Chapter 10
Understanding and Reasoning with Text

M. Anne Britt
Northern Illinois University, USA

Joseph P. Magliano
Northern Illinois University, USA

Katja Wiemer
Northern Illinois University, USA

Patty Wallace
Northern Illinois University, USA

Keith K. Millis
Northern Illinois University, USA

Peter Hastings
DePaul University, USA

ABSTRACT

Consider the assignment that teachers have been giving their students for years: "Write an expository essay on a scientific topic. Example topics may include global warming, human memory, or the spread of infectious diseases. You must have at least three references." The instructor makes it clear that the paper should have a thesis or claim that is supported by evidence. Claims might be that global warming will be disastrous only for some nations, why it is futile to teach mnemonics to young children, or that cell phone use causes cancer. From the perspective of the student (and cognitive psychologists), this assignment is challenging at any grade. The challenge is that the assignment entails a number of complicated and interconnected tasks. For example, reading a research paper requires the reader to make inferences that span sentences and paragraphs (in addition to a whole host of other processes), and to understand the logical and rhetorical structure of the text as a whole. If the paper describes an experiment, the student must additionally understand how to determine whether the data support the conclusion (i.e., the scientific method). In most cases, the student must also integrate the content of several papers (sources) into a coherent structure. This process involves evaluating the credibility of the sources, selecting relevant pieces of information from each, and putting them into a coherent argument structure. No wonder such assignments are met with groans.

DOI: 10.4018/978-1-61350-447-5.ch010

Copyright © 2012, IGI Global. Copying or distributing in print or electronic forms without written permission of IGI Global is prohibited.

INTRODUCTION

At a fundamental level, each of these processes entail reasoning, the process of specifying how one idea logically leads to or supports another. That is, sentences support inferences; data support conclusions; reasons support claims; quality credentials support credibility, and so on. However, reasoning is difficult to teach. Students need practice with individualized feedback, which is not always possible in the classroom. Members of the Discourse and Technology Group at Northern Illinois University are designing applications to help students assess and improve their ability to reason with texts. The applications include assessing reading comprehension strategies (RSAT), enhancing scientific reasoning (CT Tutor and Operation ARIES!), teaching appropriate sourcing and integration skills (SAIF), and improving argument comprehension and evaluation skills (CASE).

One common aspect of all of these applications is that they use relatively simple algorithms to assess students' performance based on their verbal input. By *simple*, we mean approaches that provide reasonable estimates of whether student products reflect attainment of key constructs using the least computationally demanding methods. The primary goal of these projects is *not* to advance the state of the art in Natural Language Processing. On the contrary, we are using the simplest, most computationally feasible techniques we can find that, in concert with cognitive instructional principles and discourse processing theories, enable us to provide effective assessment and feedback for learning. We believe simple methods are possible and appropriate when one can develop *models* of the students' cognitions and the task, and a range of student products that should underlie the key constructs in the student and task models (Mislevy, 1993; Pellegrino & Chudowsky, 2003; Pellegrino, Chudowsky, & Glaser, 2001). The objective of this chapter is to describe the methods used by each of these applications along with a discussion of the strengths and weaknesses of each. Then we discuss some general issues that are common to all of the applications.

BACKGROUND

A student's interaction with text in the context of learning involves a complex sequence of cognitive processes, some of which are shared across tasks and some are task-specific. On a basic level, the student reading to acquire knowledge for writing a research paper must simply understand the material from each text. Text comprehension is itself a complex task that requires comprehension of individual statements, recognizing connections between statements, relating statements to prior knowledge, and integrating these elements into a coherent representation of the text (e.g., Graesser, Singer & Trabasso, 1994; Kintsch, 1988; 1998). The Reading Strategy Assessment Tool (RSAT) was designed to assess high-school and college students' use of successful reading strategies during comprehension. Such reading strategies are important because they help readers construct a coherent mental representation of a text and have been shown to be predictive of comprehension (Magliano & Millis 2003, Magliano, Trabasso, & Graesser, 1999).

On a more global level, our hypothetical student must be able to reason more deeply with and about the texts they are reading (Rouet, Britt, Mason, & Perfetti, 1996). Our other projects focus on the student's use of their text representation in various reasoning tasks. The Critical Thinking tutor (CT Tutor) helps students critically evaluate scientific studies, Operation ARIES! (*A*cquiring Research Investigative and Evaluative *S*kills) teaches the scientific concepts needed to evaluate studies, Cultivating Argument Skills Efficiently (CASE) teaches students to comprehend, evaluate and produce arguments, and Sourcer's Apprentice

Intelligent Feedback (SAIF) helps students write essays from multiple texts.

The specific task focus varies across these projects. As such, the assessment component uses techniques geared at quite different processes, and will therefore be presented within the separate sections to follow. In this introduction, we will outline the general principles derived from models of discourse comprehension and learning which we follow and incorporate into all the projects. Broadly speaking, all projects aim to achieve effective and efficient *assessment* of the student's success in the given task. The projects geared at intervention further aim at effective *feedback*, and as a result, improved performance.

Assessment

Development of the assessment components for all these projects can be described within the framework of the evidence-based approach (Mislevy, 1993; Pellegrino & Chudowsky, 2003; Pellegrino, et al., 2001). This approach incorporates a student model, which identifies cognitive states and processes of the student that predict task outcome; a task model, which specifies the task requirements and how they are served by the cognitive processes that are part of the student model; and guidelines for how to interpret a student's task performance with respect to the underlying cognitive processes. With this framework in mind, each project identifies the steps in the complex comprehension process that are of key interest to the specific task. Table 1 provides the key processes that are identified (see Elements Identified) for each project that support this evidence-based approach.

Methods

For each application, students type input in response to a prompt such as *What are your thoughts?* (RSAT), *Are there any flaws with this design?* (CT tutor), *Why is it important to control for confounding variables in research?* (Operation ARIES!), *What was the predicate of the main claim?* (CASE), or *To what extent was Carnegie responsible for breaking the union at Homestead?* (SAIF). As shown in Table 1 (See Method of Identification), the methods for assessment range from word (Literal word matching and soundex) or string matching procedures, in which student input is compared to ideal answers to more complex semantic evaluation using Latent Semantic Analysis (LSA) (Landauer & Dumais, 1997). Methods are chosen to serve various objectives (see Use of Information Identified in Table 1). The overarching goal is to tailor assessment methods to effectively and efficiently measure the success students have at very specific tasks with minimal computational effort. By focusing on selective cognitive processes in each project, it is possible to provide effective feedback with relatively little computational expense. This goal, however, is challenging for at least two reasons. The first challenge is to balance computational prowess with pedagogical objectives. At times, computational techniques can place limits on the assessment and feedback processes in ways that may hinder optimal learning. The second challenge is to create applications that can be adaptable to new domains or material sets. For example, SAIF uses LSA to evaluate coverage of content as well as the extent to which essays are independent from the wording of the sources; however, it appears that this success is in part dependent on having LSA trained on a topic-specific text corpus. This makes it impractical for students to select their own topic for practice. We will return to such challenges in the discussion, for now it is enough to consider the trade-offs of computational effort, pedagogical objectives, and program adaptability. In the next section, we briefly describe each application and discuss its strengths and limits.

