Using Intelligent Feedback to improve Sourcing and Integration in Students' Essays

M. Anne Britt, *363 Psychology-Math Building, Northern Illinois University, DeKalb, Il 60115, USA*
britt@niu.edu
Peter Wiemer-Hastings*, School of Computer Science, Telecommunications, and Information Systems, DePaul University, 243 S. Wabash, Chicago, IL 60604, USA*
Aaron A. Larson, *363 Psychology-Math Building, Northern Illinois University, deKalb, Il 60115, USA*
Charles A. Perfetti, *Learning Research & Development Center, University of Pittsburgh, Pittsburgh, PA 15260 USA*

Abstract. Learning and reasoning from multiple documents requires students to employ the skills of sourcing (i.e., attending to and citing sources) and information integration (i.e., making connections among content from different sources). Sourcer's Apprentice Intelligent Feedback mechanism (SAIF) is a tool for providing students with automatic and immediate feedback on their use of these skills during the writing process. SAIF uses Latent Semantic Analysis (LSA), a string-matching technique and a pattern-matching algorithm to identify problems in students' essays. These problems include plagiarism, uncited quotation, lack of citations, and limited content integration. SAIF provides feedback and constructs examples to demonstrate explicit citations to help students improve their essays. In addition to describing SAIF, we also present the results of two experiments. In the first experiment, SAIF was found to detect source identification and integration problems in student essays at a comparable level to human raters. The second experiment tested the effectiveness of SAIF in helping students write better essays. Students given SAIF feedback included more explicit citations in their essays than students given sourcing-reminder instructions or a simple prompt to revise.

INTRODUCTION

Essay writing is an important skill for students to learn and use. It provides an opportunity for students to transform and integrate information (Hemmerich, & Wiley, 2002; Voss, & Wiley, 2001; Wiley, 2001) and to self-explain which leads to better learning (Chi, DeLeeuw, Chiu, & LaVancher, 1994). Of course, writing is important in its own right as evidenced by the abundance of programs for writing across the curriculum in U.S. colleges and universities (Ackerman, 1993) and the inclusion of writing assessment in post-secondary (SAT) and graduate program (GRE) admissions (The College Board, 2004; ETS, 2004). Students learn to write by engaging in writing often and receiving feedback from teachers. Sometimes teachers include an opportunity for students to revise their work after constructive feedback.

However, because it is very time consuming to prepare helpful feedback, students receive fewer opportunities for revision than would be desirable. In this paper, we describe an automated tool that provides students with feedback about their citations and use of sources. The program is not proposed as a grader but as a tool that the students can use to evaluate one aspect of their writing and give them specific instructions on what should be revised, just as they do with automated spelling and grammar checkers. This type of tool encourages students to revise, thereby leading to higher quality essays.

Research by Britt and colleagues has found that students' research papers can greatly benefit from modest amounts of additional instruction and practice at sourcing and integrating information from various documents (Rouet, Britt, Mason, & Perfetti, 1996; Britt, Rouet, & Perfetti, 1996; Britt & Aglinskas, 2002). Sourcing is the skill of attending to a document's source and later making explicit citations to a document when mentioning information from that document (Wineburg, 1991). An explicit citation means that the writer mentioned enough information about the source (i.e., author and document) to indicate to whom the content is attributed. In an informal assessment of college students' general skill in sourcing, 108 undergraduates were asked to read a variety of primary and secondary documents (approximately 1000 words) that supported different perspectives on a historical controversy. Each document was introduced by detailed information about the author and document type. Participants were able to take notes while reading. When they decided they had learned enough, the texts were removed and students wrote an opinion essay with only their notes available. We found that, on average, participants included only 0.71 references per essay and only about half of these (M = 0.36) were explicit, i.e., clearly attributed to an identifiable source. Only 28% of the essays included at least one explicit reference. Considering that no participants made more than 2 explicit references, it appears that undergraduates are not fully proficient at sourcing. A second problem identified in these essays was failure to explicitly cite the source of a quotation. None of the essays that included quotations (7.4%) explicitly marked the source of the quotation. Without an explicit citation, the reader is not able to verify the accuracy of the quotation or evaluate its credibility. This general failure to cite sources lead to the development of the Sourcer's Apprentice tutorial and practice environment (Britt, Perfetti, Van Dyke, & Gabrys, 2000). This environment addresses students' knowledge and awareness of document and source characteristics while studying a document but it does not directly address students' use of sources while essay writing. We describe in the next section, our use of intelligent analysis to provide students with specific feedback on their use of sources in their essays.

