Behav Res (2012) 44:622–633
DOI 10.3758/s13428-012-0214-0

Assessing the use of multiple sources in student essays

Peter Hastings · Simon Hughes · Joseph P. Magliano · Susan R. Goldman · Kimberly Lawless

Published online: 1 June 2012

Abstract The present study explored different approaches for automatically scoring student essays that were written on the basis of multiple texts. Specifically, these approaches were developed to classify whether or not important elements of the texts were present in the essays. The first was a simple pattern-matching approach called "multi-word" that allowed for flexible matching of words and phrases in the sentences. The second technique was latent semantic analysis (LSA), which was used to compare student sentences to original source sentences using its high-dimensional vector-based representation. Finally, the third was a machine-learning technique, support vector machines, which learned a classification scheme from the corpus. The results of the study suggested that the LSA-based system was superior for detecting the presence of explicit content from the texts, but the multi-word pattern-matching approach was better for detecting inferences outside or across texts. These results suggest that the best approach for analyzing essays of this nature should draw upon multiple natural language processing approaches.

Keywords Multiple documents integration · Natural language processing · Reading · Writing

P. Hastings (✉) · S. Hughes
College of Computing and Digital Media, DePaul University,
243 South Wabash Avenue,
Chicago, IL 60604, USA
e-mail: peterh@cdm.depaul.edu

J. P. Magliano
Department of Psychology, Northern Illinois University,
DeKalb, IL, USA
e-mail: jmagliano@niu.edu

S. R. Goldman
Learning Sciences Research Institute, Departments of Psychology and Curriculum & Instruction, University of Illinois at Chicago,
1007 W. Harrison Street,
Chicago, IL 60607-7137, USA
e-mail: sgoldman@uic.edu

K. Lawless
Department of Educational Psychology,
University of Illinois at Chicago,
1040 W. Harrison M/C 147,
Chicago, IL 60607, USA
e-mail: klawless@uic.edu

Imagine a situation in which a student is asked to write a research paper on the causes of climate change and, in particular, to argue that the primary causes are based on human activities. Presumably, the student would need to identify and integrate information from multiple text sources to write such a paper. The cognitive representation resulting from these reading comprehension processes would likely reflect how information in each of the texts informs the student's position and the role of this information in the argument presented in the research paper (Rouet, 2006; Rouet & Britt, 2011). The representation would likely reflect both intra- and intertextual relationships. Many of these relationships would have to be inferred by the student, because the texts would have been written by different authors, at different times, and for different purposes.

Understandably, this type of reading and writing task is challenging for many students, in part because they have not had opportunities to learn how to read and write with multiple sources of information (Bråten, Strømsø, & Britt, 2009; Goldman, in press; Goldman et al., 2010; Lawless, Goldman, Gomez, Manning, & Braasch, 2011; Rouet & Britt, 2011; Wiley et al., 2009; Wiley & Voss 1999). The skills required to do so go well beyond those of simple comprehension. But success in modern society emphasizes the functional value of

reading for accomplishing personal, academic, and professional tasks (McNamara & Magliano, 2009; Organization for Economic Co-operation and Development, n.d.; Rouet, 2006; Snow 2002). In addition, the Internet has become a ubiquitous source of information, much of it unfiltered by traditional gatekeepers (e.g., teachers, librarians, publishers, and peer reviewers). The burden of selecting reliable and relevant information and determining how to connect information across multiple, often seemingly contradictory or unrelated, sources of information has become part of reading and writing proficiency. The recently developed U.S. Common Core Standards reflect these societal needs (www.corestandards.org/in-the-states). The standards delineate literacy skills of critical reasoning within and across multiple sources of information in literature, history, and science.

Efforts to provide opportunities for students to move beyond simple comprehension necessarily require assignments that involve open-ended (constructed) responses that may have multiple answers. These kinds of performances are time-consuming to evaluate and provide feedback on. At the same time, a growing body of research is validating the viability of computer-based assessments of student essays and other forms of constructed responses (Attali & Burstein, 2006; Britt, Wiemer-Hastings, Larson, & Perfetti, 2004; Burstein, Marcu, & Knight, 2003; Foltz, Gilliam, & Kendall, 2000; Franzke, Kintsch, Caccamise, Johnson, & Dooley, 2005; Landauer, Laham, & Foltz, 2003). For the most part, these efforts have focused on essays generated from single texts and employ computational algorithms to compare the semantic content of the students' responses to the presented text or assessment targets. (See Britt et al., 2004, for an exception.) These comparisons provide the basis for automatic classification of students' responses. The assessment targets can be semantic information that is indicative of cognitive processes (e.g., Magliano & Millis, 2003; Magliano, Millis, the RSAT Development Team, Levinstein, & Boonthum, 2011), specific expectations of student responses (e.g., Graesser et al., 2000), or a range of exemplar responses that reflect different levels of quality (Foltz et al., 2000).

