.293**	.220*	.178
Time reliable/ unreliable					.366**	.374**	.226*	.103	.191*
Returns to reliable						.348**	.219*	.089	.210*
Navigation patterns							.282**	.095	.056
Ranking discrimination								.363**	.349**
References to source									.263**

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

Table 5

Correlations Among Composite Factors and Pretest Scores for Experiment 1

	Processing	Evaluation	Learning
Pretest	.10	.17	.37**
Processing		.24*	.35**
Evaluation			.44**

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

entered followed by processing scores, there were significant R^2 changes of .14, $F(1, 100 = 20.2, p < .0001$, and .05, $F(1, 100) = 7.4, p < .0001$, respectively. These results show that both processing and evaluation behaviors predicted learning outcomes, even after taking the effects of prior knowledge into account. This is an important finding because it shows that it was not simply differences in prior knowledge that led to the more effective processing and evaluation behaviors. The way a student processed the information and engaged in evaluation both had independent influences on learning resulting from the inquiry activity.

Comparing the results across these two orders of entry, in both cases, larger unique effects can be seen for evaluation behaviors. Standardized β weights in the final step also suggest that evaluation scores explained the most independent variance in this model (evaluation β = .33, pretests β = .30, and processing β = .23). This finding indicates that source evaluation is especially important for learning outcomes and indicates that the creation of strong intertext models is related to better comprehension in multiple-source inquiry tasks.

Qualities of Successful Learners

The above analyses suggest that evaluation is critical for better learning from multiple sources. To provide a more descriptive sense of the quality of the intertext models that were created, a final set of analyses contrasted the evaluation behaviors of more and less successful participants specifically in relation to their reliability rankings and justifications. More successful students ($n = 20$) were operationally defined as those who included more than four correct concepts in their essays and no erroneous causes (this criterion was based on the average performance on the essay task). Less successful students ($n = 20$) included more than one erroneous cause and fewer than four correct concepts in their essays.

As would be expected from the regression analysis, the results of independent-samples *t* tests (Table 6) showed that more successful students ranked the "reliable" sites as more reliable and the "unreliable" sites as less reliable than the less successful students, making the difference between the rankings greater for more successful students than for the less successful students. In their justifications for their rankings of reliability for each site,

Table 6
Comparison of More and Less Successful Students in Experiment 1

	More Successful	Less Successful	η^2
Source rankings	*M* (*SD*)	*M* (*SD*)	
Ranking on reliable pages**	2.75 (.65)	3.63 (.86)	.27
Ranking on unreliable pages**	5.25 (.76)	4.35 (.67)	.29
Ranking discrimination**	2.70 (1.62)	0.80 (2.38)	.31
Volcano Live site ranking	4.30 (1.34)	4.25 (1.71)	
Categories for reliability justifications	% (condition with frequency)	% (condition with frequency)	
References to source in justifications*	10.1% (20)	4.3% (6)	.11
References to evidence in justifications	12.6% (25)	12.1% (17)	
References to agreement in justifications	7.0% (14)	3.6% (5)	
References to explanation in justifications**	17.6% (35)	10.7% (15)	.19
References to relevance in justifications	25.1% (50)	43.6% (61)	
References to liking/ interest in justifications	3.2% (5)	5.2% (17)	
References to style of site in justifications	6.0% (12)	7.9% (11)	
Other	19.6% (39)	15.0% (21)	

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

more successful students were more likely to mention source information about sites. They also were more likely to make rankings on the basis of whether a site explained why or how volcanoes erupt. However, the most common basis for evaluation among both more and less successful students was the relevance of the material, specifically the presence of information about Mt. St. Helens. This suggests the use of a keyword-matching heuristic as a basis for evaluation. Although there were some references to the presence of evidence, there were relatively few references to the quality of the evidence presented in the sites in either more or less successful learners. In particular, it was rare that students spontaneously mentioned the corroboration or convergence of evidence across sites as a basis for their reliability judgments. Only six students in the contrastive sample explicitly mentioned this strategy for assessing reliability (four were more successful learners and two were less successful). In most cases, students simply stated that information "kinda went along with," "matched," or had the "same" information as another site, without explicating why this relationship might be important. One successful learner rated the NASA site as high in reliability because it allowed the learner "to apply what they learned in the Savage Earth site" (the PBS site). However, another successful learner explicitly

noted the repetition of information across the NASA, *Scientific American*, and Savage Earth sites, and used this to justify low assessments of reliability for the latter two sites "because the information was redundant and didn't add anything."

In addition, the large number of paraphrases and vague statements represented by the "other" category indicates that students had difficulty articulating their reasons for judging the reliability of Web sites. Even the more successful learners, who were able to discriminate good from poor sources, failed to look for converging information across sources and did not seem to be able to appreciate the value of corroboration. This latter point also seems related to the overall lack of sensitivity to the reliability of the information on the Volcano Live site. These results suggest that evaluative information related to the reliability of the multiple sources was not well represented in students' intertext models.

Conclusions From Experiment 1

In Experiment 1, we tested whether a writing task manipulation would affect the development of intertext models during the completion of an Internet-based science inquiry activity involving multiple information sources and, more generally, how intertext models were related to learning from multiple sources. The results indicated a few medium to large effects of the argument writing manipulation on learning, but also some nonsignificant effects. Although there were some tendencies toward improvement in some evaluation and processing behaviors, the argument writing task did not improve the ability to discriminate between reliable and unreliable sources, nor did it prompt selective rereading and comparison of information across sites. This suggests that the argument writing task alone did not promote the creation of especially strong intertext models.

There was some evidence that the students who engaged in this inquiry task included a greater number of correct causal concepts about volcanic eruptions included in their essays over students in a no-reading comparison condition. However, there was also evidence that students in the inquiry task also acquired some erroneous conceptions from the unreliable texts. This result stresses the importance of students' developing intertext models of sources to help them discriminate reliable from unreliable Internet sites during inquiry learning tasks. Otherwise, students may "learn" information from Internet inquiry tasks that does not improve their understanding of a correct causal model of scientific phenomena. Furthermore, the presence of so few Type 3 models suggests that few students were moving beyond the representation of the documents in independent situation models to integrate their understanding into a single coherent situations model.

On the other hand, when looking across the sample as a whole, better learning was associated with better evaluation behaviors. Although only about 15% of the students constructed Type 3, coherent situations models of volcanic eruptions, the regression analyses indicated that it was these

students who tended to allocate more of their processing to reliable information than unreliable information and who tended to make greater discriminations in their evaluations of reliable and unreliable sources. Both regression and contrastive analyses confirmed that there was a relation between stronger intertext models and learning.

That said, it is noteworthy that even among the best learners in this sample, the level of critical analysis of the information across sites was relatively impoverished. Few students went much beyond looking for the verbatim answer to the question that was posed, and students selected sites primarily on the basis of the presence of the phrase "Mt. St. Helens." Others took the approach of trying to understand why eruptions happen and simply searched for a good explanation in the information sources. Even the most successful learners in this experiment seemed to have a fragile understanding of how to judge the quality of information: Few used notions of source or evidence quality to justify their evaluations of reliability. Few noted the value of corroborating information across sources or actively compared information across reliable sources. So, although these better students may have had relatively better intertext models of the sources than their peers, these models still were not well developed. These findings attest to the importance of giving greater prominence to the American Association for the Advancement of Science (1993) benchmark related to understanding sources of bias and how it may influence evidence.