Table 1. Overview of projects from evidence-based approach perspective and the methods for identifying key processes

	Element Identified	Method of Identification	Use of Information Identified
RSAT	Active processing strategies to support comprehension (bridging inferences, elaborations, paraphrase). Unit targeted includes the current sentence, the immediately prior sentence, and distant textual information.	Student answers to questions are compared to semantic benchmarks (e.g., current sentence, prior text, ideal answer) that either reflect processes or comprehension. Early versions used LSA, but RSAT currently uses literal word matching and soundex algorithms.	Assesses processes while reading that support comprehension and comprehension level inferences. Currently used to predict comprehension skill.
CT Tutor	Target critical flaws in reports of experiments (e.g., lack of control group, experimenter bias, invalid measure).	Uses LSA and word overlap, in a dialogue with 2 agents, to detect if target flaws are correctly identified following a general prompt.	Assesses identified flaws and, when incorrect, provide hints and prompts to help students learn to correctly identify those flaws.
ARIES	Research methods concepts. Focus on definition, why it is important, and an example.	Overlap with key words and synonyms in target expectations. Modified Autotutor with multiple agents.	Assesses acquisition of key concepts and determines whether material is learned. When incorrect, determines whether student should watch another student being taught, prompt to explain material, or teach another student.
SAIF	Sourcing (citations, lack of plagiarism, appropriate use of quoting) and the degree of integration (coverage of material and extent of plagiarism) in essays written from reading multiple documents.	Sources were detected by a combination of LSA and string matching. Content was detected with LSA.	Assesses unsourced quotes, plagiarism, insufficient number of explicit citations, insufficient number of distinct sources mentioned, and excessive quoting. Provides feedback and help with revision by including a dynamically modeled explicit source citation.
CASE	Precision in representation of the predicates and themes in main argument claims.	String matching.	Assesses accuracy of claim predicates in recall while comprehending or evaluating arguments and provides feedback to encourage precise representations.

SPECIFIC APPLICATIONS

The Reading Strategy Assessment Tool (RSAT)

Background. Comprehension emerges as a result of inference and strategic processes that support the construction of a coherent mental model for a text (e.g., Graesser, Singer & Trabasso, 1994). However, the vast majority of tests of comprehension skills adopt a format that does not afford an assessment of these processes as they operate during reading (Magliano, Millis, Ozurur, & McNamara, 2007). Specifically, readers comprehend texts and then answer multiple-choice questions regarding different aspect of their understanding. Although it is certainly possible to develop multiple-choice comprehension tests that address important theoretical constructs associated with comprehension (Mislevy, 1993; Pellegrino & Chudowsky, 2003; OECD 2002; Pellegrino, et al., 2001), we have argued that the format of asking questions after the text is read and while it is still available compromises the ability of these tools to directly target the processes that support comprehension (e.g., Magliano & Millis, 2003).

We have been exploring the viability of developing computer-based systems that analyze verbal protocols that are produced *while* a student is actually reading a text. This work is motivated by a substantial amount of research that has demonstrated that thinking aloud protocols produced while reading and question answering (Magliano & Millis 2003, Magliano, Trabasso, & Graesser,

1999; Millis, Magliano, & Todaro, 2006; Olson, Duffy, & Mack, 1984; Trabasso & Magliano, 1996) are predictive of ones' comprehension level. Moreover, they are indicative of inferences and strategies that support comprehension (e.g., Magliano, 1999).

How RSAT works. R-SAT is a computer-administered test that is designed to assess a student's level of comprehension and the processes that support it while reading (Gilliam, Magliano, Millis, Levinstein, & Boonthum, 2007; Magliano, Millis, The RSAT Development Team, Levinstein, & Boonthum, under review). The elements identified and methods for identification are summarized in Table 1. The initial version of RSAT simply asked students to type their thoughts after reading pre-selected target sentences. The more recent version prompts users with one of two types of open-ended questions: indirect and direct. *Direct questions* are "wh-" (e.g., why, what) questions about the text and provide an assessment of a reader's comprehension. *Indirect questions* require readers to report thoughts regarding their understanding of the sentence in the context of the passage. Specifically, participants are instructed to provide answers that are akin to thinking aloud (e.g., Trabassso & Magliano, 1996). These answers provide assessments of the processes that support comprehension. RSAT targets three types of processes: making bridging inferences between the current sentence and the prior discourse context, elaborating based on world knowledge, and paraphrasing the current sentence.

The crux of RSAT is to identify what the person is thinking by comparing the typed input to different types of information representing different responses. For assessing performance on the direct questions, student responses are compared to ideal answers to the questions representing different types of strategies supporting target inferences. Although we have used LSA to match student responses to content words representing different strategies and examples of strategies (Millis, Kim, Todaro, Magliano, Wiemer-Hastings, & McNamara, 2004), the current version of RSAT now relies solely on content word overlap. The direct and indirect answers are analyzed via word count algorithms (literal matching and soundex) to detect and count only content words (nouns, pronouns, verbs, adverbs, and adjectives). The direct answers are compared to ideal answers to the "wh-" questions. The indirect answers are compared to two semantic benchmarks. The first benchmark is the information in the current sentence, which provides an assessment of paraphrasing. The second is the information in the prior text sentences, which provides a measure of bridging. All content words in the protocol that did not get counted as occurring in the current sentence or prior text sentences are counted as elaboration words. RSAT computes comprehension, paraphrasing, bridging, and elaboration scores by computing mean word counts (i.e., averaging across items) for each score.

Strengths and limitations of RSAT. Although research is continuing regarding the validity and reliability of RSAT, a substantial amount of data indicate the viability of RSAT (Gilliam et al., 2007; Magliano et al., under review). First the RSAT approach was shown to predict measures of comprehension comparable to standardized tests, which demonstrated convergent validity between comprehension scores and other well-established measures of comprehension (Magliano et al., under review). Moreover, the processing measures (i.e., paraphrase, bridging, and elaboration scores) are correlated with these measures of comprehension, which validates the underlying cognitive model that provided the basis for RSAT (Gilliam et al, 2007; Magliano et al., under review). Finally, we have demonstrated respectable construct validity in that the RSAT scores are highly correlated with human judgments of the verbal protocols with Pearson correlations ranging from .75 to .48 (Magliano, et al., under review). The lowest correlation typically occurs for the measure of elaboration, which has proven to be the most difficult inference to detect with computer-based

assessment (Millis, Magliano, Todaro, & McNamara, 2007).

That said, RSAT has several limitations that require further refinement. Detecting elaborations through automatic coding has proven to be a challenge (Millis et al., 2007), which, in part, stems from the fact that there is a greater variety of responses that can be produced in the context of elaborations (based on general knowledge) than bridging inferences (based on the text). One initial strategy to improve this coding could be to distinguish between different types of elaborations (e.g., relevant, irrelevant, inferences, recollections). Another limitation is that RSAT does not provide assessments of the quality of the paraphrases, bridges, and elaborations reported by students. For example, a good paraphrase should not simply repeat the sentence that was just read, but rather summarize it. A good bridging inference should establish how the sentence is related to the discourse context in a manner that reflects the important implicit and explicit relationships between sentences. Finally, another limitation -- in the context of RSAT's use in a future intervention -- is that there are no mechanisms for providing feedback to the users regarding the quality of their protocols. Future research will attempt to address these limitations.

The Critical Thinking Tutor (CT Tutor)

Background

Development of critical thinking skills is a major objective of college education, and at times, entire courses are devoted to teaching these important skills. Critical thinking involves the mindful application of reasoning, problem solving, metacognitive knowledge, and decision-making skills in an effort to find a desirable outcome to a problem (Halpern, 2003). The CT Tutor was constructed to enhance students' ability to reason within three domains: everyday arguments, persuasive communications, and scientific research. At the heart of each of these domains, the CT Tutor teaches the student how to identify flaws in reasoning. In the domain of persuasive communication, the tutor teaches students to identify logical fallacies (e.g., arguments against the person, appeal to pity). For everyday arguments, the tutor teaches students the components of arguments (claim, reasons, and warrants). In scientific research, the CT tutor guides students in locating flaws in research summaries, such as the lack of a control group where one would be expected.

CT tutor provides training in all three areas through interactive evaluation of short verbal descriptions (of studies and arguments) that are flawed to varying degrees. Providing training for recognizing flaws is important because most people do not read in a "deep" fashion. Most people do not spontaneously question information as long as it sounds convincing (NAEP, 2006). In terms of the cognitive processes involved, the task demand here goes markedly beyond mere comprehension. The student has to actively question information [i.e., they have to use a model of correct scientific methodology (as well as consider possible flaws) to evaluate the components of a study, such as confounding variables or limitations on the validity of a measure]. In the case of arguments or persuasive fallacies, the student needs to be able to recognize reasoning flaws that they have previously learned or to recognize and evaluate support presented for an argument. In all these cases, the student thus actively applies knowledge to a novel problem presented in the text (Bransford, Brown, & Cocking, 2000), also known as problem-based learning (Hmelo-Silver, 2004; Kolodner et al., 2004).