SOURCER'S APPRENTICE INTELLIGENT FEEDBACK MECHANISM

Sourcer's Apprentice (SA)

The Sourcer's Apprentice is a computer environment that provides direct instruction and practice to help students develop the skill of sourcing while reading multiple documents (Britt, et.al., 2000). Students are first given explicit instruction on how to identify important source features (e.g., who the author is, when the document was written, etc) and they then are given a controversial episode in history to learn about by reading excerpts from actual documents. They select books to read from a virtual bookshelf on the computer screen and use structured notecards at the bottom of the screen to take notes on source and content information for each document. Students enter information into the notecards by dragging a text segment from a document and dropping it onto screen targets. For example, to obtain points for a correct answer for one document, Andrew Carnegie's Autobiography, would require the student to drag the author's name "Andrew Carnegie" into the *who bucket* and "autobiography" into the *document type bucket*. Immediately after dropping a response into the notecard bucket, participants were given positive feedback and points for correct answers. If the answer was wrong, they were given graduated hints to aid them in correctly identifying the source information. After learning about the controversy by reading the texts and filling in the notecards, students then answer several source and content questions and write a short essay on the controversy.

The Sourcer's Apprentice was built on a foundation of six instructional-design principles derived from the literature on successful tutoring: support student learning via problem solving, support expert representations, support task decomposition, support transfer, provide explicit instruction, and motivate engagement. To date, the Sourcer's Apprentice has been used by all levels of high-school history classes: from "mainstream" to advanced placement (Britt, et al, 2000) and has been shown to improve sourcing on a transfer test compared to regular classroom activity or a textbook-centered version of the same material (Britt & Aglinskas, 2002). Analysis of the students' essays found that students who used the system included more explicit citations and integrated material from more distinct sources. These students also mentioned an average of 3.38 (of seven relevant and informative documents) citations per essay while students reading the textbook-centered material only included 0.47 citations.

One limitation of the Sourcer's Apprentice is a lack of feedback and instruction for essay writing. Students were simply instructed to write an essay and were not provided with support. SA could be improved by providing immediate and automatic feedback to students on the quality of their essays. For example, prompting students to re-word possible plagiarized phrases or explicitly cite a minimum number of sources (e.g., "According to Carnegie's autobiography"). Providing immediate feedback requires a way to automatically process student's essays. We adopt a strategy of combining string matching and similarity comparison using Latent Semantic Analysis (LSA).

Latent Semantic Analysis

LSA uses a large body of text to derive a high-dimensional space to describe semantic relatedness (Foltz, Laham, & Landauer, 1999). The resulting metric of semantic similarity enables the computation of a variety of factors which are useful for giving immediate feedback on students' use of source material.

Latent Semantic Analysis begins by constructing a large co-occurrence matrix of words and documents from a large and varied set of texts within some domain. Each cell in this matrix is a weighted count of the number of times the corresponding word occurs in the corresponding document. This matrix is then processed with a matrix algebra technique called Singular Value Decomposition (SVD). SVD transforms the original data by reordering the dimensions and sorting them. The original data can then be approximated by maintaining only the most significant dimensions. The exact number must be empirically derived, but many studies have used approximately 300 dimensions. The rationale for this reduction of data is that it captures the semantically important concepts, and eliminates the noise from the training data. In this resulting space, each word and each text is represented as a vector. The cosine between vectors in this space gives a measure of the semantic distance between the corresponding texts. In practice, the cosines range from 0 to 1 and can be treated like correlations. If two words or phrases have a high cosine (e.g., greater than 0.70), then they can be considered to be highly similar or related. If they have a low cosine (e.g., less than 0.20), then they are unrelated. A complete description of LSA can be found elsewhere (Deerwester, Dumais, Furnas, Landauer, & Harshman, 1990; Landauer & Dumais, 1997).

LSA has been used to provide 6th grade students feedback on revising written summaries (Kintsch, Steinhart, Stahl, LSA Research Group, Matthews, &, Lamb, 2000), to evaluate and grade student essays (Foltz, Britt, & Perfetti, 1996; Foltz, Gilliam, & Kendall, 2000; Foltz, et.al., 1999), to give feedback on student stories and essays (Wiemer-Hastings and Graesser, 2000; Foltz, et.al., 2000), and to evaluate the quality of student responses to an automated tutor (Graesser, Wiemer-Hastings, Wiemer-Hastings, Harter, Tutoring Research Group & Person, 2000).

Linking LSA capabilities to the Sourcer's Apprentice enables the system to give students automatic and immediate feedback on their essays (see below) before submitting them for grading. We refer to this tool as the Sourcer's Apprentice Intelligent Feedback mechanism (SAIF). The LSA space was created for the 1892 Homestead Steel strike controversy described in Britt and Aglinskas (2002). This document set contains excerpts from seven authentic documents including excerpts from a high-school textbook, historian essays, participant accounts, and a congressional committee report. The documents describe the events that happened in the summer of 1892 at the Homestead steelworks and to what extent the owner, Andrew Carnegie, was responsible for breaking the workers' union. The LSA semantic space was derived from the 7 target documents (2586 words) and 10 general texts on the Homestead Steel strike (29244 words) including excerpts from encyclopedias, newspaper articles, textbooks, historians' essays and participants' accounts.