Thus, the findings of Experiment 1 indicate that few students engaged in behaviors that would afford them deep comprehension of the material from across the sources, with a notable lack of development in intertext models. In Experiment 2, we tested whether explicit instruction in evaluating the quality of information could improve the quality of intertext models and, as a result, lead to more effective learning during Internet inquiry tasks in which students need to choose among a selection of reliable and unreliable sources.

Experiment 2: The Effect of Evaluation Instruction on Learning From Inquiry Tasks

The primary goal of Experiment 2 was to investigate if instructing students how to evaluate the reliability and usefulness of information on Web sites would positively affect their comprehension and learning from an Internet inquiry task. The results of Experiment 1 are consistent with prior research indicating that students do not seem to have a stable or coherent understanding of how they should evaluate sources (Brem et al., 2001; Britt & Aglinskas, 2002). Other studies have found that simply prompting students to evaluate sources is not enough to lead to improvements in learning outcomes (Rouet et al., 1997; Stadtler & Bromme, 2007).

One possible reason for both of these kinds of results is that the typical undergraduate may not have ever received instruction in source evaluation. To get a sense of this, we conducted a survey of 56 undergraduates in the introductory psychology subject pool. We asked, "Have you ever received

instruction in assessing Web site reliability, and if so, in what context?" and "When doing research, how do you determine if a Web site is reliable?" A majority, 79% (44 of 56), of students indicated that they had never received instruction in assessing the reliability of an Internet source. The 21% (12 of 56) who had received instruction indicated that they learned it "sometime in high school" ($n = 4$), in English courses in high school and/or college ($n = 7$), or in a foundations of computer applications course ($n = 1$). The students who had prior instruction all mentioned appropriate strategies for evaluating the reliability of information sources found on the Internet. They reported considering the author's identity and credentials, the reputation of the source, or accreditation of the source by institutional affiliation. Of the 44 students who had not had instruction, 30 reported strategies similar to those who had instruction (e.g., author and source credentials) but also mentioned using suggestions of friends. The remaining 14 reported relying on surface characteristics of sites, such as their layout and comprehensibility, when deciding which sites to use for their research. Only a very small number of students (7 of 56 [13%]) mentioned the importance of corroborating information or gathering converging evidence from comparison across sources. These data indicate that most students probably have some sense of where to begin evaluating sources, but for the most part, they lack an understanding of how to assess the quality of the information they find on the Internet. Our survey results are consistent with previous research (e.g., Korpan et al., 1997) and suggest that students need opportunities to learn a more scientifically based understanding of reliability.

Previous efforts to provide students with opportunities to learn about or scaffold source evaluation have met with moderate success. For example, Britt, Perfetti, Van Dyke, and Gabrys (2000) developed the Sourcer's Apprentice environment to support student learning from multiple sources in history. The main scaffold in this environment is an on-screen note-taking facility that requires students to make entries about authorship and the evidence available within each source as students read a set of texts. Sourcer's Apprentice has been shown to improve skills of sourcing, contextualization, and corroboration, as well as learning from primary and secondary sources in history (Britt & Aglinskas, 2002). In science content areas, the Web-based Inquiry Science Environment has used online peer discussions and "recommended site" discussion boards to support student evaluation of Web sites used for inquiry activities (Slotta & Linn, 2000). In the context of conducting Internet searches for health-related information, Stadtler and Bromme (2007) created met.a.ware, which prompts readers to evaluate the sources that they encounter. The instructional environment we developed and tested in this experiment is similar to these efforts but was more explicitly structured to support the use of source evaluation strategies outside of the instructed context. The instructional unit developed for Experiment 2 involved the presentation of declarative knowledge of what to consider when evaluating sources, multiple opportunities to apply that information, and feedback on in the form of expert evaluations. Opportunities to apply

knowledge in a variety of situations and to receive feedback on these efforts are features of instructional environments that have been shown to support the acquisition of more flexible knowledge and transfer to new learning situations (cf. Anderson, Corbett, Koedinger, & Pelletier, 1995; Bransford et al., 2000; Pashler et al., 2007). Experiment 2 explicitly tested whether participating in these instructional activities would transfer to a new inquiry task and produce positive effects on inquiry learning in a content area other than the one used in the instruction.

On the basis of prior research and the results of Experiment 1, we identified four key aspects of source evaluation that students needed to learn to systematically consider, identified by the acronym SEEK: (a) the Source of the information, (b) the nature of the Evidence that was presented, (c) the fit of the evidence into an Explanation of the phenomena, and (d) the fit of the new information with prior Knowledge. This formed the basis of the declarative information provided to students.

The context for the SEEK instruction and the comparison group was a task in which students were to ascertain which sources from among a set of six would be the best for deciding whether the Atkins low-carbohydrate diet is healthy or harmful. As in the volcano unit, the sites varied in their reliability. The SEEK instructional materials provided a description of how to consider and use each of these aspects to evaluate sources of information (see Appendix B for details of the declarative information that was provided). Students in the SEEK condition were also shown expert rankings of the reliability of the sources once they had made their own evaluations. The comparison group read and evaluated the same set of sources that were used in the SEEK instruction but did not receive either the declarative instruction or the expert evaluations.

The main hypothesis of Experiment 2 was that students who had an opportunity to learn to evaluate sources using the SEEK instructional unit would make better evaluations of the sources used in the instructional setting and, more important, would be better able than a comparison group to use effective evaluation strategies in a new content area and inquiry task. Theoretically, we predicted that the SEEK instruction group would be more likely to construct better intertext models of the sources, which should lead to improvements in their comprehension and learning.

Participants in both the SEEK instruction and the comparison groups returned for a second session, during which they participated in the Mt. St. Helens Internet inquiry task from Experiment 1. As in Experiment 1, we collected navigation logs, measures of learning, and reliability rankings and justifications. We expected that the SEEK instructional group would be more likely than the comparison group to differentially devote processing time to reliable sites over unreliable, show greater differentiation in rankings between reliable and unreliable sources, and, as a result, develop deeper understanding of a causal model of volcanic eruptions.

Method

Participants

Participants were 60 undergraduates from an introductory psychology subject pool. Half were randomly assigned to the SEEK instruction and the other half to the comparison group. The mean age of participants was 19.12 years (*SD* = 1.09 years), and 50% were female. The average number of years in college was 1.67 (*SD* = 0.88 years). By class, 35 participants were freshmen, 17 were sophomores, 5 were juniors, and 3 were seniors. The average number of college science courses taken was 1.89 (*SD* = 1.90), including courses in which students were currently enrolled. The average number of earth science courses taken was 0.37 (*SD* = 0.49). There were no significant differences between the instructional and comparison groups on these variables.

Prior knowledge as measured by the volcano concept pretest was 18.40 (*SD* = 2.81) and did not vary across groups ($F < 1$).

Note also that no selection criterion was used in this experiment, whereas Experiment 1 had selected for only low-knowledge participants.

Materials

Information sources. Students were told that we did a search on Google using the phrase "low carbohydrate diets" and that they were being presented with the top six hits (the first page of results) for this search. The sources were presented on what appeared to be a Google search results page, with the page titles listed as hotlinks, with original uniform resource locators, and with short descriptions of the contents of the pages. The six sites (in order of intended reliability) were (a) the *Journal of the American Medical Association* (*JAMA*; specifically, a published article reporting a controlled, randomized trial); (b) *BusinessWeek* (http://www.businessweek.com; specifically, an article describing the *JAMA* study); (c) Atkins Diet Alert (http://www.atkinsdietalert.org), a not-for-profit organization of vegetarian physicians providing information for laypeople with questions and concerns about high-protein diets; (d) WeightWatchers (http://www.weightwatchers.com; specifically, a page on "the truth about carbs"); (e) the official Atkins site (http://www.atkins.com); and (f) a personal home page that featured a testimonial on the effectiveness of the Atkins diet.