How CT Tutor Works

The CT tutor was built on the platform of AutoTutor (Graesser, Chipman, Haynes, & Olney, 2005; Graesser, Wiemer-Hastings, Wiemer-Hastings, Kreuz, & the TRG, 1999; Graesser et al., 2004). Problems are presented by two pedagogical agents:

a "teacher" and a "student." In the domain of scientific reasoning, the student agent summarizes a flawed study. One example involves a study showing the positive effects of a new diet pill, but the study lacks a control group. The teacher agent states that there might be a problem with the study and asks the human to type in any flaws of the study. The teacher guides the student towards the correct answer by providing hints (e.g., "Think about making comparisons.") and prompts (e.g., "In the study, there was no control what?"). When all of the parts of an ideal answer are covered, the student agent provides a summary what he had learned as a result of the tutorial exchange.

A key aspect of this tutor is that it incorporates active learning (Aleven & Koedinger, 2002; Chi, de Leeuw, Chiu, & LaVancher, 1994; Graesser et al., 2004). Instead of providing the student with definitions and examples of a flaw, the student is immersed in authentic samples of studies or arguments with the goal of detecting various flaws in specific applications. This pedagogical objective requires online assessment of the student's evaluation of a problem, and the incremental learning process that leads to successful identification and explanation of a flaw. Assessment occurs by matching the student's input against ideal answers for each scenario, which range from about 3 to 7 answers. For each expected answer, the tutor has hints and prompts that can help a student access or acquire the information they need to complete the answer.

Strengths and Limitations of CT Tutor

A strength of the CT Tutor is that we have shown that it contributes to learning. Student learning was measured offline by presenting pre- and posttests to evaluate general concept knowledge, as well as the ability to apply the methodological concepts to novel problems (Storey, Kopp, Wiemer, Chipman, & Graesser, in press). Compared to a condition in which students only read a relevant text on the methodological concepts, and a control condition that was engaged in a topic-irrelevant text, students who interacted with the tutor showed significant improvement on the concept knowledge and application to novel problems over the two control conditions.

Through our research, we noted two main limitations to the tutor. First, it does not explicitly provide much of the requisite knowledge for successful interaction with the tutor. This occurred by design because the tutor was meant to provide practice opportunities for someone taking a course (or reading a book) on critical thinking. We assumed that the course (or book) would provide the necessary knowledge of research design and argumentation. This raised the issue that students may not have uniform knowledge from which to successfully interact with the tutor. It also indicated that it could not be a stand-alone module or learning environment.

Second, many students appeared to get bored interacting with the tutor because it took almost two hours to go through all of the problems. Apparently, our students did not share our appreciation for spending hours of practice in order to master complex skills. The problem of the time-consuming interactions is a general problem for many ITSs that teach through active learning. One approach to this problem is to intersperse less interactive practice opportunities within the more time-consuming full-dialogue trials. We have recently shown such a method to be both effective and more efficient (Kopp, Britt, Millis, & Graesser, in preparation). In this study, students receive scientific reasoning problems presented with varying levels of dialogue: all in full dialogue (CT Tutor), mixed dialogue (half in full dialogue and half with only a single opportunity to answer with feedback but no hints or prompts), or no dialogue control. We found that the mixed dialogue method lead to the most learning per unit time. Thus, combining a number of limited dialogue trials with deeper, more active trials may be one solution to students becoming bored during training. Another option is to provide a more engaging game-like environ-

ment to help maintain the student interest. This led us to develop Operation ARIES!.

Operation ARIES!

Background

Operation ARIES! is currently being developed among researchers at Northern Illinois University, University of Memphis and Claremont McKenna College. ARIES is an acronym for *a*cquiring *r*esearch *i*nvestigative and *e*valuative *s*kills. The goal of Operation ARIES! is to teach students how to critically evaluate research that they may encounter on the Web and in various media outlets. Operation ARIES! has three main modules that users progress through sequentially. The first module is Interactive Text in which users read an online text that covers all of the concepts to be taught (e.g., need for control groups, independent and dependent variables, etc). The second module is Case Studies, in which users evaluate flawed research. This is not too different from the scientific reasoning component of the CT tutor. The third module, Interrogation, teaches the user how to uncover and evaluate implicit information from a research study by asking questions. In many cases, to evaluate a study, a person must actively seek out answers to questions (e.g., "Were the answers scored objectively?" or "Is there a conflict of interest?"). In this module, the user is given the opportunity to ask scientists questions about their research in an effort to determine whether the research contains flaws.

As one can see, Operation ARIES! addresses the problem of requisite knowledge by providing an online textbook to students before they start applying their knowledge on authentic problems. What about the problem of boredom or lack of engagement? The answer, of course, lies in video games. It is well known that video games can be very appealing to students (Yee, 2006), and many games contain attributes that are pedagogically appealing (Gee, 2007). Therefore, Operation ARIES! uses many features commonly found in video games, including narratives, multiple agents, goals, scaffolding, just-in-time information, and points. The storyline is that the player is asked to join the Federal Bureau of Science (FBS) to help locate extraterrestrials from the Aries constellation that are secretly on Earth. It appears that the aliens, who are disguised as humans, are publishing bad research in an effort to undermine our knowledge of the scientific method, along with other nefarious plans. The player then must learn good science so that he or she can identify the faulty research being published by the aliens. The player is guided by Dr. Quinn, who is an FBS handler, and by Glass Tealman, who is a fellow student. The player learns about the aliens' plans throughout the game, and helps Glass to pass the course. Needless to say, by the end of the game, the player and Glass defeat the aliens and save the Earth from destruction.

How ARIES Works

The scope of Operation ARIES! is rather large, so we will only focus on the Interactive Text module here. As mentioned above, the player first reads an on-line book about the scientific method. One unique aspect of the online text is that it is a manual written by the aliens given to their spies (Diane Halpern is the human that actually spearheaded the writing of the manual). Sprinkled throughout the text are references to their home world, Thoth, and to Human Beings, which they refer to as Human Beans. The hope is that this background story will help maintain interest in the material. To further maintain engagement in the material, Dr. Quinn and Glass begin each chapter with a short dialog, which introduces the chapter material but in the context of the overall storyline. The player receives emails from Glass and helps Glass make decisions in various situations (of course, Glass falls in love with one of the alien spies). Furthermore, the player answers

open-ended questions throughout the book and takes multiple-choice questions on the material.

One unique aspect of the Interactive Text module is that the player engages in *trialogues*. A trialogue is a conversation between the human player, Dr. Quinn, and Glass. Trialogues use a keyword-matching algorithm to determine the completeness of an answer. They occur immediately after the human player answers each of the last three multiple-choice questions associated with a chapter. There are three types of trialogues. In the Teaching type, the human player is asked to teach Glass, who got the question wrong. These are triggered when there is evidence (based on performance on the multiple choice items) that the human player has high knowledge of the targeted concept. In the Tutor type, Dr. Quinn teaches the human player, in much the same way AutoTutor coaches students in answering a problem. These occur when there is evidence that the human has intermediate knowledge of the targeted concept. In the Vicarious Learning type, the human player watches as Dr. Quinn teaches Glass. This trialogue type is reserved for times when the human shows evidence of low knowledge because there is evidence that low knowledge students are helped by watching others interact (Craig, Sullins, Witherspoon, & Gholson, 2006).