Essay deficiencies targeted by SAIF

We identified several problems with students' essays that we wanted to address using SAIF. These are listed in Table 1 along with the intended goal for feedback that addresses each problem. First, students frequently include *unsourced copied material* in their essays. By this we mean either plagiarism (1a in Table 1) or quoting material without explicitly stating the source of the quote (1b in Table 1). Feedback encourages students to state content in their own words and explicitly cite the source of quoted or paraphrased material. Second, students often fail to include an acceptable number of *explicit citations* such as "According to Carnegie's autobiography" (2 in Table 1) in their essays. Feedback prompts students to make at least 3 distinct citations. We suggest a minimum number of three citations because an informal review of requirements for student research papers shows that high schools teachers expect about 3 citations and college teachers expect between 3 and 5 citations. This parameter could of course be easily set depending on expectations by the instructor. Third, the number of *distinct sources mentioned* in student essays (3 in Table 1) is generally low. That is, the cited sources are limited to one or two documents and therefore the essay will not be well-rounded. Feedback reminds students to integrate information from more than a single source. Fourth, students may also rely on *excessive quoting* rather than attempting to put the material in their own words. Feedback prompts students to paraphrase more rather then relying on quoting. Paraphrasing will support a deeper understanding of the material in cases where the quote is used to present content rather than as supporting evidence. There are differences among disciplines in the acceptable number of quotes, so this may be an important parameter to adjust. Finally, students often lack adequate *content integration from multiple sources* in that they fail to sample material from a variety of different documents. Feedback reminds students to integrate information from multiple sources.

Table 1
Types of problems SAIF addresses and the intended goal of feedback

Problem	Feedback prompts student to:
1a. Unsourced copied material (plagiarism)	Reword plagiarism and model proper format.
1b. Unsourced copied material (quotation)	Explicitly credit source and model proper format.
2. Explicit citations	Explicitly make a minimum of 3 citations.
3. Distinct sources mentioned	Cite at least 2 different sources.
4. Excessive quoting	Paraphrase more instead of relying on quotations too heavily.
5. Integration from multiple sources	Include a more complete coverage of the documents in set.

We selected problems general to learning to write research papers in history. It is expected, however, that these types of problems are common to research papers in many domains (e.g., History, English, Psychology). Many domains require students to learn to cite sources and integrate information from multiple documents. Domain differences would not be so much a matter of the required skills in general, but rather the actual parameters expected for each skill. For example, a research paper for an English class may expect more quotation while the same type of paper for a psychology class may expect little to no quotation. Another possible difference among disciplines is the required format of the citation (e.g., MLA or APA style). This would entail differences in the pattern selected for string matching. Therefore, many domain differences will be a matter of parameter setting rather than differences in the types of skills expected to be mastered.

SAIF identifies each of these problems in a student's essay using a variety of mechanisms. It uses a robust string-matching technique to identify the names of authors and books that occur in the student's essay. It uses a regular expression pattern matching algorithm to recognize phrases used in citations. Example citation indicators include the following expressions: "according to" or "as stated in" or "in his book" followed by an author's name. Finally, it uses LSA both to further identify citations and to identify material that was taken from the source texts. Using these sources of information, a rule-based decision mechanism determines what problems exist in each essay and what suggestions to make on how the student can improve his or her essay. The 108 student essays mentioned previously were used to develop and calibrate SAIF. In the next section, we describe how SAIF identifies each essay problem and the type of corrective feedback it provides when a problem was identified in an essay.

Plagiarism or unsourced copied material

SAIF identifies prospective cases of plagiarism by comparing each essay sentence to the source sentences in the LSA space. The higher the cosine, the more similar the two sentences are. A cosine of 1.00 indicates the two statements are identical (or at least have identical words). For our purposes, we identified pairs that had a LSA cosine of 0.75 or higher and lacked an explicit citation (e.g., "In his letter, Carnegie writes"). This value will be justified in the next paragraph. This sentence-by-sentence comparison using LSA cosines can identify the two most common types of plagiarism. One type occurs when students merely try to re-order information within a sentence. LSA can easily identify this type of plagiarism because LSA makes order-independent comparisons. A second type of plagiarism occurs when students simply substitute one or more synonyms for content words in the target sentence. LSA can also identify these instances because LSA can abstract similarities among concepts to detect synonyms.

The threshold cosine of 0.75 was used to detect possible plagiarism. We empirically selected a value that would catch all very near matches, not just those that were verbatim matches. This was done to prompt students to re-word even close matches. Students learning to write supported essays, especially on an unfamiliar topic, may find it difficult to put statements in their own terms and are often unaware of what actually counts as plagiarism. Therefore, identifying possible instances of plagiarism is an important initial step in avoiding it. Early in the learning process, we feel it is best to err on the side of prompting for paraphrasing and transforming the information from the documents. This will force students to construct their own representation of the material and lead to deeper understanding (Hemmerich & Wiley, 2002; Voss & Wiley, 2001; Wiley, 2001; Chi, et al, 1994). It will also help students become more aware of what constitutes plagiarism.