A major distinction between different computational algorithms is whether they include any consideration of word order. "Bag-of-words" approaches, such as latent semantic analysis (LSA; Landauer & Dumais, 1997), do not consider word order, whereas pattern-matching approaches, such as the text classification systems developed by Zhang and colleagues (e.g., Zhang, Yoshida, & Tang, 2007), do. (See Graesser & McNamara, 2012, for an extensive review of approaches to analyzing constructed responses.) Regardless, the assessments are probabilistic rather than absolute and can be seen as general estimates of the quality and nature of the responses. However, new challenges arise when attempting to use computational approaches to evaluate students' responses that are intended to be based on multiple sources of information.

The two most significant challenges are semantic overlap among sources and cross-source inferences. Semantic overlap is a natural result of the fact that sources of information on the same topic are likely to involve many of the same concepts and words. Another typical characteristic of multiple-source situations is that the connections across sources are not explicit: The reader must infer them. These two characteristics of multiple-source reading situations introduce two complexities for computational algorithms: increased ambiguity in the "match" of a student response to a specific text/source, and the increased importance of how words and sentences are ordered and related to one another, especially across sources. The latter consideration increases the importance of relational terms (e.g., causals or logical connectors) in determining the quality of constructed responses. If one aim of analyses of student essays is to determine the degree to which a student has drawn on multiple sources in constructing the essay, and has done so appropriately, these two challenges must be tackled. The work reported in this article is an initial attempt to develop computational approaches to tackling these two challenges of multiple-source comprehension situations.

Specifically, in this article, we report on our efforts to use three types of computational approaches to analyze student essays that were generated as part of a project whose goal was the development of assessment tools for multiple-source comprehension (Goldman et al., 2011; Lawless et al., in press). In the context of the assessment tool development project, students read three texts that contributed complementary information on the inquiry topic and wrote an essay using the texts to address the inquiry question. The reading and writing tasks were conducted via a Web-based application, and data were collected on reading patterns and on the essays. Coding of the essays was done by human scorers with two purposes in mind: determining what information students included in their essays (*relevance*) and how they organized it (*integration*). Organization was evaluated against a template of how the source information related to a complete answer to the inquiry question. This template can be thought of as an "ideal" or "expert" map of the information in each of the sources and of the relationships across sources, and is referred to as an *integrated model*. Just as a mental representation of a text might serve as the basis for a response, the integrated model serves as a basis for constructing an essay that responds to the inquiry question. Using the integrated-model template, human coders determined which elements and relationships were present in the essays.

There is variability in how students respond to this task, in terms of how they use the texts to construct their essays (Goldman et al., in press): Some students simply produce content from one text; others provide information from multiple texts, but do so without constructing an integrated

argument; finally, some students engage in the task as intended and write an integrated argument that combines content from the texts in a novel and appropriate manner. These different approaches can be discerned through time-consuming qualitative analyses, and are therefore unwieldy for teacher use. However, the development of computer-based automated essay analysis could form the foundation of a classroom-friendly system that would provide this kind of information. With that goal in mind, we explored the viability of three computational approaches to coding the content of essays: pattern matching, latent semantic analysis, and support vector machines (SVMs; Hastie, Tibshirani, & Friedman, 2009; Joachims, 2002).

Computational approaches

Pattern matching is a variant of string matching. It involves identifying patterns of key words that should be relatively diagnostic of the extent to which the different elements of the integrated model are reflected in the essays. This approach generally involves identifying a family of potential patterns, which are derived from a development sample of essays. This step is critical, because it helps ensure that the patterns reflect the language actually used by the students. As will be discussed below, we developed a variant of the *multi-word* approach (Zhang et al., 2007) that automatically identifies simple patterns—sequences of consecutive words—that are associated with different integrated-model nodes. This approach has been successful in a variety of applications, including document classification and the creation of indices for information retrieval systems (e.g., Chen, Yeh, & Chau, 2006; Papka & Allan 1998; Weiss, Indurkhya, Zhang, & Damerau, 2005; Zhang, Yoshida, & Tang, 2007, 2008, 2011; Zhang, Yoshida, Tang, & Ho, 2009). The primary merit of this approach is that it should be sensitive to the language used by the students and the order of words used in the essays. There is no guarantee, however, that the patterns developed from one sample of students and/or topics will transfer to a new sample.

The other two approaches are so-called *bag-of-words* approaches, which completely ignore word order and treat words as the distinguishing features of their respective texts. The first uses LSA (Landauer & Dumais, 1997) to assess whether student essays reflect the semantic information in the source texts. LSA has previously been used in a multiple-document context to identify the overall source document invoked by student sentences at the college (Britt et al., 2004; Foltz, Britt, & Perfetti, 1996) and middle school (Hastings, Hughes, Magliano, Goldman, & Lawless, 2011) levels. We adapted an approach used by Magliano and colleagues (Magliano & Millis, 2003; Magliano et al., 2011), which we call *mapped LSA*. Specifically, LSA was used to compare each of the sentences in the student essays to the sentences of the original source texts. LSA yields a cosine that functionally varies between 0 and 1 and reflects the proximity in the semantic space between the student text and the source text. The LSA cosines between the sentences in the text set and the sentences that comprise the student essays are used to determine how students used the information in the text to construct their essays.