SEEK declarative materials. The first part of the SEEK instructional unit consisted of a three-page description of the instructions for the session along with declarative information about which aspects of sources to consider when evaluating the reliability of a site (see Appendix B). It began as follows:

> Making a decision about which sites to use depends on evaluating how reliable or trustworthy they are. There are several things to

consider in evaluating the usefulness of a site. *Who is the author? How reliable is the information? How well does the site explain the information?*

For each of these criteria, procedures for answering the questions were then provided.

SEEK evaluation template. In addition to the descriptive information, the SEEK unit included opportunities for students to apply the declarative information to the evaluation of the six Web sites on the low-carbohydrate diet topic. They used the template of questions shown in Appendix C to guide their application. The seven questions in the template corresponded to the criteria or evaluation process that was included in the initial descriptive information.

SEEK expert feedback. The third aspect of the SEEK instructional unit was feedback on the evaluation of each of the six Web sites. The feedback consisted of a rank ordering of the sites that was attributed to 10 experts, listed right next to the rank ordering provided by the participant. An example of this feedback is included in Appendix D. As shown in Appendix D, the feedback also included four questions that were designed to focus the participants on critical dimensions of reliability evaluation. The expert rankings were the same across all SEEK participants but the participant rankings were those of the specific participant.

Volcano unit. The materials for the Mt. St. Helens Internet inquiry task were identical to those used in Experiment 1, including all of the information sources and all of the learning assessments and instructions.

Procedure

The experiment consisted of two separate sessions. During the first session, all participants worked in a computer classroom on information sources about low-carbohydrate diets. This first session took approximately 1 hour to complete for both groups. Participants then returned to the classroom for a second session 2 to 7 days after the first session, during which they participated in the same Mt. St. Helens inquiry task used in Experiment 1. Importantly, each participant in the experimental group was yoked to a participant in the comparison group in terms of the amount of time between the first and second sessions between groups. Thus, the overall amount of time between the first and second sessions was matched for the SEEK and comparison groups.

The volcano concept recognition pretest from Experiment 1 was administered to all participants at the beginning of the first session.

For the first session, all students were told that their goal was to ascertain which sites would be the best for deciding whether the Atkins low-carbohydrate diet is healthy or harmful. Students in the SEEK condition were provided with the declarative materials on determining source reliability. These instructions indicated that they were to read through each source and that there were also several things to consider when making their decisions about which sites would be best. They were also instructed to use the information in these declarative materials when asked to complete a template for each site. Participants were then asked to read through all of the sites and also complete an evaluation template for each site while reading. After students were done reading through all the sites, students then ranked each site in terms of its reliability (with a rank of 1 meaning *most reliable*). After all rankings were complete, students were prompted to justify their rankings ("You ranked this site as number one, please explain why").

After students in the SEEK condition justified all their rankings, they were then given feedback on their site rankings. Students were presented with the rankings of 10 hypothetical experts and were asked to compare their ratings to those of the experts.

Students in the SEEK instruction condition had access to their templates and the descriptive information throughout the entire first session.

The comparison group was given the same task of deciding which sources would be best to decide if the Atkins diet is healthy or harmful and had access to the same six Web sites as the SEEK group. However, the major difference between the groups was that the comparison group was not provided with the declarative information, the template for evaluating sites and sources, or the expert rankings. The comparison group was instead instructed as follows:

> People frequently consult sources they find on the World Wide Web to get information. As you know a web search often brings up multiple sites that you could look at. This study is about how you decide which sites to use. Which sites to use depends on what you want to do. In this study we want you to decide which sites are the best for deciding whether the Atkins low carbohydrate diet is a healthy or harmful diet. We have given you 6 websites that were the output of a Google search we did on "low carbohydrate diets." We want you to read through each one in order to decide which sites you think would be most useful for this task, and will help you to understand whether or not a low carbohydrate diet is healthy or harmful.

The comparison group then read through the Web sites and completed the rankings. They were permitted to take notes and could use these notes while they completed their site rankings and the justifications, just as the SEEK group could use their evaluation templates. The procedure during the second session was the same for all participants: They each participated in the argument version of the volcano inquiry task of Experiment 1.

Results and Discussion

Atkins Inquiry Task Reliability Rankings

All students, regardless of condition, ranked the *JAMA* site as the most reliable and the *BusinessWeek* site as the second most reliable. Both groups ranked the personal testimony site as the least reliable. Thus, all students seemed to have some basic ability to recognize the sources with the most authenticity or credibility (*JAMA*) and the most subjective information (personal testimony). However, the average rankings of the reliable and unreliable sources were more extreme for the SEEK instruction group. On average, for the SEEK instruction group, the average median of the rankings was 2.33 for reliable sources and 4.40 for the unreliable sources, while for the comparison group, the average median of the rankings was 2.83 for reliable sources and 4.10 for unreliable sources. (Lower scores indicate higher reliability.) The difference between reliable and unreliable rankings was significantly larger for the SEEK instruction group ($M = 2.07$, $SD = 1.34$) than for the comparison group ($M = 1.27$, SD = 1.76), $t(58) = 2.10$, $p < .04$, $\eta^2 = .07$. This effect demonstrates that the SEEK instruction had the predicted effect on reliability judgments for sites used in the reliability instruction unit (i.e., the Atkins diet).

Transfer to the Mt. St. Helens Inquiry Task

The SEEK instructional materials were present only during the first session. Therefore, performance on the Mt. St. Helens inquiry task tested whether skills learned from the SEEK instruction would transfer to a new content domain and new inquiry task context.

A MANCOVA with pretest scores entered as a covariate revealed a significant main effect for SEEK instruction, $F(10, 48) = 3.10$, $p < .01$, $\eta^2 = .39$. The follow-up analyses for each measure are presented in Table 7. As predicted, SEEK instruction improved performance on all evaluation and learning outcome measures. No differences were seen in the processing measures.

In addition, the effects of the SEEK instruction were also examined for several measures requiring nonparametric tests (model type, navigation patterns, and corroboration).

The distributions of conceptual models of volcanic activity reflected in the student essays are presented in the rightmost columns of Table 1 for the SEEK and comparison groups. The distribution for the comparison group was fairly even across the four models, with the majority of essays coded as Type 1 (single cause). In comparison, the majority of essays produced by those in the SEEK group were coded as Type 3. The differences in these distributions was related to whether students had participated in SEEK instruction during the first session, $\chi^2(3) = 12.4$, $p < .006$, $\varphi = .46$.

Table 7
Results of Experiment 2 Multivariate Analysis of Covariance and Follow-Up Tests

	Main Effect of SEEK Instruction
Core causes	SEEK > NO ($\eta^2 = .06$)
Erroneous causes	SEEK < NO ($\eta^2 = .07$)
Posttest scores	SEEK > NO ($\eta^2 = .09$)
d' on posttest	SEEK > NO ($\eta^2 = .08$)
Time on reliable pages	—
Time on unreliable pages	—
Returns to reliable pages	—
Ranking discrimination	SEEK > NO ($\eta^2 = .12$)
Volcano Live site ranking	SEEK < NO ($\eta^2 = .08$) (site seen as more reliable)
References to source	SEEK > NO ($\eta^2 = .10$)

Note. SEEK = source, evidence, explanation, knowledge (see text); NO = no instruction. Only significant effects are reported.