SAIF provides corrective feedback for two types of *unsourced copied material*, plagiarism and uncredited quotations. First, if the "copied" material is not in quotation marks and exceeds the LSA threshold (0.75), then it is marked as possible plagiarism. All such sentences are then listed along with the sentence that is determined to be too similar to the student's statement. For example, an essay that included the sentence "The mischief was done, the works were in the hands of the Governor; it was too late", would receive the following corrective feedback from SAIF:

> *Unsourced copied material (plagiarism)*. This might be plagiarism (unsourced copied material): (p=1.0: "Carnegie Autobiography", The mischief was done, the works were in the hands of the Governor; it was too late)
> To correct this, either make this a quotation with an explicit source citation as in the example below or restate the information in your own words, also with an explicit source citation.
>
> *Example citation*: As Carnegie stated in his autobiography, "The mischief was done, the works were in the hands of the Governor; it was too late".

Second, if the "copied" material includes quotation marks and exceeds the threshold, then it is marked as an uncredited quotation. Students are prompted to credit the source for this quoted material. For example, if the same essay sentence was in quotation marks, the corrective feedback would be:

Unsourced copied material (uncredited quotation). May want to credit the source of this information "in the hands of the Governor; it was too late". (p=1.0: "Carnegie Autobiography", The mischief was done, the works were in the hands of the Governor; it was too late)

To correct this, add an explicit source citation to the quotation as in the example below.

Example citation: "The mischief was done, the works were in the hands of the Governor; it was too late" (Carnegie Autobiography).

If there is at least one instance of either type of unsourced copied material, then SAIF also presents a dynamic or constructed example of a proper citation of the material for the first instance (as shown in the *Example citation* above). To create this example, SAIF determines the source document and then formats the author information correctly and attaches the quoted material. The student can then use this example as a model to transfer the details of sourcing to the other instances of possible plagiarism.

Citation

SAIF locates references to sources using three complementary methods. The first is an *author-name approach* that checks for a string match with any form of the author's name excluding any authors who are also characters in the narrative. For instance, Carnegie is both a participant and an author so SAIF does not include his name in the comparison set. Any essay sentence that includes the name of an author is assumed to be a citation. The second source-identification method checks material in parentheses for author or title information by either a string match or an LSA match above 0.80. Here, the expected match with the corresponding text was relatively high, so we empirically determined a threshold that caught most of the matches, but allowed for a small amount of noise. The third source-identification method checks for special source identifiers such as "according to", "as stated in", "stated/states", "claimed/claims", "explained/explains", or "in book". While none of these methods will result in perfect source identification, the combination should ensure that nearly all explicit citations are counted.

The author-name approach presents several interesting challenges that are worth mentioning. First, it may over-estimate source citations in cases where the student discusses the credentials or credibility of an author or document (e.g., "Bridge is more believable because he has nothing to lose by telling the truth about Carnegie"). Such meta-source comments evaluating the source (e.g., author credentials or bias, type of document) are an important element of sourcing, we do not consider this a negative. Occasionally, however, students also mention a primary document as an object without mentioning the content. For example, one document that students read was a draft notice that Carnegie sent to Frick to present to the Homestead workers. If a student mentions this document as part of the story without mentioning the document's content (e.g., "April 1892 he sent Frick the draft of notice meant for Homestead workers."), then it should not be considered a source citation. In such cases, which occur very infrequently, SAIF will incorrectly classify these object mentions as a case of sourcing. Second, the author-name approach may be a problem when an involved participant is an author. For instance, the essays used to test SAIF were written after students read a document set that discussed Andrew Carnegie's role in the strike which included two primary documents written by Carnegie himself. This makes it impossible to consider all instances of the author's name as a citation. In such cases, document type (e.g., autobiography, notice) was used to indicate these citations rather than author name. Finally, the author-name method may not work in its present form if one wants to also verify the accuracy of the cited material. On several occasions, the student cited the author in an introductory sentence without an explicit mention of the content of the document. Then in subsequent sentences the content was mentioned with the only source marking being a pronoun. This does not cause a problem for SAIF because our primary goal is to ensure that a minimal number of citations are included in the essay. This would, however, pose a problem for a program that attempts to check the accuracy of the cited material when the source information is mentioned in one sentence and the content is mentioned in another sentence.