The third approach involves machine-learning algorithms called SVMs (Joachims, 2002; Hastie et al., 2009; Medlock, 2008). SVMs are one of the most widely used machine-learning techniques in use today for a wide range of tasks (Hastie et al., 2009). For example, Medlock used SVMs to perform four natural language processing tasks: topic classification, content-based spam filtering, anonymization, and hedge classification. SVMs use annotated examples to induce a classification based on the features in the examples. In our approach, which we label *SVM multiclass* herein, the training examples are the sentences from the student essays, the features are the words in the sentences, and the classes to be learned are the integrated model codes for the inquiry task assigned by the human raters. Our SVM approach is similar to mapped LSA, in that it filters out "stop words" (generally function words that carry little discriminative semantic content), and it weights the remaining words in the documents to reduce the effects of words that occur widely across documents and highlight those that are more discriminating. Also like LSA, SVMs treat the data as points in a high-dimensional space. SVMs do not use singular value decomposition, though. Instead, they identify hyperplanes that create the largest separations between the different classes of data.

These three systems have different potential strengths and weaknesses in the context of assessing essays (Magliano & Graesser, 2012). The multi-word approach is particularly useful when particular words or phrases are used to discriminate the different types of semantic content associated in the texts. Although LSA ignores word order, it does have the advantage of being trained on a large corpus of texts, so it should be able to identify semantic content without requiring the appearance of particular words. The SVM multiclass approach likewise ignores word order, but has the advantage of learning the classification from the actual texts that the students have produced. It should have an advantage, therefore, if students use particular combinations of words that differ from the patterns found in a larger corpus. Specific hypotheses regarding the outcomes of employing these three approaches will be best addressed, however, in the context of an explanation of the essays that were used in this study.

Data corpus and hypotheses

The essays that constitute the data corpus for the present study were the result of an inquiry question that students

were to answer on the basis of multiple text sources. The inquiry question was "In 1830 there were 100 people living in Chicago. By 1930, there were three million. Why did so many people move to Chicago?" This inquiry question focuses on migratory patterns, which according to historians involve "push," "pull," and "enabling" factors. That is, there are reasons why people leave their current location, reasons why certain places are attractive to relocate to, and circumstances that facilitate people getting from the current to the new locations. Because we were interested in how students used multiple sources to address inquiry questions, we constructed the text set so that a complete answer required students to use information from three different texts. One text described only "pull" factors (e.g., jobs in Chicago), another only "push" factors (e.g., poor farming conditions in Europe), and the third the "enabling" factors (e.g., transportation infrastructure for getting to Chicago). Furthermore, in anticipation of attempting to develop automated, computer-based scoring of the essays, we intentionally designed the texts to have relatively little overlap of specific words and semantic content. Some overlap was unavoidable, however, given the shared intersection of the text subjects. Furthermore, the presence of some common words could serve as a trigger for integrating inferences by the students.

To quantify the lexical overlap between the texts, we used the percent vocabulary overlap metric (PVO; Medlock, 2008). For two documents D_1 and D_2, in which $\mathbf{T}_1$ is the set of terms in D_1 excluding "stop words" and $\mathbf{T}_2$ is the set of terms in D_2, PVO is defined as**[COMP: Set built fractions in most eqs., as described at top. Cf. pdf p. 13 for this one.]**

$$\mathrm{PVO}(D_1, D_2) = \{|\mathbf{T}_1 \cap \mathbf{T}_2| // |\mathbf{T}_1 \cup \mathbf{T}_2|\}.$$

For comparison, we measured the PVO between our "enabling"-factors text and a general text about the same topic: the section from the Wikipedia Chicago History page (http://en.wikipedia.org/wiki/Chicago_history, accessed October 30, 2011) entitled "Emergence as a Transportation Hub." The PVO was 15.0 %. The PVO between this section of the Chicago history Wikipedia page and the section of the Wikipedia page on the history of New York City from the same time period (http://en.wikipedia.org/wiki/New_York_City_History, accessed March 25, 2012) was 6.5 %. The PVO values between the source texts used in this study fell squarely within this range: Between the "push"- and "pull"-factors texts, the PVO was 7.4 %; between the "push" and "enabling" texts, it was 7.9 %; and between the "pull" and "enabling" texts, it was 10.8 %.

The integrated model (template) for human coding of the essays was driven by a representation of a causal model that addressed the inquiry question, as shown in Fig. 1. The integrated model reflects a "complete" map of the text content, although we did not expect all of this information to be included in the essay. The representation places each idea in the text in a relationship to the inquiry question—a claim, evidence for a claim, or a detail or elaboration about the evidence. The questions of interest were what students included in their essays (i.e., what "level" of information) and how they organized it, including whether or not they connected "push," "pull," or "enabling" factors across texts.