Strengths and Limitations of ARIES

We are in the midst of building and testing Operation ARIES!, so we do not know the full extent that students will learn from it and whether it will be as appealing as we hope. However, we have some preliminary evidence that the trialogues will be effective. We had students answer six multiple choice questions associated with five chapters. After they chose their answer for each question, all were given corrective feedback. The first three questions were meant to gauge the prior knowledge of the topics. For participants in a *trialogue* condition, the last three multiple choice questions were each followed by a trialogue based on their performance on the questions. For participants in the *no trialogue* condition, there was no trialogue after each of the last three multiple choice questions, just corrective feedback. After they completed all questions associated with the five chapters, the students were given an open-ended test on all of the concepts. The test asked for a definition of the concept (e.g., "What is an independent variable?), the importance of the concept (e.g., "Why are independent variables important in science?), and an example (e.g., "Write down a novel example of an independent variable?"). It should be noted that the multiple choice questions covered these aspects of each concept. Adjusting for prior knowledge, the participants in the trialogue condition scored an average of .42 on the open-ended posttest, whereas participants in the nontrialogue condition scored an average of .37, $F(1, 87) = 5.95$, $p < .01$; $d = .41$.

Sourcer's Apprentice Intelligent Feedback (SAIF)

Background

One of the most challenging tasks that students encounter in school is reading multiple documents and writing an essay such as a research paper. Students have to attend to and evaluate the source of the content prior to reading and using the content (i.e., sourcing), integrate the information across documents, evaluate information for consistency across documents (i.e., corroboration), and then use this information in their essay. High-school and college students, however, do not spontaneously engage in many of these sourcing and integration skills (Brem, Russell, & Weems, 2001; Britt & Aglinskas, 2002; Rouet, Britt, Mason, & Perfetti, 1996; Rouet, Favart, Britt, & Perfetti, 1997; Wiley, Goldman, Graesser, Sanchez, Ash, & Hemmerich, 2009; Wineburg, 1991; Wolfe & Goldman, 2005). For example, high school students failed to encode or evaluate source information prior to reading the content of a document

Table 2. SAIF essay detection methods & rules

Problem	SAIF ID methods	SAIF Rule	SAIF Feedback
Plagiarism	*Plagiarism*: LSA cosine > 0.75; *Source citation*: String match or LSA cosine > 0.80; (NO *Quotation marks*)	If plagiarism with no citation	Lists suspect sentence(s) and prompts to reword; presents a transformed sentence modeling the proper format.
Unsourced quotes	*Plagiarism*: LSA cosine > 0.75; *Quotation marks*: Pattern match; (NO *Source citation*)	If quote marks with no citation	Prompt for an explicitly credit to source and model proper format.
Lack of explicit citations	*Source citation*: String match or LSA cosine > 0.80	If citations > 3	Prompt to make a minimum of 3 explicit citations.
Under use of distinct sources	*Source citation*: String match or LSA cosine > 0.80	If different sources > 2	Prompt to cite at least 2 different sources.
Excessive quotation	*Quotation marks*: Pattern match	If quoted sentences > 50% of total essay	Prompt to paraphrase more instead of relying on quotations too heavily.
Insufficient amount of integration	*Content covered*: LSA > 0.60 with sentence for doc	If covered doc > 3	Prompt to include a more complete coverage of the documents in set.

(Wineburg, 1991), viewed the textbook as more trustworthy than primary documents (Wineburg, 1991), and used information from novels and films to support their claims (Britt & Aglinskas, 2002; Seixas, 1994; Wineburg, 2000). We are not surprised with this lack of multiple-document reading skills since students often do not receive explicit instruction on evaluating sources (Wiley et al., 2009). Recently, researchers have designed several interventions, such as Sourcer's Apprentice (Britt & Aglinskas, 2002), Met.a.ware (Stadtler & Bromme, 2007) and Seek (Wiley et al., 2009), to teach these skills to students.

The Sourcer's Apprentice, SA, is a computer environment designed to help students develop these multiple document skills (Britt, Perfetti, Van Dyke, & Gabrys, 2000). In SA, students are given instruction on how to identify important source features (e.g., who the author is, when it was written, etc) and then are given a set of documents that relate to a controversial topic in history. We found that students given training and practice with SA included more explicit citations and integrated material from more distant sources than students not given such training (Britt & Aglinskas, 2002). SA is limited, however, in that it does not provide feedback on the quality of their essays. To provide such support, we developed Sourcer's Apprentice Intelligent Feedback (SAIF) to accompany SA (Britt, Wiemer-Hastings, Larson, & Perfetti, 2004).

How SAIF Works

SAIF assesses the quality of the students' use of sources, in terms of explicit citations and appropriate use of quoting, and the degree of content integration, in terms of coverage of material and the lack of plagiarism. Students write an essay and then submit it for analysis. SAIF automatically detects common problems in the essay (see Table 2) and then provides immediate feedback. If there is no problem with that aspect of the essay, they are given positive feedback. If there is an apparent problem, they are given corrective feedback with a modeled correct usage of source information. For example, a sentence that SAIF determines is unsourced copied material may receive the following feedback:

This sentence might be plagiarism (unsourced copied material)

Your sentence: The strike dragged on until November, but by then the union was dead and thousands of workers had lost their jobs.

Possibly plagiarized from: (p=1.0: "King", The strike dragged on until November, but by then the union was dead and thousands of workers had lost their jobs)

Example appropriate citation: According to King, "The strike dragged on until November, but by then the union was dead and thousands of workers had lost their jobs"

To create SAIF, we needed to be able to detect 4 types of information: *Plagiarism*, *Source citation*, *Quotation marks*, and *Content covered*. Plagiarized sentences were identified as those that exceeded a LSA cosine of 0.75. Source citations were identified by (1) a string match to the name of a non-character author, (2) pattern matching to detect parentheses and then comparing the contents to source information (e.g., author name or book title) from the documents using string matching or a high LSA cosine, or (3) using string matching to identify citation starters (e.g., "according to", "in his book", "claims"). Quotation marks were identified through simple pattern matching. Finally content covered was determined by comparing each essay sentence to every sentence from a document. A document was considered covered if an essay sentence achieves an LSA cosine was greater than 0.60 with any sentence from that document. It is beyond the scope of this chapter to explain why LSA parameters or essay quality parameters were selected; see Britt et al, (2004) for these more information and justification of these parameters.

As shown in Table 2, plagiarism and unsourced quotes were detected by comparing each essay sentence to each sentence from the documents. Plagiarized sentences were defined as those that exceeded a LSA cosine of 0.75 without a source citation and without quotes (see SAIF rule in Table 2). Unsourced quotes were identified by quoted verbatim material that does not include a source citation. Feedback included a listing of all suspected plagiarized or unsourced quotes along with the sentence it was too close to. The student was instructed to reword these sentences or to quote with an explicit citation. One of the student's problematic sentences was dynamically transformed to demonstrate appropriate sourcing conventions and to serve as a model for the other problematic sentences (as shown in the "King" example above). SAIF also identified the absolute number of explicit source citation and distinct sources using the three methods above. If an insufficient number of citations was found, SAIF instructed the student to either try to include a minimal number of 3 or to make their implicit or vague citations more explicit. If they did not explicitly refer to at least two different documents, they were prompted to do so. The problem of excessive quoting was identified by the number of sentences that included quotes, ignoring single word quoting. If at least 50% of text in the essay was quoted, they were told that they should work to put things in their own words. Finally, insufficient content integration was determined by examining the amount of content covered. If SAIF determined that the essay mentioned information from two or fewer documents (of 7), SAIF suggested that the student should not rely on only a couple of documents.

Strengths and Limitations of SAIF

In comparing SAIF's detection of problems to that detected by an expert human rater (Britt et al, 2004), we were able to show high agreement (Cronbach's alpha) in identifying plagiarism (81%), unsourced quotations (80%), identification of explicit citations (76%), and identification of which document mentioned the information (91%). Thus, SAIF can be used to automatically classify source elements from student essays written on a historical controversy. We also found that SAIF was effective in helping students write better essays. In this experiment, students received SA's

Table 3. Agreement between SAIF-G and human raters in identifying sourcing problems in science essays

	Size of space		Generality of space	
Problem type	**SAIF small space - Human rater**	**SAIF large space - Human rater**	**Topic-specific space - Human rater**	**General space -Human rater**
Plagiarism	87%	89%	87%	28%
Unsourced Quotes	98%	98%	98%	94%
Explicit citations	98%	85%	98%	85%

tutorial and wrote an essay. Then students received SAIF feedback, a reminder about proper sourcing, or were just told to revise the essay. There were significantly more explicit references to sources in the essays given SAIF feedback.