The goal in designing SAIF was to identify only explicit references and meta-source comments. Because our target users are just beginning to learn to cite properly, the decision was made to not identify implicit (e.g., "Carnegie said he was hurt deeply by this."), vague (e.g., "Some believe Andrew Carnegie, the owner of the steel works, is to blame.") or incomplete citations (e.g., "Another excerpt portrayed Carnegie as a conspirator with Frick who went to Scotland and made himself unreachable on purpose because he knew what was going to happen."). Although an analysis of historians' essays indicates that many citations are implicit, the reader can give the expert the benefit of doubt that they have actually read all the primary documents and studied the evidence very carefully. As such, when an expert merely states that a participant made a statement without explicitly mentioning how they knew this, the reader can somewhat safely assume that the expert knows the source but is not mentioning it presently for the flow of the text, or perhaps mentioned the source in a previous chapter. This is one case in which it is best for novices not to model their writing completely after experts. This bias in SAIF toward identifying only explicit references is mentioned when students are given corrective citation feedback.

In addition to providing feedback for plagiarism and uncredited quotations, SAIF also evaluates the total number of explicit citations. If the essay did not include at least 3 explicit citations, the following corrective feedback is provided:

> *Number of Citations*. We did not find a base level of explicit source citation such as "According to Krause". Citing sources gives more weight to your interpretation, helps others locate this information for themselves, and gives credit to the author. If you did cite sources, they may be vague, incomplete, or in a form that the Sourcer's Apprentice can't identify. Double check your essay for these problems and for places where citations and quotations can be added.

Note that students are told that this feedback does not mean that citations were not included; it may mean that they weren't explicit enough. The feedback also reminds the student of the purpose of citing and provides an example of an explicit citation. If the essay included the minimum of 3 explicit citations, then the feedback reports that sourcing was positive (i.e., The number of citations looks good).

Sources

The sources category counts the number of different sources cited explicitly in the students' essays to ensure that at least two different sources are used. For example, mentioning Krause and Carnegie's autobiography would count as two sources while mentioning Krause twice would count as one source. This will help the student write an essay that incorporates more than a single perspective on the topic. SAIF uses the same methods as mentioned in the citation section to identify an explicit source and checks if this number is greater than 1. If it is not greater than 1, the student is given corrective feedback which directs the student to consider information provided in additional sources. For example, if the essay only includes an explicit reference to Carnegie's Autobiography, then the corrective feedback would be:

> *Number of Sources*. This essay appears to have citations from ("Carnegie Autobiography"). Be sure to use other sources to put this author's ideas in context or present a well-rounded account.

If the student's essay includes a citation from at least two different sources, SAIF provides positive feedback (i.e., The number of sources cited looks good.).

Quoting

Quoted material was identified through a simple pattern match with the constraint that the string must exceed a single word. From a cursory analysis of past student essays using SA, we found single-word quoting served a different function, that is, to emphasize or comment on the meaning of the word. Conversely, single word quoting is seldom appropriate for source. The threshold for determining whether an essay had excessive quoting was set at 50%. This means that if more than half of the sentences in an essay included a quote then this essay would be given a warning for too much quoting. An essay with more than 50% quoting would receive the following corrective feedback:

> *Number of Quotations*. This essay seems to have a lot of quoted material. Don't lose your own voice and ideas by quoting too much. Try paraphrasing more and quoting less.

If, however, 50% or less of the essay sentences include quotation marks, then the student is told that the number is not too high (i.e., The number of quotations seems appropriate). This threshold level was decided upon because our target users are either novice or intermediate essay writers in the domain of history. As such, they may have a tendency to rely on others' words, thinking that they could never put it as perfectly as the author they are quoting. Therefore, the threshold is set low to make sure that students are prompted to rely on their own words. Different disciplines would most likely set the acceptable proportion of quoted sentences at a different level.

Integration of sources

A final problem with student essays is incomplete coverage of documents. Students may rely too heavily on a single document or perspective. SAIF can prompt students to sample content from several sources to ensure that they have adequately covered the evidence that must be explained. If at least one sentence from a student's essay has an LSA cosine above 0.60 with a sentence from a document from the studied set, then that document is included. The student must include information from at least three different sources in order to not receive corrective feedback. Example of corrective feedback for essays with insufficient integration of multiple sources is:

> *Integration of Sources*. It looks like this essay has information from a limited number of specific sources: "Krause". This may give your reader a limited view of the story. Go back and look for places to introduce information from other sources.

Here they are told which documents SAIF considers covered and are prompted to include others from the document set. If the essay samples from enough of the documents, then positive feedback is given (i.e., It looks like you have done well in integrating material from a number of sources into your essay.).

Finally, SAIF provides special additional feedback to any essay that receives corrective feedback on fewer than two of the five problems.

> It appears as though you have a comfortable grasp of the basics of citing sources in an essay. Now try to explore more stylistic concerns such as varying citation formats and providing interesting contexts to set up your quotations.

This feedback tells students that they have only passed a minimal level of sourcing and are invited to continue to develop their skill. The special feedback for essays that did not have several problems identified was intended to remind the student that SAIF only identifies serious errors and that there are more subtle aspects to the development of each skill.