The integrated model shown in Fig. 1 reflects the "pull," "push," and "enabling" conditions as three main claims that address the inquiry question about why people moved to Chicago between 1830 and 1930. Each text in the set can be used to support a claim that could be made about the "pull," the "push," or the "enabling" factor. The substance of the claim, the evidence, and the details underlying that evidence reflect three hierarchically organized elements under each claim. In Fig. 1, the "pull" text conveys the claim (CL1) that there were jobs in Chicago and presents two major lines of evidence for this claim (EV1A and EV1B), along with details that elaborate on the evidence. Similarly, the "push" text conveys the claim that people were looking for a better life (CL2) and the evidence for why this was so, along with details elaborating on the evidence. Finally, the third claim (CL3) was about the transportation system that made Chicago accessible. The expressions of these claims, the evidence, and the details are generalizations and paraphrases of information presented in the text. Thus, within each claim hierarchy, some information was explicit in the text (text-based elements) for each of the idea nodes depicted in Fig. 1. Usually, multiple sentences appeared in a text for each node in Fig. 1. The specific way in which the idea might be expressed in an essay could vary from an exact copy of one or of multiple sentences to a phrase that summarized across multiple sentences. Causal relations were also explicit *within* a text. These are indicated by RC codes (e.g., RC1A, RC1B, etc.) and solid lines showing the nodes that these lines connect. Cross- and extratext relationships and connections were not explicit in the texts. Figure 1 depicts these inferred relationships using codes starting with "I" (e.g., IR12, meaning an inferred relationship between Claims 1 and 2) and dotted lines showing the nodes that they connect.

The different elements of the representation shown in Fig. 1 pose different challenges for automated scoring systems. Given that the text-based elements correspond to specific sentences in the texts, there is relatively rich semantic context to make comparisons between the essays and semantic benchmarks reflecting the model elements. The challenge is in recognizing summaries or transformations that depart from the exact words that were presented in one of the sources. On the other hand, by design, the three text sources contained semantically distinct words and ideas, making it somewhat easier to determine which text a particular essay sentence reflected. There were, however, surface text matches across texts that complicated the text source