While SAIF appears to be effective in increasing citations for history topics, we wanted to expand SAIF to cross-disciplinary instruction. Modifying the tutorial instructions and practice texts was not difficult. However, it was unclear whether the same parameters we used for history essays can be used to detect problems in essays written in a science domain. To test the parameters, we had 63 students read the cross-disciplinary tutorial or a control tutorial. Then students read several articles on global warming (Bråten, Strømsø, & Samuelstuen, 2008) and wrote an argument essay. We found that plagiarism (M=1.59 per essay and 63% of essays) and lack of explicit citations (M=0.69 per essay and 78% of essays had this problem) are common problems with students' science essays just as they were in history essays. SAIF was able to accurately identify plagiarism (87% agreement), unsourced quotes (98% agreement), and explicit citations (98% agreement) using the same parameters as in the history essays.

We used this data to test two other issues that could potential effect the usefulness of this tool. First, we used a LSA space of approximately 30,000 words for both the history and science databases. We questioned whether a significantly larger LSA space would improve detection of each plagiarism, unsourced quotes, and explicit citations. To create a larger space we added approximately 276,000 words to the 30,000 from the smaller space, resulting in over 306,000 words. As shown in Table 3, the relatively smaller space was as good or better than the larger space.

We also wanted to examine whether a topic-specific space is required. The LSA spaces used to test SAIF were created from texts on the topic (e.g., 1892 steel strike Homestead in Pennsylvania and global warming). It would be ideal, in terms of adaptability, if a single general space could work well for all topics. Otherwise, teachers would have to create their own space or we would have to create and include an automatically generated specific space. We compared our topic-specific global warming space to the Colorado space. As shown in Table 3, the topic-specific space was superior to the general space. From this data, it seems that it may be necessary to create a space for the particular topics on which the students are writing the essay.

The good news for a general application is that we found that general parameters do apply to two very different disciplines and that a relatively small LSA space is "good enough". This good news is tempered by the findings that a topic-specific space may be required. This limits the utility of the current version of SAIF since one would have to create a new space for each new topic assigned. Currently Peter Hastings is working to create an automated space creator.

The current version of SAIF is limited also in that it does not verify the accuracy of the sourced information. This may not be too difficult to add

in the next version for quoted material and explicit citations. It will be much more difficult for implicit citations and citations in which the source and the content span multiple sentences. Currently we are using the sentence as a unit and this would have to be modified. However, automatic testing of the limits or scope of a source across sentences or paragraphs may prove too difficult. SAIF also does not assess the quality of the information that is integrated. This could be addressed in much the same way as Summary Street (Kintsch, Caccamise, Franzke, Johnson, & Dooley, 2007) by hand coding key evidence or facts that should be integrated in the argument essay and using these as benchmarks or materials that must be covered.

Cultivating Argument Skills Efficiently (CASE)

Background

Understanding, evaluating and writing arguments are key skills for our hypothetical student asked to write a research paper. Arguments (such as 1a-1d) have minimally a claim and one supporting reason (Toulmin, 1958; Voss & Means, 1991). The claim is a controversial assertion that includes both a predicate (e.g., is unfair, is unnecessary, will be ineffective) and a theme (e.g., banning cell phone use while driving) (Britt, Kurby, Dandotkar, & Wolfe, 2008).

1a. Banning cell phone use while driving **is unfair** because everyone should not be penalized just because a few people can't do it responsibly.
1b. Banning cell phone use while driving **is unnecessary** because laws against unsafe and inattentive driving already exist.
1c. Banning cell phone use while driving **will be ineffective** because people will just break the law and it will be too hard to catch violators.

Creating a precise representation of the predicate of a claim and keeping that active while reading an argument is an important skill for several reasons. First, the predicate of the claim is what the reasoner is trying to persuade one to believe or do. So without memory for the claim, the representation is an inadequate representation of the argument. Second, the claim predicate will dictate the set of reasons that can be put forth as support. For example, switching the reasons in the above pairs leads to unwarranted arguments (i.e., reasons that do not provide support for a claim).

*2a. Banning cell phone use while driving **is unfair** because people will just break the law and it will be too hard to catch violators.
*2b. Banning cell phone use while driving **is unnecessary** because everyone should not be penalized just because a few people can't do it responsibly.
*2c. Banning cell phone use while driving **will be ineffective** because laws against unsafe and inattentive driving already exist.

Finally, this skill of keeping the claim predicate active may require effort because there may be significant intervening textual information between the statement of a claim and each of the supporting reasons. Thus, the reader may have to keep track of the claim predicate in order to evaluate whether that reason supports the claim.

In recent studies, we had undergraduates read simple claim-reason arguments and immediately recall the claim. We found relatively poor recall of the claim predicate ($M = 76\%$) but good recall of the theme ($M = 95\%$). We also found that those readers with the most precise representation of the claim predicate were more skilled at distinguishing well-structured arguments from poorly structured arguments. These findings suggest that students form a gist representation of the claim predicate (Brainerd & Reyna, 1998, 1992; Kintsch & van Dijk, 1978, 1988, 1998) even though a verbatim representation may be required to evaluate the

quality of the argument. Furthermore, it is not a problem of simple careless encoding of the predicate. In two probe studies (Kurby, Britt, & Dandotkar, 2006), we found that both skilled and less-skilled reasoners were equally accurate at recognizing the predicate and the theme immediately after the claim. The problem comes in maintaining the representation. After reading the reason, less-skilled reasoners were less accurate at recognizing the predicate than the theme but skilled reasoners are still equally accurate at recognizing the predicate and the theme. Thus, while less-skilled reasoners do encode the predicate, they are less likely to keep it active even though it is necessary for representing the complete argument and judging quality.

How CASE Works

To teach students to represent and evaluate the quality of arguments, we have developed several CASE (Cultivating Argument Skills Efficiently) web-based modules (Larson, Britt, & Kurby, 2009). We will present only two modules here. The *Predicate Identification module* provides interactive instruction and practice in attending to the claim predicate. It begins by teaching students to recognize the claim elements with special attention to the claim predicate. Students are asked to click on the predicate or theme of the claim of a set of short (2-clause to a paragraph) arguments. If they are incorrect, the program highlights in red the incorrectly selected element and highlights in green the correct answer. Then, to encourage the maintenance of the claim predicate during the reading and evaluation of the argument, students have to read an argument and click a button to remove the argument. The student evaluates the argument and then types the predicate or theme into a textbox.

Before creating this module, it was necessary to determine how accurate the scoring of recall needs to be to help students. In this work, we coded the recall of responses according to a verbatim match (exact match) or a more liberal textbase match (that allows synonyms). It would be much easier to automatically score a verbatim match using only string matches than to also allow synonyms as in a textbase match, which would require LSA. Fortunately, a comparison of the two criteria found that the verbatim criterion was more sensitive in distinguishing skilled reasoners than the textbase criterion. Because the verbatim scoring is extremely easy for us to do automatically and it is better in terms of discriminating skilled from less-skilled argument evaluators, it was used as the scoring system for the module. Therefore, scoring can be accomplished using simple string matching. If the response is incorrect, the argument is presented again with the correct segment in green font.

The second module, *Evaluate Quality module*, was designed to teach students to distinguish structurally acceptable arguments, such as 1a-d above, from structurally flawed arguments (i.e., unsupported and unwarranted), such as 2a-d. In the first practice set of this module, students evaluate the quality of the argument while the argument is still present on the screen. Because the predicate is critical to this evaluation process, we have a second practice set that requires students to hold the claim predicate in memory while making their quality judgment. For these practice items, the student reads each argument, clicks to remove it, makes their flawed judgment, and then types the claim predicate in a textbox.