EVALUATION OF SAIF IN SCORING OF ESSAYS

In order to test the effectiveness of SAIF in detecting problems with actual student essays, we did a detailed analysis of 23 high-school students' essays from one of the experiments described in Britt & Aglinskas (2002). These essays were not used in the SAIF development process. In this experiment, high-school students read excerpts from seven documents (e.g., textbook, historian essays, participant accounts, and a committee report) describing the events surrounding the 1892 Homestead steel strike. Students read these documents either using the Sourcer's Apprentice or paper booklets. Then they wrote an essay taking a stance on the controversy, "To what extent was Carnegie responsible for breaking the union at Homestead?" The first author of this paper conducted a detailed analysis of these essays and those judgments were compared to judgments made by SAIF. These results are shown in Table 2. Considering all judgments of plagiarism by either the human rater or SAIF, the interrater reliability was high (Cronbach's alpha of 0.84). Now considering only the 42 statements that SAIF identified as plagiarism (i.e., Unsourced Copied Material - plagiarism), they were also judged by the rater as plagiarism 81% of the time. A high agreement is important because we do not want SAIF to unnecessarily warn students that they may be plagiarizing which could be annoying if the student was not actually doing so. Second, for all material identified by SAIF as a quotation, again the interrater reliability was high (Cronbach's alpha of 0.91). Of those unsourced quotations identified by SAIF (i.e., Unsourced Copied Material - quotes), SAIF and the rater agreed 80% of time. Finally, an evaluation of each sentence was conducted to determine which document was the source of the information. Considering all essay statements, the interrater reliability between the actual source identified by the two raters was high (Cronbach's alpha of 0.99). SAIF identified 106 essay statements that had an above-threshold cosine, signifying that it overlapped with a sentence from the material read. Thus, SAIF had concluded that this document was at least minimally integrated into the essay. This information is useful in determining the number of sources and the integration of sources. SAIF and the rater agreed on the source of these statements 91% of time.

Table 2
Agreement between SAIF and human raters in identifying sourcing problems in essays

	Agreement	
Problem type	SAIF- Human	
Unsourced Copied Material (plagiarism)	81%	
Unsourced Copied Material (quotation)	80%	
Sources and Integration	91-96%	
	SAIF- Humans	Human-Human
Explicit citations	76%	86%
Incomplete citations	23%	91%

The student essays did not include a large number of citations, so we had an additional 27 undergraduates from Northern Illinois University read the texts and write essays on the controversy. This produced a total of 50 essays to use for detecting citations. Our goal for SAIF was to count all explicit citations without counting the incomplete citations which students should be prompted to make more explicit. To evaluate the accuracy of SAIF's detection of explicit citations, comparisons were made with the ratings of two of the authors of this paper. Judgments were made by each rater independently of each other and before SAIF scored the essays. SAIF's identification of explicit citations was acceptably close to the raters (76%) compared to the agreement of the two raters with each other (86%). The interrater reliability for the 3 raters was high (Cronbach's alpha of 0.85). This high agreement is important because students may become frustrated receiving feedback that they are not citing sources when in fact they are.

SAIF only identifies explicit citations in an attempt to help students make their citations more clear. The human raters, however, were able to identify incomplete citations. The two raters agreed on the classification of incomplete citation 91% of the time, while SAIF rarely agreed that these were citations (23%). This low identification by SAIF is also indicated by a low interrater reliability coefficient (Cronbach's alpha of 0.51). The combination of these results for citation identification is important. First, it is critical that SAIF reliably identify those citations that humans deem to be explicit enough. This will prevent unnecessary prompting for improving already acceptable citations. As shown, the reliability here is high. Second, it is critical that citations that were not explicit enough, due to vagueness or incompleteness, were not counted as explicit. These are precisely the type of citation that students need to be prompted to improve. If SAIF identified these as explicit, then the students would not be properly prompted. As shown, the reliability here is low, meaning that many of these citations deemed implicit by humans were not identified as explicit and would result in prompting by SAIF for a more complete citation.

These results are very encouraging, suggesting that a rather simple mechanism can be employed to detect common problems with student essays. The next obvious question is whether it would be effective in helping students write better essays.

EVALUATION OF SAIF'S EFFECTIVENESS

In order to test the effectiveness of SAIF in helping students write better essays, 60 Northern Illinois University students used the Sourcer's Apprentice to receive a tutorial and read the seven documents of the Homestead problem. Participants were able to take paper and pencil notes and then, using only their notes, they were asked to write an essay on the controversy. All conditions were identical until this point. After they clicked a button to submit their essay for grading, the subjects were randomly given one of three feedback instructions. In the Revise feedback condition, participants were told to take this opportunity to revise their essay before submitting it. The actual instructions were:

> At this time you are given an opportunity to go back and revise your essay before submitting it. When you feel you are done revising your essay, click the Grade button.

In the Sourcing Reminder feedback condition, participants were reminded about how to source and why it is important. Then they were told to take this opportunity to revise their essay.

> As a reminder, citing sources such as "According to Krause…." gives more weight to your interpretation, helps others locate this information for themselves, and gives credit to the author. At this time you are given an opportunity to go back and revise your essay before submitting it. When you feel you are done revising your essay, click the Grade button..