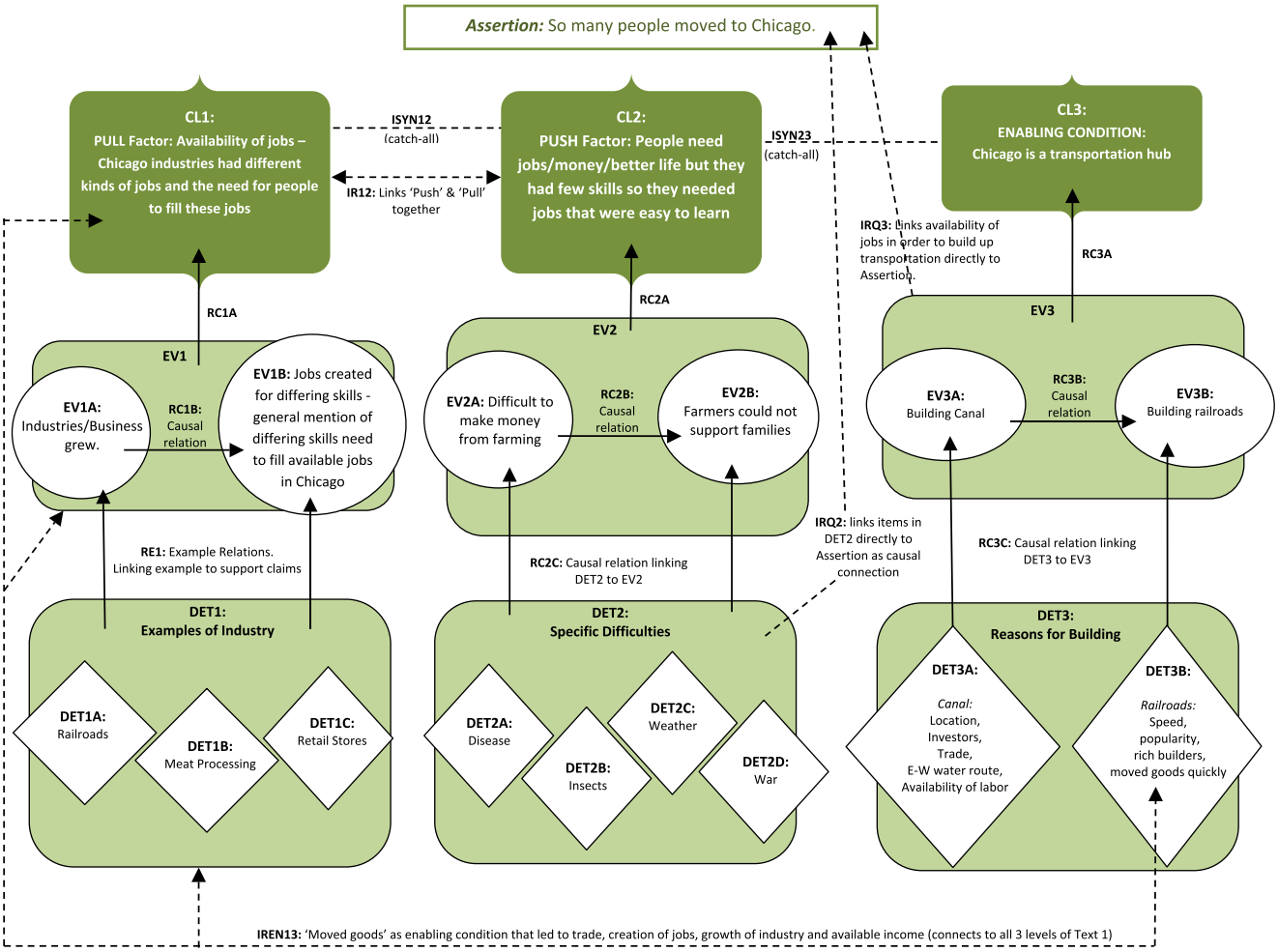

Fig. 1 The integrated model for the Chicago text set

identification issue. For example, both the "pull" text (EV1) and the "enabling" text (EV3) describe railroads as a factor, but they do so in very different ways. The "pull" text describes how railroads supported industry in Chicago, whereas the "enabling" texts described the development of the railroad infrastructure. Determining whether students were discussing railroads in the context of the "pull" or of the "enabling" text likely would require taking word order and context into consideration. For example, the student sentence *One of the businesses was to build the railroad cars* was labeled with the DET1A code, because it focuses on jobs in the railroad industry. The sentence *They also build railroads, this made it a faster way to travel* shares the words *build* and *railroad*, but was labeled with EV3B because its emphasis is on travel.

Much more challenging was automated scoring of the relational elements. For relations that are explicit in the text, there are a variety of ways that those connections might be expressed in an essay, including simple sequential order and the use of cue words such as *so*, *as a result*, or *because.* These connectors, as well as the word or phrase order, are critical to the meaning of the sentence. Inferred relations are particularly challenging because they do not directly correspond to explicit content (Magliano et al., 2011; Millis, Magliano, Todaro, & McNamara, 2007). There should be a relatively greater degree of variation in student responses for these inferences than for the text-based elements, which would make it challenging to develop semantic benchmarks indicative of the inferences (Millis et al., 2007). It may be the case that detecting the presence of function words that are indicative of linking relationships (e.g., causal and logical connectives such as *because*, *therefore*, and *furthermore*) and of important content words would present the optimal solution for automatically detecting linking relationships.

There are four hypotheses regarding the relative performance of the three computational approaches to automated classification. A *semantic-precision* hypothesis assumes that the best solution to the problem of automatically determining the coverage of the integrated model in the essay will be sensitive to word order and context. That is, given shared semantic overlap across the texts, it is critical to determine what words co-occur and in what order. According to this

hypothesis, the measure of accuracy will be higher for the pattern-matching approach than for either of the bag-of-words approaches.

A *general-semantic-overlap* hypothesis assumes that one of the bag-of-words approaches will provide a more optimal solution. The reason is that the variability of student responses may be sufficiently high that it is a challenge to identify patterns that are diagnostic and generalizable. The mapped-LSA approach arguably is the most generalizable, because it involves a semantic comparison between the texts sentences that all of the students have read and their essays. LSA should be sensitive to the gradients in semantic overlap that can occur in natural language. According to this hypothesis, the measure of accuracy will be higher for one of the bag-of-words approaches than for the pattern-matching approach.

A *specific-semantic-overlap* hypothesis holds that, given the semantic overlap between the source texts, students are likely to produce sentences that include many of the same words. Thus, a general method like mapped LSA will be unable to distinguish many of the concepts, but an approach that learns to classify sentences from an annotated corpus will more successfully identify the concepts in the greatest number of student sentences. Like the general-semantic-overlap hypothesis, it assumes that the ordering of words within sentences is less important than the combinations of words that occur in the sentences.

Finally, a *functional-semantic-overlap* hypothesis assumes that a hybrid solution will provide the best classification of the essays. Specifically, different elements of the integrated model vary in their semantic richness. Each of the nodes of the integrated model maps onto a specific set of sentences in the texts. For that reason, specific content words and related words should be indicative of students' producing content from these sentences in their essays. Systems like LSA have been shown to fairly accurately indicate the use of content from the discourse context (e.g., Magliano & Millis, 2003; Wolfe & Goldman, 2003). On the other hand, linking relationships are indicative of very specific sentences in the text and of the use of semantically depleted causal connectives (e.g., *because* and *therefore*). Mapped LSA and our multiclass SVM approach both ignore function words. Therefore, it may be that multi-word pattern matching would be better able to detect the extent to which students are producing linking relationships.

Additional information about the essay corpus

The corpus of essays to which the automated approaches were applied consisted of 460 essays that had been collected from students in Grades 5, 6, 7, and 8. These students attended two urban public schools in the Midwest.