Strengths and Limitations of CASE

These modules were shown to be effective in teaching students to distinguish structurally bad arguments from structurally good arguments. In two experiments testing the effectiveness of these modules, we found that the modules led to a 20% increase in quality judgment accuracy for college students and an 18% increase for high-school students compared to a no-treatment control. In contrast, without immediate feedback, students

did not learn to make this fine distinction (Larson, et al, 2009). Therefore, it is critical that the tutor can accurately provide immediate individual feedback. We were even more encouraged by a follow-up study, which showed that this improvement was resilient over at least a short period of time – 1 week (Britt, Storey, Kopp, Dandotkar, & Larson, 2007). In this study of college students, we replicated our earlier findings in that participants given the modules were more accurate in judging arguments (M = .81) than the control group (M = .63) on an immediate test. We also found that the module group (M = .83) was still more accurate on than the control group (M = .63). Thus, there was no loss of skill after a one-week delay in testing.

Although we have shown learning gains as a result of these modules, we still have several challenges to consider. First, not all students reached a minimal level of mastery and we failed to find evidence of transfer to the more complex skill of comprehending others' arguments. Given that at least some students will require multiple exposures, one challenge is how to motive students to complete additional exposures. Second, we would like to use NLP techniques to identify claims and reasons in student-produced essays. Simple methods such as LSA could be used to detect whether reasons are semantically or thematically related but would probably not be useful in detecting unwarranted or poorly structured arguments. Automatic detection of claims and reasons is a very difficult task. Claims that differ by a single term cannot be supported by the same set of reasons (such as 2a-2b above). Furthermore, negation is very important to the meaning of claims but negation poses a significant problem for LSA. Thus, such simple automatic methods of assessing and providing feedback may prove untenable.

ISSUES, CONTROVERSIES, PROBLEMS

Across a variety of text-based reasoning tasks, we have shown that simple methods are usually sufficient to assess information from student responses and in some cases, guide directive feedback. There are several general issues that arise from such endeavors. In this next section, we address some issues that emerge in using NLP in helping improve students' reasoning with texts.

Feedback

One primary advantage of intelligent tutoring systems is that they provide an opportunity to give individualized training and immediate feedback that may not otherwise be practical. This feedback can come from an automated agent (e.g., ARIES and the CT tutor) or be provided as part of the tutor environment (e.g., CASE and SAIF). The appropriateness of the feedback is critically dependent on the accuracy and precision of the system's assessment of the student's response or text. The simple text-processing techniques used in our projects generally provide good assessments, but they are not perfect. This classification inaccuracy may lead students to become frustrated or to disengage from the tutor. Thus, a critical issue is what standard of accuracy will students expect and accept? While getting students to engage in the target processing activity may be all that matters, if students come to decide that the tutor is not accurate enough, they may lose motivation or try to game the system (Baker, Corbett, Koedinger, & Roll, 2006). Gaming the system refers to the situation in which students strategically exploit properties of the tutor to get the answer or advance rather than using the tutor as it was designed to be used (Baker et al., 2006). In fact, frustration was one of the primary reasons students reported they gamed an intelligent tutor (Baker, Walonoski, Heffernan, Roll, Corbett, & Koedinger, 2008).

To avoid these non-productive behaviors, the system must either significantly increase the accuracy of its judgments --- which would be computationally prohibitive --- or temper its feedback to the student. Fortunately for us, people generally tend to attribute more "intelligence" to computers than they should (Weizenbaum, 1966). Agent-based approaches have mitigated this tendency by including agents that also need to learn more about the content. Such approaches include peer agents/vicarious learning (e.g. ARIES; Craig et al, 2006) and teachable agents (Leelawong, & Biswas, 2008; Reichherzer, Cañas, Ford, & Hayes, 1998). Non-agent systems must textually temper their feedback by indicating that there is a possibility that what they have identified as problematic is actually acceptable. Similarly, these systems should avoid giving direct negative feedback, as in fact, human peer tutors do (Graesser & Person, 1994). Avoiding negative feedback may be especially important if students spend time with the system in proportion to need. Less-skilled or less-knowledgeable students may be the most sensitive to negative feedback and most affected in terms of their subsequent feelings of self-efficacy or interest and motivation in the domain.

Pedagogical Objectives Guided by Available Computational Tools

Using imperfect automatic methods to guide feedback may also lead students to "learn to the tool". By this we mean, students may think that the particular skills or knowledge that the system provides feedback on reflects their importance in the discipline, not what is computational feasible. With extended use of an application, the student may become successively attuned to the methods of assessment rather than toward learning the target skill. This may influence what students think is important. For instance, if an ITS scores student responses using key words, then students may begin to guess the key words rather than composing a coherent causal explanation or well-structured claim. In many reasoning tasks, however, form and details are very important. For example, order of information within a sentence, negation, and correct causal relationships are all very important. LSA and content word methods will not be able to distinguish the quality of different utterances using the same key words. Students may stop attending to this level of detail or structure if they are not receiving feedback on those aspects of the response. Furthermore, it is unclear how they will interpret feedback from a teacher on information that is not assessed by the tutor. For instance, if SAIF evaluates the student essay and doesn't provide feedback on the quality of the information from a source, the student may be reluctant to accept criticism from a teacher that the quality of the information should be improved. We believe this can be adequately dealt with by telling students that the tutor only gives feedback on a circumscribed set of essay features but students' understanding of such circumscribed feedback has not been tested.

Highly Constrained Tasks

The systems presented in this chapter are based on the cognitive model approach (Ritter, Anderson, Koedinger, & Corbett, 2007). The success of simple methods is in part due to conducting a detailed task analysis of successful task performance and then creating highly constrained targeted activities to help students engage in the type of processing that leads to successful performance. One must develop benchmarks that reflect successful performance (or varying levels of success), but the ability to accurately detect that performance may vary depending on how many constraints are placed on the activity. RSAT provides an excellent example of this principle given that it employs two types of questions: indirect and direct. By their nature, the indirect questions (e.g., What are you thinking now?) engender a wide variety of responses, which can complicate scoring. For example, it is extremely challenging to identify a

successful elaborative inference based on world knowledge (Millis et al., 2007). In contrast, direct questions have a clearly defined correct answer, and determining how well students overlap with this answer is less complicated.

Systems that require students to converse with an "intelligent" agent also add a level of complexity that requires a detailed task analysis. This analysis involves identifying possible student products (answers of varying levels of completeness and misconceptions) and specifying how the agents will respond after the assessment systems match student responses to these anticipated products. It is rare that a user will produce the complete and correct answer on the first response; therefore, an approach has to be scripted that guides the student to the correct response. Given the open-ended nature of these conversations, a central challenge is to identify when to "move on", either because the student has produced an adequate response or is not likely to do so. Situations in which the student is producing the right response on a conceptual, but not linguistic, level or simply cannot produce the right answer can be frustrating. Our approach in systems such as the CT tutor and ARIES is to have some protracted, guided exchange followed by a summary-recap of the correct response. This gives users the opportunity to compare their responses to ideal ones. Of course, the value of this comparison depends upon the extent to which the user engages with the system.

This highly structured, skill-on-demand approach appears to work well for many of the skills that we have targeted in our tutoring and training systems. In these systems, the instructor determines the learning or reasoning goals and there are few paths to success (an exception is the indirect questions in RSAT). However, such simple methods will not likely work in more naturalistic situations where there are generally multiple paths to reasoning from texts. It will also be less useful when reading goals are driven by the student rather than by the instructor.

LSA-Based vs. Keyword Based Approaches

One theme that has emerged from the development of these applications is the question of when LSA-based approaches significantly improve accuracy. Both RSAT and CT tutor started with a mixed, LSA and keyword-based approach but the more current version (RSAT and ARIES) rely only on keywords. The lack of a significant improvement in assessment accuracy for these applications may be a result of the length of the student response. Short text units are difficult for LSA (Foltz, Kintsch, & Landauer, 1998). Additionally, precision may matter. With the claim recall task in CASE, a verbatim match was required, so word matching was the preferred technique based on expert performance. In contrast, SAIF used a combination of LSA and string matching. We suspect the length of the response (M=284.5 words) and the desired fuzzy matching led to the success of the LSA approach.