Finally, in the SAIF feedback condition, participants were given individual feedback computed by SAIF and then were told to revise their essay. Possible corrective and positive feedback was described earlier.

The explicit citations in each essay were scored by two independent raters, scored blind to condition. One rater was the first author of this paper. The other rater was independent of this project and skilled at evaluating student essays, having taught undergraduates to write research papers in both the English and Psychology departments at

Northern Illinois University. The inter-rater agreement between the two raters was high (94% agreement and a Cronbach's alpha of 0.92). As shown in Table 3, the SAIF essays included approximately 2 more source statements than did the other two feedback conditions. A between-subjects ANOVA revealed a significant main effect of Type of Feedback (*F* (2, 57) = 3.77, *p* < .05). A Newman-Keuls post-hoc test (experiment-wise alpha = .05) showed that the SAIF essays had significantly more explicit references to sources than either of the other two feedback groups, which did not differ from each other. Directly pointing out problems with the students' essays and providing an example led to improved use of sources over even a general sourcing reminder. This is a remarkable improvement given that all participants were given the sourcing tutorial, an environment to support sourcing (SA), and instructions to revise their essay.

Table 3
Descriptive statistics for the number of source citations mentioned in the student essays depending on feedback instructions

Feedback Instructions	Mean	Std. Dev	Range
Revise	2.05	2.14	0 - 7
Sourcing reminder	2.35	1.84	0 - 5
SAIF feedback	3.80	2.44	1 - 8

SUMMARY

SAIF is a method for providing students with immediate and automatically generated feedback on the adequacy of sourcing and integration in their essays. Using LSA and pattern matching techniques, SAIF was able to satisfactorily identify sentences that were plagiarized or quoted without a citation (approximately 80%). It also was very good at identifying which document particular essay's statements came from (approximately 90%) to provide feedback on coverage and plagiarism. Finally, SAIF's classification of explicit citations agreed with human raters relatively well and, equally important, SAIF did not count as explicit citation statements that raters judged to be implicit or incomplete. Furthermore, we found that SAIF actually did provide helpful feedback to students in the revision process. The essays written after SAIF feedback included more explicit citations than essays written after sourcing reminder instructions or a simple prompt to revise. This approximately one and a half additional citations is important because the increase is not compared to students receiving no training. Both comparison groups had SA training and a prompt to revise. Furthermore, if students are expected to cite about 3 to 5 sources in college research papers, then SAIF can help the student move into that range.

The improvement due to SAIF in this study and due to SA in the Britt and Aglinskas (2002) study leads to the question of what the "optimal" level of sourcing is in a research paper. Our informal survey of research paper requirements suggests a minimum of 3 is references for high school students and 3 to 5 for college students. Moving students beyond the minimum number is important and perhaps by gradually increasing the minimum number expected by SAIF over the course of the term could help students continually improve their sourcing skills. In terms of absolute numbers, it is impossible to compare the current citation level to those from the Britt and Aglinskas (2002) study because the populations were so different. In the current study, the students were undergraduate students in an introductory Psychology class in an experimental setting. In the Britt and Aglinskas (2002) study, the students were completing a year-long History course that required a research paper with a minimum of 3 citations and the controversy was covered in the part of the term that it would normally occur. Students using SA without feedback made an average of 3.38 citations compared to 0.47 without SA (Britt & Aglinskas, 2002). It is an empirical question whether this average could be improved further by SAIF but we suspect it would be based on the current results.

A program such as SAIF enables students to receive personal and immediate feedback on two important aspects of essay writing: sourcing and integration. It enables students to mark their progress in the development of these skills. It also enables students to give teachers more polished products in much the same way that grammar and spelling checkers raise the minimal requirement. Teachers can then focus more of their attention on essay content and higher level writing skills.

SAIF can be further extended in several ways. Presently, SAIF does not verify the accuracy of the sourced information. It would be a relatively easy process to check sourced quotations. It would be much more difficult for implicit citations. As previously mentioned, the scope of a citation-content link may often span more than one sentence. Using the sentence as a unit makes this impossible presently.

Another area for improvement is in checking student's incorporation of information from the author and document pages. Each document in the Sourcer's Apprentice contains a page describing the author and the type of document. SAIF could provide feedback as to whether this information was included as qualifiers to citations or as separate meta-source statements (e.g., "*Carnegie is not a reliable source*, and therefore can not be believed on what he said").

Finally, SAIF could be extended to a third related skill: corroboration. This is when a reader verifies the accuracy of information by checking whether it is consistent or inconsistent with information from an independent source. From the perspective of a writer, corroboration is an important factor in knowing when to cite. For instance, it is more important to cite the source of unique information (i.e., mentioned only in one source) than facts or events mentioned by multiple authors. A program such as SAIF could check whether certain target statements, identified by a human as important and unique, include an explicit citation if mentioned in an essay.