Two-thirds of the students participated individually in a computer-based task in which they were asked to use the three passages to answer the inquiry question *Why did so many people move to Chicago between the years 1830 and 1930?* During the reading phase of the task, the three texts were available for reading in a "reading screen," but the texts could be open only one at a time. As described previously, each text provided different information relevant to the inquiry question: "pull" factors—the development of industry; "push" factors—the search for a better life; or "enabling" conditions—transportation systems. The students spent approximately 15–20 min reading each of the three Chicago texts. In Phase 2 (writing), the students were asked to write an essay to address the inquiry question using the information in the three texts, each of which could be opened (one at a time) in a window next to the writing window. The students could not copy and paste from the reading to the writing window; they had to type what they wanted in the writing window. The students spent approximately 15–45 min on the writing portion of this task.

The other third of the students participated in a paper-and-pencil version of the task and hand-wrote their responses.

Analyses of the essays

Human coding The goal of the data analysis was to examine the students' essays in order to see how they were making sense of the texts and how the texts were being processed with respect to the posed inquiry question. However, the present study focused on only one aspect of the coding system—in particular, the extent to which the student essays reflected the different aspects of the integrated model. The essays were initially spell-checked and parsed into sentences, which were the unit of analyses for the human raters and the automated systems. There were two phases in the coding scheme. The first phase involved human coders identifying whether the content of the text sentences was represented in the student essays. Each essay sentence was coded with respect to each sentence in each of the three source texts by two coders with an obtained interrater reliability of 85 %.

The second phase involved mapping the sentences to the elements of the integrated model. The difference between the two phases of coding was that in the second phase, sentences in the essays might be determined to "match" a

node in the model but not necessarily a specific sentence in any one of the texts. This tended to occur when the students summarized paragraphs rather than including details from specific presented sentences. Although this is not clearly specified in Fig. 1, the integrated model contained 37 components that corresponded to the nodes and links. Each element was linked to a set of text sentences, and if an essay sentence was determined to reflect a text sentence, it was also determined to reflect its corresponding element of the integrated model. It is important to note that a sentence could reflect multiple nodes or links in the integrated model, especially if it contained multiple clauses.

Computer-based analyses Three metrics were used for evaluating the computer-based approaches. For a given document class, *recall* is defined as the proportion of documents (sentences, in this case) belonging to that class in the entire data set that were correctly assigned to that class in this approach. In other words, if *truePositives* is the number of human-coded sentences for a class which were also coded with that class by the computer-based approach, and if *falseNegatives* is the number of human-coded sentences which were not given that code by the computer-based method, then**[COMP: Set built fractions. Cf. pdf pp. 21–22.]**

$$recall = \{truePositives//truePositives + falseNegatives\}.$$

Precision is the proportion of documents assigned to that class by the approach that actually belonged to that class—or, if *falsePositives* is the number of sentences assigned to a class by the computer method that were not given that code by the humans, then

$$precision = \{truePositives//truePositives + falsePositives\}.$$

Typically, as recall increases, precision decreases, and vice versa. Therefore, to capture both, a combined measure, the F_1 score, is computed as follows (van Rijsbergen, 1979):

$$F_1 = \{2 \cdot recall \cdot precision//recall + precision\}.$$

If the recall and precision values are similar, the F_1 score will approximate an average of the two. If they differ, the F_1 score will be less than the average. As there is normally a trade-off between recall and precision, different F scores can be chosen that privilege one over the other. The choice of which to use may depend on the application. For example, assume that you want to give feedback to students about sentences that are classified with a particular code, and you want to be fairly certain that the example sentence actually falls in this class, but you do not mind if you miss some examples. Then an F score can be chosen that privileges precision over recall, like the $F_{.5}$ measure. The general measure is F_β, with

$$F_\beta = \left\{\left(1 + \beta^2\right) \cdot precision \cdot recall//\left(\beta^2 \cdot precision\right) + recall\right\}.$$

Tenfold cross-validation (e.g., Mitchell, 1997) was used with both the multi-word pattern-matching and the SVM multiclass approaches. This involves segmenting the data into separate training and test sets to verify that the algorithms were not overfitting the data and that the classifier can generalize well to unseen data. The algorithms are then trained on the training data, and the test data set—previously unseen by the algorithms—is used to evaluate their performance.

The multi-word pattern-matcher classifies student sentences as belonging to a particular element in the integrated model in Fig. 1. It extracts reoccurring phrases consisting of one or more consecutive words from the set of sentences assigned to a particular node, and then builds up a pattern set using those phrases, starting with very specific patterns that only match a few sentences and gradually generalizing the pattern set by adding more phrases as alternatives.

Iterating through all of the sentences in the training data set for a particular integrated-model node, the multi-word algorithm initially creates all possible sequences of consecutive words with lengths between one word and the length of the sentence. All sequences occurring only once are then removed. Next, the algorithm matches each sentence in the training data set against each multi-word sequence and selects the sequence that maximizes the F_p score below:**[-COMP: Set square-root radical over *recall*. Cf. pdf p. 23.]**

$$F_{\mathrm{p}} = precision \cdot \sqrt{recall}.$$

All sentences matching the selected multi-word pattern are then removed, including those that belong to a different integrated-model node. Removing matched sentences, even those that would be misclassified, ensures that future patterns are chosen that perform well on the remaining, unclassified sentences. This process is then repeated, each time picking the multi-word pattern that maximizes the F_{p} score on the currently unmatched sentences. After each iteration, the selected multi-word pattern is disjunctively combined with the previously selected patterns to form a larger composite pattern. This pattern matches any text containing one or more of the previously selected multi-word patterns.