The distribution of navigation patterns for the two groups, shown in Table 2, indicates that the comparison group was more likely to reread the sources either nonselectively or with a bias for returning to unreliable sites. In contrast, those who had participated in the SEEK instruction were more likely to selectively reread the reliable sources. The distribution of navigation patterns was significantly different because of SEEK instruction, $\chi^2(3) = 12.98$, $p < .01$, $\varphi = .49$.

Finally, the presence of comments made specifically about the overlap of evidence and information across sites was analyzed. In total, 18 students made at least one comment about corroboration or the consistency of information across sites in justification of their source reliability rankings. Twelve of these were in the SEEK group, and 6 were in the comparison group. This distribution approached significance, $\chi^2(1) = 2.86$, $p < .09$, $\varphi = .21$.

Conclusions From Experiment 2

The results of Experiment 2 demonstrate the effectiveness of the SEEK instructional unit on improving the comprehension of scientific phenomena from Internet research tasks. Instructing students on the importance of considering the source of information, evaluating the evidence and explanations that are provided, and relating the new information to prior knowledge and among sites in the context of an Internet research task on the Atkins diet led to better source evaluation skills. Furthermore, these evaluation skills carried over to a new Internet inquiry task on the causes of the eruption of Mt. St. Helens. The evidence that students in the SEEK group had better evaluation skills comes from their better discrimination

between reliable and unreliable sites in their rankings, as well as their better ability to articulate their reasons for their evaluations in terms of reliability and corroborating evidence. These results demonstrate that the SEEK instruction led to the construction of better intertext models during a subsequent inquiry task. Furthermore, with improved understanding of how to evaluate the quality of Web sites, students who received SEEK instruction showed more successful learning from the subsequent Web-based inquiry task. They were more likely to engage in strategic rereading of reliable sources and more likely to learn correct concepts. Furthermore, students in the SEEK instruction condition included fewer erroneous causes in their final essays than the comparison group and were more likely to have integrated causal accounts (Type 3 essays), suggesting that they created more accurate and coherent situations models. Without SEEK instruction, the comparison group was more likely to mention erroneous causes, suggesting, as in Experiment 1, that without the ability to discriminate between reliable and unreliable information, some students were failing to create accurate intertext models or coherent situations models of the sources, resulting in the acquisition of incorrect information from the unreliable sites during the inquiry task. Thus, the results of this experiment demonstrate the critical role that intertext models play in effective multiple-source comprehension.

General Discussion

Over the past 15 years, there has been a growing awareness of the importance of reading and writing in science and of the need to provide opportunities for students to learn the literacies of science (Airey & Linder, 2008; Goldman & Bisanz, 2002; Holliday, Yore, & Alverman, 1994; Otero et al., 2002; Yore et al., 2003). All individuals, whether they are practicing scientists or not, need a level of science literacy that allows them to participate in public discourse and debate about current issues and controversies in science. They need to be able to evaluate purported scientific information relevant to personal decision making, especially in areas of physical and mental health. Such science literacy includes knowledge of the norms and genres of argumentation in science and the criteria that must be met for information to be deemed scientifically reliable (American Association for the Advancement of Science, 1993). The accessibility of information on the World Wide Web and the lack of review or regulation of the content means that it is now even more important that people possess the knowledge and skills to distinguish between content that meets criteria for reliable scientific findings and content that does not.

As noted earlier, prior research has shown that readers generally find it very difficult to determine the quality of information, especially when they lack knowledge of the topic. Thus, the present results expand our knowledge of important dimensions for learning from secondhand investigations (Palincsar & Magnusson, 2001) such as Internet-based inquiry activities.

Although the present research examined college students with relatively low domain knowledge about volcanic eruptions, the results of Experiment 1 revealed that they could distinguish the reliability of sources found on the Web to some extent. However, few students could articulate the reasons for their evaluations of reliability, and most appeared to have no particularly systematic way of evaluating the reliability of the Web sites. In general, readers' intertext models seemed fairly impoverished during the comprehension of multiple sources. Furthermore, an argument writing task only moderately supported better attention to the reliability or quality of sources. There was still much room for improvement in terms of the consistency of effects across measures.

However, an analysis of the navigation and evaluation data in relation to learning outcomes suggested that some aspects of intertext models may have been in place for the more successful learners. In particular, successful learning was related to greater discrimination in rankings between more and less reliable sites, spending a larger proportion of time reading the more reliable sites than the less reliable sites, and a selective focus on reliable sites, especially during rereading. What is unclear from the present set of studies is why the more successful learners behaved in this way. That is, we cannot tell from the present data what the basis of the more successful learners' decision making was. How did they know what were the more reliable sites? What cues or information were they using? Why did they return to reliable sites more than unreliable and spend more time on them? We are pursuing in-depth analyses of the think-aloud protocols to attempt to address these questions in greater depth.

What is clear from Experiment 1 is that only a small percentage of students seemed to engage in selective processing and evaluation spontaneously. Without explicit support, the majority of the students did not engage in many of the behaviors that are part of an ideal self-regulated learning process, consistent with previous findings in the reading-to-learn literature (Azevedo & Cromley, 2004; Graesser et al., 2007; Pressley, 2002; Thiede et al., 2003; Winne, 2001). Although all students in this study were provided with an inquiry question that asked for a causal explanation, many students approached this question on a very superficial level and selected sites primarily on the basis of their keyword relevance. Even when students did engage in rereading and source comparison processes, only a small portion of students were focused on selectively integrating just the reliable information into a coherent situations model. These results show how the lack of a well-developed intertext model can impair comprehension in a multiple-source context.

Interventions That Support Future Learning

The critical, causal relation between the quality of intertext models and multiple-source comprehension was demonstrated even more clearly in Experiment 2, in which explicit instruction on evaluating source reliability

was manipulated. When students had experienced the SEEK instructional unit in the context of evaluating sources about the Atkins diet, they were more able to engage in the target behaviors of constructing intertext models of the sources in a new inquiry context, which in turn led to more effective comprehension.

We found it very encouraging that the relatively brief instructional activity used in Experiment 2 to support source evaluation skills produced noticeable learning gains in an inquiry task in a new domain that occurred several days later. We believe there are several reasons the SEEK instructional unit produced a positive effect on learning in a new domain. First, SEEK instruction integrated prompts and scaffolds used in previous research to construct an ordered series of evaluative dimensions. These dimensions were realized as a set of questions about the source, the evidence, the explanation, and the fit of the explanation with prior knowledge. Previous research has explored some of the dimensions used in the SEEK unit (Brem et al., 2001; Britt & Aglinskas, 2002; Rouet, 2006; Slotta & Linn, 2000), but the SEEK condition consolidated them and tested for the effects of such instruction on learning outcomes from a new unit.

Second, the SEEK instructional unit also provided some explanation of how to go about answering the questions, emphasizing why the dimension was important and how to arrive at an answer to each of the specific SEEK questions. This approach is consistent with previous research showing that inducing the use of new strategies is facilitated by explaining why the new strategy is important or might be effective (Paris & Winograd, 1990; Pressley, 2002). Without such explanation, there is the risk that learners are left with procedures that are not causally connected to desired outcomes or their uses. Third, the application of the declarative information to the task of making reliability rankings, in combination with feedback on the rankings, may have led to the transfer of skills that was observed (Anderson, Reder, & Simon, 1996; Bransford et al., 2000). Learners used the set of questions to determine the most reliable and relevant sites for making a decision about the target domain during the Atkins diet task. Having done so, they were given feedback on their decisions in the form of expert decisions and the reasoning that led to those decisions. Hence, the SEEK instructional unit provided several key features of robust learning that we think account for the fact that learners who participated in the intervention were able to transfer their new source evaluation skills to a new inquiry context, which in turn supported better learning in a new content domain.