CONCLUSION

We return to our hypothetical student trying to learn to how to write a research paper. We have shown the effectiveness of using simple approaches to automatically assess a subset of the component skills. Our applications show the generality of these methods for a variety of complex text-based reasoning skills and we expect that additional skills related to writing a research paper might be amenable to this treatment. As text processing techniques improve, tutoring systems will be able to make more accurate assessments and provide more refined feedback. For pedagogical goals, there is a delicate balance between assessment accuracy, directive feedback, motivation / interest in the task, computational power and cost, and adaptability. The more accurately one can assess students' responses, the more detailed and directive the feedback can be. However, to the extent

that assessment accuracy is dependent on creating materials and tasks that lead to a restricted set of acceptable responses, students may find the training less motivating and feel less interest toward the domain. But the more freedom available in the environment, the more difficult it is to accurately assess student responses and provide appropriate feedback. Greater freedom also requires more computational power and cost. One can try to increase motivation by creating a computer-based learning environment that exploits state-of-the-art natural language processing, artificial intelligence, cool interfaces, and graphics. Such a system, however, takes millions of dollars to develop and it is unclear how effective it would be in terms of learning gains and how well it would meet pedagogical goals. We do not believe simple is always better but it is sometimes good-enough to provide effective learning activities for well-understood skills.

ACKNOWLEDGMENT

This research was supported by the Institute of Education Sciences, U.S. Department of Education, through Grants R305H05133, R305G040055, and R305B070349 to Northern Illinois University. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of IES.

REFERENCES

Aleven, V., & Koedinger, K. R. (2002). An effective metacognitive strategy: Learning by doing and explaining with a computer-based cognitive tutor. *Cognitive Science*, *26*, 147–179. doi:10.1207/s15516709cog2602_1

Baker, R., Walonoski, J., Heffernan, N., Roll, I., Corbett, A., & Koedinger, K. (2008). Why students engage in "gaming the system" behavior in interactive learning environments. *Journal of Interactive Learning Research*, *19*(2), 185–224.

Baker, R. S. J. d., Corbett, A. T., Koedinger, K. R., & Roll, I. (2006). Generalizing detection of gaming the system across a tutoring curriculum. *8th International Conference on Intelligent Tutoring Systems*, (pp. 402-11).

Brainerd, C. J., & Reyna, V. F. (1992). Explaining "memory free" reasoning. *Psychological Science*, *3*(6), 332–339. doi:10.1111/j.1467-9280.1992.tb00042.x

Brainerd, C. J., & Reyna, V. F. (1998). When things that were never experienced are easier to "remember" than things that were. *Psychological Science*, *9*(6), 484–489. doi:10.1111/1467-9280.00089

Bransford, J. D., Brown, A. L., & Cocking, R. R. (Eds.). (2000). *How people learn*. Washington, DC: National Academy Press.

Bråten, I., Strømsø, H. I., & Samuelstuen, M. S. (2008). Are sophisticated students always better? The role of topic-specific personal epistemology in the understanding of multiple expository texts. *Contemporary Educational Psychology*, *33*, 814–840. doi:10.1016/j.cedpsych.2008.02.001

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 identify and use source information. *Cognition and Instruction*, *20*, 485–522. doi:10.1207/S1532690XCI2004_2

Britt, M. A., Kurby, C. A., Dandotkar, S., & Wolfe, C. R. (2008). I agreed with what? Memory for simple argument claims. *Discourse Processes*, *45*(1), 52–84. doi:10.1080/01638530701739207

Britt, M. A., Perfetti, C. A., Van Dyke, J., & Gabrys, G. (2000). The sourcer's apprentice: A tool for document-supported history instruction. In Stearns, P., Seixas, P., & Wineburg, S. (Eds.), *Knowing, teaching and learning history: National and international perspectives* (pp. 437–470). New York, NY: NYU Press.

Britt, M. A., Storey, J. K., Kopp, K., Dandotkar, S., & Larson, A. A. (2007). *Web-based tutorials to improve argumentative evaluation*. Poster presented at the 2007 IES National Research Conference. Washington, D.C.

Britt, M. A., Wiemer-Hastings, P., Larson, A. A., & Perfetti, C. A. (2004). Using intelligent feedback to improve sourcing and integration in students' essays. *International Journal of Artificial Intelligence in Education*, *14*, 359–374.

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

Craig, S. D., Sullins, J., Witherspoon, A., & Gholson, B. (2006). Deep-level reasoning questions effect: The role of dialog and deep-level reasoning questions during vicarious learning. *Cognition and Instruction*, *24*(4), 563–589. doi:10.1207/s1532690xci2404_4

Foltz, P. W., Kintsch, W., & Landauer, T. K. (1998). The measurement of textual coherence with latent semantic analysis. *Discourse Processes*, *25*(2&3), 285–307. doi:10.1080/01638539809545029

Gee, J. P. (2007). *What video games have to teach us about learning and literacy*. New York, NY: Palgrave Macmillan.

Gilliam, S., Magliano, J. P., Millis, K. K., Levinstein, I., & Boonthum, C. (2007). Assessing the format of the presentation of text in developing a reading strategy assessment tool (RSAT). *Behavior Research Methods, Instruments, & Computers*, *39*, 199–204. doi:10.3758/BF03193148

Graesser, A. C., Chipman, P., Haynes, B. C., & Olney, A. (2005). AutoTutor: An intelligent tutoring system with mixed-initiative dialogue. *IEEE Transactions on Education*, *48*, 612–618. doi:10.1109/TE.2005.856149

Graesser, A. C., Jackson, G. T., Mathews, E. C., Mitchell, H. H., Olney, A., & Ventura, M. … TRG. (2003). Why/AutoTutor: A test of learning gains from a physics tutor with natural language dialog. In R. Alterman & D. Hirsh (Eds.), *Proceedings of the 25th Annual Conference of the Cognitive Science Society* (pp. 1-6). Boston, MA: Cognitive Science Society.

Graesser, A. C., Lu, S., Jackson, G. T., Mitchell, H., Ventura, M., Olney, A., & Louwerse, M. M. (2004). AutoTutor: A tutor with dialogue in natural language. *Behavior Research Methods, Instruments, & Computers*, *36*, 180–193. doi:10.3758/BF03195563

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. doi:10.1037/0033-295X.101.3.371

Graesser, A. C., Wiemer-Hastings, K., Wiemer-Hastings, P., Kreuz, R., & The, T. R. G. (1999). AutoTutor: A simulation of a human tutor. *Journal of Cognitive Systems Research*, *1*, 35–51. doi:10.1016/S1389-0417(99)00005-4

Halpern, D. F. (2003). *Thought and knowledge: An introduction to critical thinking*. Mahwah, NJ: Erlbaum.

Hmelo-Silver, C. E. (2004). Problem-based learning: What and how do students learn? *Educational Psychology Review*, *16*, 235–266. doi:10.1023/B:EDPR.0000034022.16470.f3

Kintsch, E., Caccamise, D., Franzke, M., Johnson, N., & Dooley, S. (2007). Summary street: Computer-guided summary writing. In T. K. Landauer, D. M., McNamara, S. Dennis, & W. Kintsch (Eds.), *Latent semantic analysis* (pp. 263-277). Mahwah, NJ: Erlbaum.