The F_{p} score emphasizes accuracy over generality by placing more emphasis on precision than on recall. In the construction of the final composite pattern, the algorithm starts with a multi-word phrase that has a very high precision but a low recall, because it only matches a small number of sentences. On each iteration, additional multi-word patterns are selected—each with a high precision—and

are combined with the previous patterns, increasing the recall while retaining a high precision. This allows the algorithm to generate a composite pattern with a high F_1 score. After each iteration, the performance of the composite pattern on the test data set is measured using the F_1 score. The algorithm terminates following ten consecutive cycles without any improvement in the F_1 score on the test data set. This ensures that the algorithm does not overfit the training data and that the learned pattern performs well on unseen data.

For example, the human coders marked 373 student sentences with the CL1 code. On the first iteration, the multi-word approach chose the phrase *to fill*, which correctly matched 108 sentences, with 15 false positives and 265 false negatives, producing a precision of .88, a recall of .29, and an F_1 score of .44. Next, the phrase *center of industry* was chosen, which correctly matched 20 sentences, with 1 false positive and 245 false negatives, for a precision of .95, a recall of .05, and a combined F_1 score of .50. Finally, the phrase *of jobs* was selected, producing 36 true positives, 47 false positives, and 209 false negatives, for a total combined F_1 score of .55. At that point, no other patterns were added, because they would have lowered the overall F_1 score.

For the mapped-LSA approach, we separated the student essays into sentences and compared each student sentence to each sentence from the original source texts by using the lsa.colorado.edu website with the "General_Reading_up_to_1st_year_college" semantic space, 300 factors, and document-to-document comparisons. We used an empirically derived threshold to determine whether a student sentence matched the source sentence. To calibrate the mapped-LSA approach with the human raters, we calculated the correspondence between the human-assigned sentence codes and the LSA-assigned sentence codes using thresholds of .4, .5, .6, and .7. The best performance was achieved with the threshold set at .7, with recall=.43, precision=.66, and F_1=.52.

As mentioned above, the integrated model includes, for every intratext component, a set of one or more sentences from the original source texts that exemplify that component. We used the LSA-derived sentence classification along with this mapping to compute which integrated-model components were covered by the student essays. For example, the second sentence of the "pull" text was *The jobs were mainly in three kinds of businesses: the railroad industry, meat processing and retail stores that sold things to people*. In the integrated model, this sentence was associated with components EV1A (Industries/Business Grew) and DET1 (Examples of Industry). If a student sentence achieved a cosine over .7 with this sentence, it would be labeled with these two codes from the integrated model.

We implemented the SVM multiclass approach using svm_multiclass from Joachims's SVM^{light} package, available from http://download.joachims.org/svm_multiclass/current/svm_multiclass.tar.gz. A training example was created from each student sentence. For a sentence, the training data consisted of a set of weights, one for each word (except stop words) that occurred in the sentence. The weights (*tfidf*; see, e.g., Medlock, 2008) emphasized words that were distinctive to this document and reduced the effect of words that occurred across all documents. Although many machine-learning classification tasks are binary—that is, they determine whether or not an example corresponds to only one class at a time—SVM multiclass simultaneously learns a classification for all of the classes that it is given—the 37 nodes of the integrated model, in this case. The SVM was trained using tenfold cross-validation, and the resulting model was used to classify the integrated-model components for each student sentence.

Results

Figure 2 shows the frequencies of the human ratings for the different types and levels of integrated-model components. The first three columns reflect the text-based categories, from the top-level claims (CL), to the evidence (EV), to the details (DET). The other two columns reflect the linking relations: the intratext categories (RC) and the extratext categories (IR) that connect to assertions or between texts. Examples of the IR category are clearly less frequent than those of the other categories, reflecting the relative scarcity of student inferences outside a single text. This can pose a challenge for automatic methods of learning classifications, because there are fewer examples to learn from.

Figure 3 gives the performance of the three classification methods on the set of student texts, as indicated by their aggregated F_1 scores. The aggregated results are sensitive to

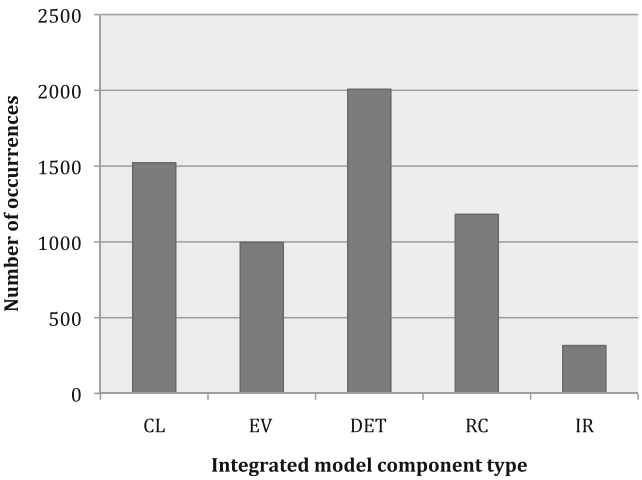

Fig. 2 Frequency counts for the production of the model components in the student essays