The collection of information across multiple sources is a characteristic of authentic research in many disciplines beyond science. Historians use multiple sources to corroborate accounts. In literary studies, analysts draw on close readings of multiple bits of information within and across texts. In educational research, researchers draw on reports of previous experimental studies to inform theories and design new empirical tests. Yet often,

instruction in these disciplines does not explore how members of the disciplines use, evaluation, and integrate information across sources. With respect to science, what students learn in science class is often the product of scientific studies, not the process. Presenting science as facts and not as a research process distorts the process and robs the discipline of much of its inherent interest. Furthermore, when only textbook accounts are presented, students do not get a full appreciation that science is about reasoning from evidence and that the real business of science is one of testing among various competing explanations. One reason for the growing popularity of Internet inquiry learning tasks is that they resemble this process of real scientific activity of reasoning from evidence and have constructive integration across information sources as their goal.

As inquiry-based approaches have become more popular, the question of how and when these approaches may lead to effective learning has become more important. In the present study, we investigated learning from an inquiry task that required integrating diverse bits of information about the earth and its dynamic crustal cycle from across several sites. If students engaged in the constructive activities of assessing the reliability of the sources, and integrating information across the reliable sources, then a coherent model of volcanic eruptions could result. Although there was an overall positive effect of this Internet inquiry task on learning, what may be most important was the significant relation that we observed between learning during an Internet inquiry task and evaluation of information sources. Complex knowledge acquisition occurred more effectively when students had been instructed to evaluate the reliability of sources in a previous learning activity. Of course, the generalizability of empirical studies is always an issue. Although this research used samples that were quite representative of university students at two public institutions, efficacy studies with other samples, particularly younger students or in the context of science courses, would help assess the further applicability of these results.

The present findings suggest that multiple-source Internet inquiry tasks can be effective learning activities but that learners need instruction in critical document evaluation and representation skills before the benefits of such inquiry activities may be realized. Furthermore, beyond their contribution to the effectiveness of inquiry learning in science, these source evaluation skills may be even more important for successful lifelong learning, especially when one considers how adult readers are increasingly relying on the Web as a source of information to inform medical, financial, and other important real-world decisions.

Appendix A
Causal Model of Volcanic Eruptions

The network in Figure A1 has a number of events and states that captures the scientific mental model. Mt. St. Helens is a subduction zone volcano, which means that it is located on a tectonic plate boundary between an oceanic and continental plate, not on a hotspot or in the middle of an ocean. The earth's tectonic plates, which are floating on the mantle, move because of convection currents in the earth's mantle or liquid layers just below the crust. When oceanic and continental plates are pushed together, or converge, one plate (the oceanic plate) "subducts" under the continental plate (see box 3 in Figure A1). This causes the crust from the subducted plate to be pushed down into areas of high temperature, causing the crust to melt (see box 7 in Figure A1). Oceanic plates are made up mainly of basalt, and basalt is lighter than the silica that mainly makes up continental plates. Thus, when oceanic crust subducts and melts below a continental plate, the new magma is more buoyant and less dense than the surrounding magma, and this causes the new magma to rise. As the new magma rises and melts part of the continental plate, the resulting magma becomes more viscous. The magma continues to rise into any weak spots or openings it finds under the earth's surface (see box 10 in Figure A1). These openings become magma chambers. As magma fills the chambers, pressure increases, which is not released until the magma shifts or finds a way to expand, and an explosive eruption occurs.

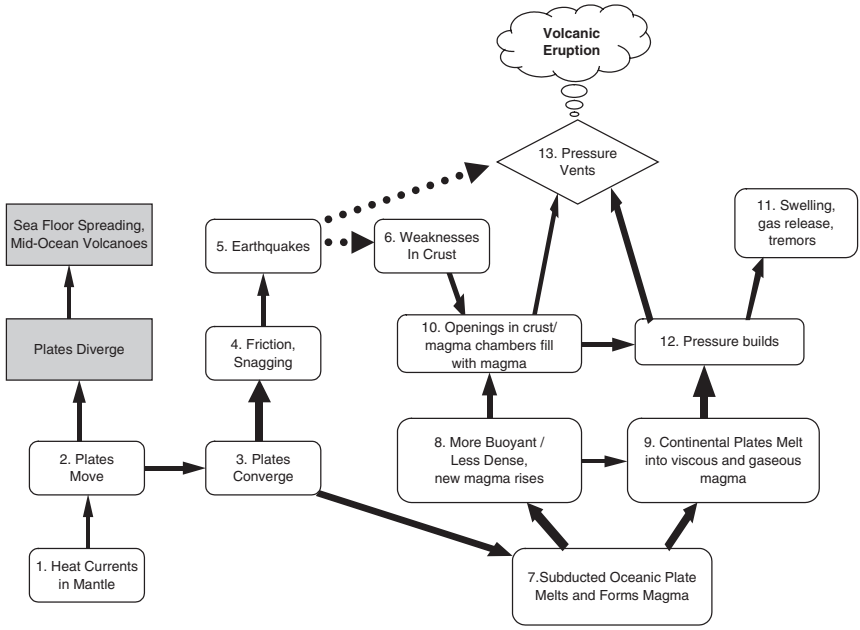

Figure A1. **Network model of causal eruptions.**

(continued)

Appendix A (continued)

In contrast, a second kind of volcano, the midocean volcano, results from diverging oceanic plates. Because midocean volcanoes do not involve continental plates, their magma is highly basaltic and less viscous. As a result, they do not erupt with the violence of subduction zone volcanoes. Thus the location of a volcano at a plate boundary has implications for how magma forms, its chemical composition, and how it reaches the surface.

Earthquake activity can also be indirectly related to volcanic eruptions. The convergence of plates can cause earthquakes if the plates snag and the oceanic plate does not subduct smoothly. Furthermore, earthquakes can cause weak spots in the crust. Earthquakes can also trigger an eruption when a system is already under pressure.

Appendix B
Descriptive Compontent of SEEK Instruction

People frequently consult sources they find on the World Wide Web to get information. As you know a web search often brings up multiple sites that you could look at. This study is about how you decide which sites to use. Which sites to use depends on what you want to do. In this study we want you to decide which sites are the best for deciding whether the Atkins low carbohydrate diet is a healthy or harmful diet. We have given you 6 websites that were the output of a Google search we did on "low carbohydrate diets." We want you to read through each one in order to decide which sites you think would be most useful for this task, and will help you to understand whether or not a low carbohydrate diet is healthy or harmful.

Making a decision about which sites to use depends on evaluating how reliable or trustworthy they are. There are several things to consider in evaluating the reliability of a site. *Who is the author? How reliable is the information? How well does the site explain the information?*

Below are some ideas to help you answer these questions.

Who is the author?

Can you figure out who the author is?

Is the person who is providing the information someone who is knowledgeable about the topic?

You can figure this out from several cues. One is the information provided about the author, what training the person has had, what their current occupation is. Sometimes you can tell this from the institution with which the page or author is affiliated (e.g., National Institute of Health, Fitness Centers Incorporated). Affiliation is sometimes shown in a logo or copyright statement on the page. Finally, the URL (web address) for a site lets you know whether the site is a profit making operation (.com), and educational institution (.edu), a government sponsored site (.gov) or an organization, usually non-profit (.org) or (.net).