In conclusion, the addition of SAIF to the Sourcer's Apprentice appears to be a valuable teaching mechanism which follows the approach of systems like Belvedere (Suthers, Connelly, Lesgold, Paolucci, Toth, Toth, and Weiner, 2001) and StoryStation (Robertson and Wiemer-Hastings, 2001). These systems engage students in authentic tasks and provide evaluative and corrective feedback on their products. By creating dynamic feedback that incorporates elements of the students' essays, SAIF produces a strong learning effect. And especially with the increasing heterogeneity of student populations and cultural backgrounds, a system that effectively teaches sourcing standards and techniques is very valuable. Such systems as a SAIF enhanced SA program can be made available in the high-school library or a homework website as a resource for teachers to send students having particular difficulty with sourcing and content integration in writing research papers.

ACKNOWLEDGEMENTS

We are grateful to the J. S. McDonnell Foundation's Cognitive Studies in Education grant that supported the work on which this article is based. The opinions expressed do not necessarily reflect the position or policy of the McDonnell Foundation, and no official endorsement should be inferred. We also wish to thank Jodie Sommer for help in scoring data and are grateful to Gareth Gabrys, Jodie Sommer and three anonymous reviewers for helpful comments on the paper.

REFERENCES

Ackerman, J. M. (1993). The Promise of Writing to Learn. *Written Communication,* 10(3), 334-370.

Britt, M.A., & Aglinskas, C. (2002). Improving student's ability to use source information. *Cognition and Instruction,* 20(40), 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, P. Seixas, & S. Wineburg (Ed.) *Knowing, Teaching and Learning History: National and International Perspectives*. New York: NYU Press.

Britt, M.A., Rouet, J.-F., & Perfetti, C.A. (1996). Using hypertext to study and reason about historical evidence. In J.-F. Rouet, J.J. Levonen, A.P. Dillon & R.J. Spiro (Eds.) *Hypertext and Cognition* (pp. 43-72). Mahwah, NJ: Lawrence Erlbaum Associates.

Chi, M.T.H., DeLeeuw, N., Chiu, M;, & LaVancher, C. (1994). Eliciting self-explanations improves understanding. *Cognitive Science,* 18, 439-477.

Deerwester, S., Dumais, S.T., Furnas, G.W., Landauer, T.K., & Harshman, R. (1990). Indexing by latent semantic analysis. *Journal of the American Society for Information Science*, 41, 391-407.

ETS (2004). *GRE for Educators: Interpreting Scores on the GRE Analytical Writing Measure.* Retrieved June 1, 2004, from http://www.gre.org/interpret.html

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

Foltz, P.W., Gilliam, S., & Kendall, S. (2000). Supporting Content-Based Feedback in On-Line Writing Evaluation with LSA *Interactive Learning Environments,* 8(2), 111-127.

Foltz, P. W., Laham, D. & Landauer, T. K. (1999). Automated Essay Scoring: Applications to Educational Technology. In Proceedings of *EdMedia '99.*

Graesser, A.C., Wiemer-Hastings, P., Wiemer-Hastings, K., Harter, D., Tutoring Research Group & Person, N. (2000). Using Latent Semantic Analysis to Evaluate the Contributions of Students in AutoTutor. *Interactive Learning Environments*, 8(2), 129-147.

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

Kintsch, E., Steinhart, D., Stahl, G., LSA Research Group, Matthews, C., &, Lamb, R. (2000). Developing Summarization Skills through the Use of LSA-Based Feedback. *Interactive Learning Environments*, 8(2), 87-109.

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.

Robertson, J., & Wiemer-Hastings, P. (2002). Feedback on children's stories via multiple interface agents. In *Proceedings of the 2002 Conference on Intelligent Tutoring Systems*, (pp. 923-932). Berlin: Springer-Verlag.

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.

Suthers, D., Connelly, J., Lesgold, A., Paolucci, M., Toth, E. E., Toth, J., & Weiner, A. (2001). Representational and advisory guidance for students learning scientific inquiry. In K. Forbus and P. Feltovich (Eds.) *Smart Machines in Education*. Menlo Park, CA: AAAI Press.

The College Board (2004). *The New SAT 2005: Writing section.* Retrieved June 1, 2004, from http://www.collegeboard.com/newsat/hs/writing.html

Voss, J. F. & Wiley, J. (2001) Developing understanding in history. In P. K. Smith & A.D. Pellegrini (Eds.) *Psychology of Education: Major Themes*. Routledge/Falmer: London.

Wiley, J. (2001) Supporting understanding through task and browser design. Proceedings of the Twenty-third annual Conference of the Cognitive Science Society. Hillsdale, NJ: Erlbaum.

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.

Wiemer-Hastings, P., & Graesser, A. (2000). Select-a-Kibitzer: A computer tool that gives meaningful feedback on student compositions. *Interactive Learning Environments,* 8*,* 149-169.