Kintsch, W. (1988). The role of knowledge in discourse comprehension construction-integration model. *Psychological Review*, *95*, 163–182. doi:10.1037/0033-295X.95.2.163

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

Kintsch, W., & van Dijk, T. A. (1978). Toward a model of text comprehension and production. *Psychological Review*, *85*, 363–394. doi:10.1037/0033-295X.85.5.363

Kolodner, J. L., Camp, P. J., Crismond, D., Fasse, B., Gray, J., Holbrook, J., & Ryan, M. (2004). Promoting deep science learning through case-based reasoning: Rituals and practices in learning by design-super (TM) classrooms. In Seel, N. M., & Dijkstra, S. (Eds.), *Curriculum, plans, and processes in instructional design: International perspectives* (pp. 89–114). Mahwah, NJ: Erlbaum.

Kopp, K. J., Britt, M. A., Millis, K. K., Graesser, A. (In preparation). *Improving the efficiency of dialogue in tutoring.*

Kurby, C. A., Britt, M. A., & Dandotkar, S. (2006, July). *Representing argument claims: Availability and Accessibility of the predicate and implications on argument evaluation*. Poster presented for the 16th Annual Conference of the Society for Text and Discourse, Minneapolis, Minnesota.

Landauer, T. K., & Dumais, S. T. (1997). A solution to Plato's problem: The latent semantic analysis theory of acquisition, induction, and representation of knowledge. *Psychological Review*, *104*(2), 211–240. doi:10.1037/0033-295X.104.2.211

Larson, A. A., Britt, M. A., & Kurby, C. (2009). Improving students' evaluation of informal arguments. *Journal of Experimental Education*, *77*(4), 339–366. doi:10.3200/JEXE.77.4.339-366

Leelawong, K., & Biswas, G. (2008). Designing learning by teaching agents: The Betty's brain system. *International Journal of Artificial Intelligence in Education*, *18*(3), 181–208.

Magliano, J. P. (1999). Revealing inference processes during text comprehension. In Goldman, S. R., Graesser, A. C., & van den Broek, P. (Eds.), *Narrative comprehension, causality, and coherence: Essays in honor of Tom Trabasso* (pp. 55–75). Mahwah, NJ: Erlbaum.

Magliano, J. P., & Millis, K. K. (2003). Assessing reading skill with a think-aloud procedure and latent semantic analysis. *Cognition and Instruction*, *21*, 251–284. doi:10.1207/S1532690XCI2103_02

Magliano, J. P., & Millis, K. K. The RSAT Development Team, Levinstein, I., & Boonthum, C. (under review). *Assessing comprehension during reading with the reading strategy assessment Tool (RSAT).* An unpublished manuscript.

Magliano, J. P., Millis, K. K., Ozuru, Y., & McNamara, D. S. (2007). A multidimensional framework to evaluate reading assessment tools. In McNamara, D. S. (Ed.), *Reading comprehension strategies: Theories, interventions, and technologies* (pp. 107–136). Mahwah, NJ: Erlbaum.

Magliano, J. P., Trabasso, T., & Graesser, A. C. (1999). Strategic processes during comprehension. *Journal of Educational Psychology*, *91*, 615–629. doi:10.1037/0022-0663.91.4.615

Millis, K. K., Kim, H. J., Todaro, S., Magliano, J., Wiemer-Hastings, K., & McNamara, D. (2004). Identifying reading strategies using latent semantic analysis: Comparing semantic benchmarks. *Behavior Research Methods*, *36*, 213–221.

Millis, K. K., Magliano, J. P., & Todaro, S. (2006). Measuring discourse-level processes with verbal protocols and latent semantic analysis. *Scientific Studies of Reading*, *10*, 251–283. doi:10.1207/s1532799xssr1003_2

Millis, K. K., Magliano, J. P., Todaro, S., & McNamara, D. S. (2007). Assessing and improving comprehension with latent semantic analysis. To appear in T. Landauer, D.S., McNamara, S. Dennis, & W. Kintsch (Eds.), *LSA: A road to meaning*. Mahwah, NJ: Erlbaum.

Mislevy, R. J. (1993). Foundations of a new test theory. In Frederikson, N., Mislevy, R. J., & Bejar, I. I. (Eds.), *Tests theory for a new generation of tests*. Hillsdale, NJ: Lawrence Erlbaum Associates, Inc.

National Assessment of Education Progress: Prepublication version of Science Framework for the 2009 NAEP. (2006). Retrieved November 06, from http://www.nagb.org/pubs/pubs.html

Olson, G. M., Duffy, S. A., & Mack, R. L. (1984). Thinking-out-loud as a method for studying real-time comprehension processes. In Kieras, D., & Just, M. (Eds.), *New methods in the study of immediate processes in comprehension*. Hillsdale, NJ: Erlbaum.

Organisation for Economic Co-Operation and Development (OECD). (2002). *Reading for change. Performance and engagement across countries*. Paris, France: OECD Publications.

Pellegrino, J. W., & Chudowsky, N. (2003). The foundations of assessment. *Interdisciplinary Research and Perspectives*, *1*, 103–148. doi:10.1207/S15366359MEA0102_01

Pellegrino, J. W., Chudowsky, N., & Glaser, R. (Eds.). (2001). *Knowing what students know: The science of design of educational assessment*. Washington, DC: National Academy of Sciences.

Reichherzer, T., Cañas, A. J., Ford, K. M., & Hayes, P. J. (1998). The giant: An agent-based approach to knowledge construction and sharing. *FLAIRS Conference*, (pp. 136-140).

Ritter, S., Anderson, J. R., Koedinger, K. R., & Corbett, A. (2007). Cognitive tutor: Applied research in mathematics education. *Psychonomic Bulletin & Review*, *14*, 249–255. doi:10.3758/BF03194060

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. doi:10.1037/0022-0663.88.3.478

Rouet, J.-F., Favart, M., Britt, M. A., & Perfetti, C. A. (1997). Studying and using multiple documents in history: Effects of discipline expertise. *Cognition and Instruction*, *15*, 85–106. doi:10.1207/s1532690xci1501_3

Seixas, P. (1994). When psychologists discuss historical thinking: A historian's perspective. *Educational Psychologist*, *29*(2), 107–109. doi:10.1207/s15326985ep2902_6

Stadtler, M., & Bromme, R. (2008). Effects of the metacognitive computer-tool *met.a.ware* on the web search of laypersons. *Computers in Human Behavior*, *24*, 716–737. doi:10.1016/j.chb.2007.01.023

Storey, J. K., Kopp, K. J., Wiemer, K., Chipman, P., & Graesser, A. C. (in press). Using AutoTutor to teach scientific critical thinking skills. *Behavior Research Methods*.

Toulmin, S. E. (1958). *The uses of argument*. Cambridge, MA: Cambridge University Press.

Trabasso, T., & Magliano, J. P. (1996). Conscious understanding during comprehension. *Discourse Processes*, *21*, 255–287. doi:10.1080/01638539609544959

Voss, J. F., & Means, M. L. (1991). Learning to reason via instruction in argumentation. *Learning and Instruction*, *1*(4), 337–350. doi:10.1016/0959-4752(91)90013-X

Weizenbaum, J. (1966). Eliza-A computer program for the study of natural language communication between man and machine. *Communications of the ACM*, *9*, 36–45. doi:10.1145/365153.365168

Wiley, J., Goldman, S. R., Graesser, A. C., Sanchez, C. A., Ash, I. K., & Hemmerich, J. (2009). Source evaluation, comprehension, and learning in Internet science inquiry tasks. *American Educational Research Journal*, *46*(4), 1060–1106. doi:10.3102/0002831209333183

Wineburg, 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. doi:10.1037/0022-0663.83.1.73

Wineburg, S. S. (2000). Making historical sense. In Stearns, P., Seixas, P., & Wineburg, S. (Eds.), *Knowing, teaching and learning history: National and international perspectives* (pp. 306–325). New York, NY: NYU Press.

Wolfe, M. B. W., & Goldman, S. R. (2005). Relations between adolescents' text processing and reasoning. *Cognition and Instruction*, *23*, 467–502. doi:10.1207/s1532690xci2304_2

Yee, N. (2006). The labor of fun: How video games blur the boundaries of work and play. *Games and Culture*, *1*(1), 68–71. doi:10.1177/1555412005281819