(continued)

Appendix B (continued)

What is their motivation?

Knowing something about the author is important because authors often have specific agendas they want to push. Frequently, web sites want to sell readers goods and services or obtain donations from them. To do so, they provide only the part of the information that supports their sales goals. Or they use the site to provide very graphic images that evoke emotional responses. Some sites may have political agendas. As you read the information on the web sites, use information about the author and site to figure out the motivations, possible biases, and purposes that the site author and host might have. You can also determine the motivation of the author by thinking about who the intended audience is.

How reliable is the information?

Is the information based on scientific evidence?

Information that has been gathered through a scientific process can be considered more accurate than personal opinion, beliefs, or anecdotes. Is evidence provided or reported for claims? Are scientific peer reviewed journals cited? Is this information likely to be evaluated well by informed scientists?

Is there similar information given across reliable sources?

If multiple sites or authors give the same information, it is more likely to be accurate than if the sites or authors disagree. This is especially true when the sites with converging information have affiliations that seem trustworthy. If information in a site contradicts other sites that you think are trustworthy, then it suggests the new information might not be reliable. Also consider if the account given is complete, or whether it fails to omit information that other reliable sources mention.

How well does the site explain the information?

Do you understand how the process works based on the information provided?

For a lot of scientific information, another important criteria for a useful site is how well the information on the site explains things.

Does the explanation fit together with your prior scientific knowledge or with information from other reliable sites?

Especially using sites which have affiliations that seem trustworthy, you should examine whether each interpretation of the evidence fits together to generate a coherent explanation of a scientific process or phenomena.

In this packet there are worksheets for each page with these 7 questions about reliability. Please read the first site (Atkins Nutritionals: Home) and fill in your answers to the questions on your first worksheet. Once you have finished filling out these questions, please raise your hand and the research assistant will move you on to the next step.

Appendix C
Example Source Worksheet for SEEK Instruction

Site: http://www.atkins.com/

1. Who is the author?
2. Is the person who is providing the information someone who is knowledgeable about the topic?
3. What is the author's motivation?
4. Is the information based on scientific evidence?
5. Is there similar information given across reliable sources?
6. Do you understand how the process works based on the information provided?
7. Does the explanation fit together with your prior scientific knowledge or with information from other reliable sites?

Appendix D
Feedback Component for SEEK Instruction

Here are your rankings again, and how a survey of 10 experts ranked these pages. Be sure to look at the differences between your rankings and theirs, and think about why they might have evaluated the pages differently.

Experts' Rank	Your Rank	Web Site
5	1	**Atkins** Nutritionals: Home Learn more about why **Atkins** works. Click here. **...** read this article. Dr. **Atkins** Weight: The Truth. The day he fell, Dr. **Atkins** weighed 195 pounds**...** http://www.atkins.com/ - 33k - Feb 28, 2004
2	2	BW Online	April 8, 2003	Count Calories, Not Carbs ... popular low-carbohydrate diets has proved, yet again, that there's no such thing ... **Atkins diets** have spawned a huge industry of books, diet aids, ... www.businessweek.com/technology/content/apr2003/tc2003048_5670_tc024.htm – 79k
4	3	Weightwatchers.com ... no and low-carb **diets**...**Atkins**...make carbs part of your healthy weight loss program... www.weightwatchers.com/truthaboutcarbs.html - 86k
1	4	Efficacy and safety of low- carbohydrate diets: a systematic review... JAMA. 2003 Apr 9;289(14):1837-50. Efficacy and safety of low-carbohydrate **diets** which include the popular **Atkins**... Bravata DM, Sanders L ... http://jama.ama-assn.org/cgi/content/full/289/14/1837 - 42k
6	5	Why you should try the **Atkins diet:** My personal success... ... I read Dr. Atkins' New Diet Revolution and tried his approach.... www.geocities.com/tinayke/atkinsdiet.html - 26k
3	6	**Atkins Diet** Alert / a Physicians Committee for Responsible **...** AtkinsdietAlert.org, for physicians and laypeople with questions and concerns about high-protein, **Atkins**-type **diets** including a registry where **dieters** can **...** www.atkinsdietalert.org/ - 12k - Feb 28, 2004

To help you think about this, please answer the following four questions:

1. Why did the experts rank the testimonial site as least reliable?
2. Why did the experts rank the JAMA site as most reliable?
3. Why was Business Week ranked less reliable than JAMA but more reliable than the Atkins Diet Alert page?
4. Why was the WeightWatchers.com "Truth about Carbs" page ranked less reliable than Business Week but more reliable than Atkins.com?

Note

This project was funded by Grant REC 0126265 ("Understanding in Science") from the National Science Foundation to Jennifer Wiley, Susan R. Goldman, and Arthur C. Graesser. The preparation of thist article was also supported by the Institute for Education Sciences through Grants R305H030170 and R305B07460 to Jennifer Wiley and Grant R305G050091 to Susan R. Goldman. All opinions expressed herein are those of the authors and do not necessarily reflect those of the funding organizations. The data presented here represent a coordinated effort by a large research team. The final stimulus materials for the Mt. St. Helens inquiry task were designed by Ivan K. Ash on the basis of materials developed previously by Jennifer Wiley and Linda Gentry for Grant N000140110339 from the Office of Naval Research to Jennifer Wiley. Data for Experiment 1 were collected with the assistance of Jason Braasch, Cara Jolly, and Tenaha O'Reilly. Melinda Jensen assisted with Experiment 2. Navigation logs and other Web-based data collection scripts were developed by Colleen Kehoe. We also thank Thomas D. Griffin for his comments on this article.

References

Airey, J., & Linder, C. (2008). A disciplinary discourse perspective on university science learning: Achieving fluency in a critical constellation of modes. *Journal of Research in Science Teaching*, *46*, 27–49.

American Association for the Advancement of Science. (1993). *Benchmarks for Science Literacy*. Available at http://www.project2061.org/publications/bsl/default.htm

Anderson, J. R., Corbett, A. T., Koedinger, K., & Pelletier, R. (1995). Cognitive tutors: Lessons learned. *Journal of the Learning Sciences*, *4*, 167–207.

Anderson, J. R., Reder, L. M., & Simon, H. A. (1996). Situated learning and education. *Educational Researcher*, *25*, 5–11.

Andriessen, J., Baker, M., & Suthers, D. (2003). *Arguing to learn: Confronting cognitions in computer-supported collaborative learning environments*. London: Springer-Verlag.

Azevedo, R., & Cromley, J. G. (2004). Does training on self-regulated learning facilitate students' learning with hypermedia? *Journal of Educational Psychology*, *96*, 523–535.

Barrow, L., & Haskins, S. (1996). Earthquake knowledge and experiences of introductory geology students. *Journal of College Science Teaching*, *26*, 143–146.

Bazerman, C. (1985). Physicists reading physics—Schema-laden purposes and purpose-laden schema. *Written Communication*, *2*, 3–23.

Bell, P., & Linn, M. C. (2000). Scientific arguments as learning artifacts: Designing for learning from the Web with KIE. *International Journal of Science Education*, *22*, 797–817.

Berkencotter, C., & Huckin, T. N. (1995). *Genre knowledge in disciplinary communication: Cognition/culture/power*. Hillsdale, NJ: Lawrence Erlbaum.

Bransford, J. D., Brown, A. L., & Cocking, R. R. (2000). *How people learn: Brain, mind, experience, and school*. Washington, DC: National Academy of Sciences.

Brem, S. K., Russell, J., & Weems, L. (2001). Science on the Web: Student evaluations of scientific arguments. *Discourse Processes*, *32*, 191–213.

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

Britt, M. A., Perfetti, C. A., Van Dyke, J., & Gabrys, G. (2000). The Sourcer's Apprentice: A tool for document-supported history instruction. In P. Stearns (Ed.), *Knowing,*

teaching and learning history: National and international perspectives (pp. 437–470). New York: New York University Press.

Brown, A. L., & Campione, J. C. (1996). Psychological theory and the design of innovative learning environments: On procedures, principles, and systems. In L. Schauble & R. Glaser (Eds.), *Innovations in learning: New environments for education* (pp. 289–325). Mahwah, NJ: Lawrence Erlbaum.

Chi, M. T. H. (2000). Self-explaining expository texts: The dual process of generating inferences and repairing mental models. In R. Glaser (Ed.), *Advances in instructional psychology* (pp. 161–238). Mahwah, NJ: Lawrence Erlbaum.

Chi, M. T. H., de Leeuw, N., Chiu, M., & Lavancher, C. (1994). Eliciting self-explanations improves understanding. *Cognitive Science*, *18*, 439–477.

Chinn, C. A., & Malhotra, B. A. (2002). Epistemologically authentic reasoning in schools: A theoretical framework for evaluating inquiry tasks. *Science Education*, *86*, 175–218.

Cohen, J. (1992). A power primer. *Psychological Bulletin, 112*, 155–159.

Coté, N., Goldman, S. R., & Saul, E. U. (1998). Students making sense of informational text: Relations between processing and representation. *Discourse Processes*, *25*, 1–53.

Craig, S. D., Sullins, J., Witherspoon, A., & Gholson, B. (2006). The deep-level reasoning questions effect: The role of dialogue and deep-level reasoning questions during vicarious learning. *Cognition & Instruction*, *24*, 565–591.

Duke, N., & Pearson, P. D. (2002). Effective practices for developing reading comprehension. In A. E. Farstrup & S. J. Samuels (Eds.), *What research has to say about reading instruction* (3rd ed., pp. 205–241). Newark, DE: International Reading Association.

Duschl, R. A., Schweingruber, H. A., & Shouse, A. W. (Eds.). (2007). *Taking science to school: Learning and teaching science in grades K–8*. Washington, DC: National Academies Press.

Garner, R. (1987). *Metacognition and reading comprehension*. Norwood, NJ: Ablex.

Garner, R., & Alexander, P. A. (1994). *Beliefs about text and about instruction with text*. Hillsdale, NJ: Lawrence Erlbaum.

Gobert, J. D. (2000). A typology of causal models for plate tectonics: Inferential power and barriers to understanding. *International Journal of Science Education*, *22*, 937–977.

Goldman, S. R., & Bisanz, G. (2002). Toward a functional analysis of scientific genres. In J. Otero, J. A. León, & A. C. Graesser (Eds.), *The psychology of science text comprehension* (pp. 19–50). Mahwah, NJ: Lawrence Erlbaum.

Goldman, S. R., Duschl, R. A., Ellenbogen, K., Williams, S., & Tzou, C. T. (2003). Science inquiry in a digital age: Possibilities for making thinking visible. In H. van Oostendorp (Ed.), *Cognition in a digital age* (pp. 253–283). Mahwah, NJ: Lawrence Erlbaum.

Graesser, A. C., & Bertus, E. L. (1998). The construction of causal inferences while reading expository texts on science and technology. *Scientific Studies of Reading*, *2*, 247–269.

Graesser, A. C., & Olde, B. (2003). How does one know whether a person understands a device? *Journal of Educational Psychology*, *95*, 524–536.

Graesser, A. C., & Person, N. K. (1994). Question asking during tutoring. *American Educational Research Journal*, *31*, 104–137.

Graesser, A. C., Singer, M., & Trabasso, T. (1994). Constructing inferences during narrative text comprehension. *Psychological Review*, *101*, 371–395.

Graesser, A. C., Wiley, J., Goldman, S. R., O'Reilly, T., Jeon, M., & McDaniel, B. (2007). SEEK Web Tutor: Fostering a critical stance while exploring the causes of volcanic eruption. *Metacognition and Learning*, *2*, 89–105.

Greene, S. (2004). The problems of learning to think like a historian: Writing history in the culture of the classroom. *Educational Psychologist*, *29*, 89–96.

Griffin, T. D., Wiley, J., & Thiede, K. W. (2008). Individual differences, rereading, and self-explanation: Concurrent processing and cue validity as constraints on metacomprehension accuracy. *Memory & Cognition*, *36*, 93–103.

Guthrie, J. T. (1988). Locating information in documents: Examination of a cognitive model. *Reading Research Quarterly*, *23*, 178–199.

Hacker, D. J., Dunlosky, J., & Graesser, A. C. (Eds.). (1998). *Metacognition in educational theory and practice*. Mahwah, NJ: Lawrence Erlbaum.

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

Hofer, B. K. (2004). Epistemological understanding as a metacognitive process: Thinking aloud during online searching. *Educational Psychologist*, *39*, 43–55.

Holliday, W. G., Yore, L. D., & Alverman, D. E. (1994). The reading-science learning-writing connection: Breakthroughs, barriers and promises. *Journal of Research in Science Teaching*, *31*, 877–893.

Janick-Buckner, D. (1997). Getting undergraduates to critically read and discuss primary literature. *Journal of College Science Teaching*, *29*, 29–32.

Jones, S. (2002). *The Internet goes to college: How students are living in the future with today's technology*. Retrieved from http://www.pewinternet.org/pdfs/PIP_College_Report.pdf

King, A. (1999). Discourse patterns for mediating peer learning. In A. M. O'Donnell & A. King (Eds.), *Cognitive perspectives on peer learning* (pp. 87–116). Mahwah, NJ: Lawrence Erlbaum.

Kintsch, W. (1998). *Comprehension: A paradigm for cognition*. Cambridge, UK: Cambridge University Press.

Korpan, C. A, Bisanz, G. L, Bisanz, J., & Henderson, J. M. (1997). Assessing literacy in science: Evaluation of scientific news briefs. *Science & Education*, *81*, 515–532.

Kuhn, D. (1993). Science as argument: Implications for teaching and learning scientific thinking. *Science Education*, *77*, 319–337.

Linn, M. C., Davis, E. A., & Bell, P. (Eds.). (2004). *Internet environments for science education*. Mahwah, NJ: Lawrence Erlbaum.

Magliano, J. P., & Millis, K. K. (2003). Assessing reading skill with a think-aloud procedure. *Cognition & Instruction*, *21*, 251–283.

Marques, L., & Thompson, D. (1997). Misconceptions and conceptual changes concerning continental drift and plate tectonics among Portuguese students aged 16–17. *Research in Science and Technological Education*, *15*, 195–222.

Marschak, S. (2001). *Earth: Portrait of a planet*. New York: W.W. Norton.

McNamara, D. (2004). SERT: Self-explanation reading training. *Discourse Processes*, *38*, 1–30.

McNamara, D., Kintsch, E., Songer, N., & Kintsch, W. (1996). Are good texts always better? Interactions of text coherence, background knowledge, and levels of understanding in learning from text. *Cognition & Instruction*, *14*, 1–43.

Mosenthal, P. (1996). Understanding the strategies of document literacy and their conditions of use. *Journal of Educational Psychology*, *88*, 314–332.

Norris, S. P., Phillips, L. M., & Korpan, C. A. (2003). University students' interpretation of media reports of science and its relationship to background knowledge, interest, and reading difficulty. *Public Understanding of Science*, *12*, 123–145.

O'Reilly, T., & McNamara, D. S. (2007). The impact of science knowledge, reading strategy knowledge on more traditional "high-stakes" measures of high school students' science achievement. *American Educational Research Journal*, *44*, 161–196.

Otero, J., Leon, J. A., & Graesser, A. C. (Eds.). (2002). *The psychology of science text comprehension*. Mahwah, NJ: Lawrence Erlbaum.

Palincsar, A. S., & Brown, A. L. (1984). Reciprocal teaching of comprehension-fostering and comprehension-monitoring activities. *Cognition & Instruction*, *1*, 117–175.

Palincsar, A. S., & Magnusson, S. J. (2001). The interplay of first-hand and second-hand investigations to model and support the development of scientific knowledge and reasoning. In S. M. Carver & D. Klahr (Eds.), *Cognition and instruction: Twenty-five years of progress* (pp. 151–193). Mahwah, NJ: Erlbaum.

Paris, S. G., & Winograd, P. (1990). How metacognition can promote academic learning and instruction. In B. F. Jones & L. Idol (Eds.), *Dimensions of thinking and cognitive instruction* (pp. 15–51). Hillsdale, NJ: Lawrence Erlbaum.

Pashler, H., Bain, P., Bottge, B., Graesser, A., Koedinger, K., McDaniel, M., et al. (2007). *Organizing instruction and study to improve student learning* (NCER 2007-2004). Washington, DC: National Center for Education Research.

Perfetti, C. A., Rouet, J.-F., & Britt, M. A. (1999). Toward a theory of documents representation. In H. van Oostendorp & S. Goldman (Eds.), *The construction of mental representations during reading* (pp. 99–122). Mahwah, NJ: Lawrence Erlbaum.

Pressley, M. (2002). Metacognition and self-regulated comprehension. In A. E. Farstrup & S. Samuels (Eds.), *What research has to say about reading instruction* (pp. 291–309). Newark, DE: International Reading Association.

Pressley, M., & Ghatala, E. S. (1990). Self-regulated learning: Monitoring learning from text. *Educational Psychologist*, *25*, 19–33.

Quintana, C., Zhang, M., & Krajcik, J. (2005). A framework for supporting metacognitive aspects of online inquiry through software-based scaffolding. *Educational Psychologist*, *40*, 235–244.

Rosenshine, B., Meister, C., & Chapman, S. (1996). Teaching students to generate questions: A review of the intervention studies. *Review of Educational Research*, *66*, 181–221.

Rouet, J.-F. (2006). *The skills of document use: From text comprehension to Web-based learning*. Mahwah, NJ: Lawrence Erlbaum.

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 & Instruction*, *15*, 85–106.

Sanchez, C., & Wiley, J. (2006). Effects of working memory capacity on learning from illustrated text. *Memory & Cognition*, *34*, 344–355.

Sandoval, W. A., & Reiser, B. J. (2004). Explanation-driven inquiry: Integrating conceptual and epistemic scaffolds for scientific inquiry. *Science Education*, *88*, 345–372.

Slotta, J. D., & Linn, M. C. (2000). The Knowledge Integration Environment: Helping students use the Internet effectively. In M. Jacobson & R. Kozma (Eds.), *Innovations in science and mathematics education: Advanced designs, for technologies of learning* (pp. 193–226). Mahwah, NJ: Lawrence Erlbaum.

Stadtler, M., & Bromme, R. (2007). Dealing with multiple documents on the WWW: The role of meta-cognition in the formation of documents models. *International Journal of Computer-Supported Collaborative Learning*, *2*, 191–210.

Thiede, K. W., Anderson, M. C. M., & Therriault, D. (2003). Accuracy of metacognitive monitoring affects learning of texts. *Journal of Educational Psychology*, *95*, 66–73.

Thiede, K. W., Griffin, T. D., Wiley, J. & Anderson, M. (in press). Poor metacomprehension accuracy as a result of inappropriate cue use. *Discourse Processes*.

Thiede, K. W., Griffin, T. D., Wiley, J., & Redford, J. (in press). Metacognitive monitoring during and after reading. In D. J. Hacker, J. Dunlosky, & A. C. Graesser (Eds.), *Handbook of metacognition in education*. London: Routledge.

Trabasso, T., & van den Broek, P. (1985). Causal thinking and the representation of narrative events. *Journal of Memory & Language*, *24*, 612–630.

Voss, J. F., & Silfies, L. N. (1996). Learning from history text: The interaction of knowledge and comprehension skill with text structure. *Cognition & Instruction*, 14, 45–68.

Voss, J. F., & Wiley, J. (2006). Expertise in history. In K. A. Ericsson, N. Charness, P. Feltovich, & R. R. Hoffman (Eds.), *The Cambridge handbook of expertise and expert performance* (pp. 569–584). Cambridge, UK: Cambridge University Press.

Wallace, R. M., Kupperman, J., Krajcik, J., & Soloway, Ed. (2000). Science on the Web: Students on-line in a sixth-grade classroom. *Journal of the Learning Sciences*, *9*, 75–104.

White, B. Y., & Frederiksen, J. (2005). A theoretical approach for fostering metacognitive development. *Educational Psychologist*, *40*, 211–224.

Wiley, J. (2001). Supporting understanding through task and browser design. In *Proceedings of the Twenty-Third Annual Conference of the Cognitive Science Society* (pp. 1136–1143). Hillsdale, NJ: Lawrence Erlbaum.

Wiley, J., & Ash, I. (2005) Multimedia learning in history. In R. Mayer (Ed.), *The Cambridge handbook of multimedia learning* (pp. 375–391). Cambridge, UK: Cambridge University Press.

Wiley, J., Griffin, T. D., & Thiede, K. W. (2005) Putting the comprehension in metacomprehension. *Journal of General Psychology*, *132*, 408–428.

Wiley, J., & Myers, J. L. (2003). Availability and accessibility of information and causal inferences from scientific text. *Discourse Processes*, *36*, 109–129.

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.

Wineburg, S. S. (1991). Historical problem solving: A study of the cognitive processes used in the evaluation of documentary and pictorial evidence. *Journal of Educational Psychology*, *83*, 73–87.

Wineburg, S. S. (1994).The cognitive representation of historical texts. In G. Leinhardt, I. L. Beck, & C. Stainton (Eds.), *Teaching and learning in history* (pp. 85–135). Hillsdale, NJ: Lawrence Erlbaum.

Winne, P. H. (2001). Self-regulated learning viewed from models of information processing. In B. Zimmerman & D. Schunk (Eds.), *Self-regulated learning and academic achievement: Theoretical perspectives* (pp. 153–189). Mahwah, NJ: Lawrence Erlbaum.

Wolfe, M. B., & Goldman, S. R. (2005). Relationships between adolescents' text processing and reasoning. *Cognition & Instruction*, *23*, 467–502.

Yarden, A., Brill, G., & Falk, H. (2001). Primary literature as a basis for a high-school biology curriculum. *Journal of Biological Education*, *35*, 190–195.

Yore, L. D., Bisanz, G. L., & Hand, B. M. (2003). Examining the literacy component of science literacy: 25 years of language arts and science research. *International Journal of Science Education*, *25*, 689–725.

Manuscript received January 21, 2008
Final revision received January 5, 2009
Accepted January 13, 2009