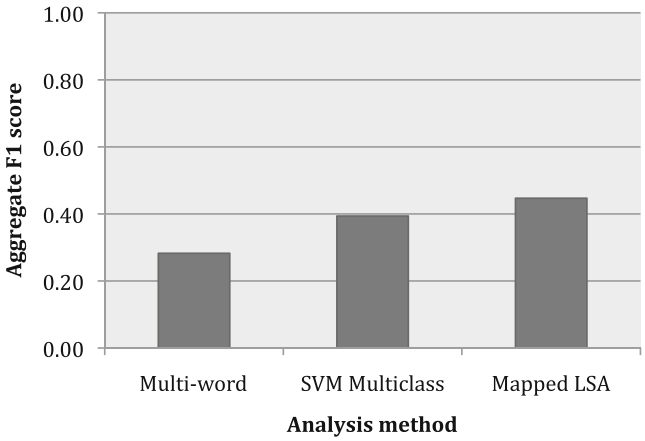

Fig. 3 Aggregate F_1 scores for the different natural language processing approaches

the frequencies of occurrence of the categories and give a good overall picture of how well the techniques did in identifying the categories of any student sentence. The mapped-LSA method performed the best overall (F_1=.45), followed by SVM multiclass (F_1=.39), followed by multi-word pattern matching (F_1=.28),

The performance of the different techniques in the different categories of integrated-model components is shown in Fig. 4. Table 1 shows the recall and precision scores along with the F_1 scores. As can be seen in Table 1, mapped LSA had higher precision scores than did the other approaches for all categories expect IR. Upon careful inspection of the recall scores, no one approach had a clear advantage, though the multi-word approach had the highest scores for the CL, RC, and IR categories. For the three text-based categories—CL, EV, and DET—mapped LSA performed as well as or better than the other techniques. It also did well on the RC category. It should be noted that although RC is a category of linking relations, some of the components are also text-based; that is, specific sentences in the original source texts

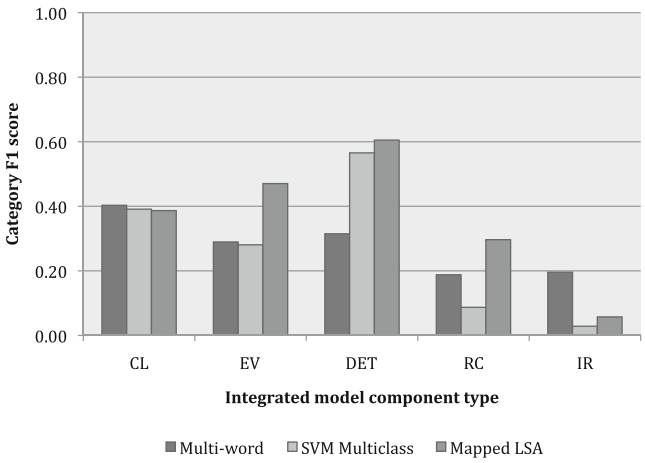

Fig. 4 F_1 scores for the different natural language processing approaches, as a function of model component

Table 1 Performance (recall, precision, and F_1) by integrated-model component (IM comp) types

IM comp	Multi-word			SVM Multiclass			Mapped LSA		
	Recall	Prec	F_1	Recall	Prec	F_1	Recall	Prec	F_1
CL	.48	.35	.40	.39	.39	.39	.31	.51	.39
EV	.45	.21	.29	.25	.32	.28	.55	.41	.47
DET	.50	.23	.31	.68	.49	.57	.59	.62	.61
RC	.42	.12	.19	.06	.20	.09	.26	.35	.30
IR	.31	.14	.20	.02	.06	.03	.04	.09	.06
Aggr	.46	.20	.28	.38	.41	.39	.41	.49	.45

do describe those relations. The IR category is inherently non-text-based, because it indicates inferences across the texts or to the overall inquiry question. Mapped LSA cannot perform well for these types of components, because there are very few sentences from the source texts to compare with the student sentences.

Discussion

The results of this study are consistent with the functional-semantic-overlap hypothesis. For conceptual sentences that were strongly text-based, mapped LSA proved the most accurate natural language processing tool. For relational sentences, however, the multi-word pattern matching approach was superior. To our knowledge, this is the one of the few studies that have directly compared the accuracy of these three natural language processing approaches for analyzing the content of essays, although some systems have used LSA and pattern matching (Britt et al., 2004). What is the advantage of mapped LSA over the other two approaches? LSA represents word knowledge in a high-dimensional semantic space, which can represent the direct and indirect semantic relationships between words (Landauer & Dumais, 1997). Given that essays such as these are written extemporaneously, they may be ill-formed, and students can use a large range of semantic content to cover the ideas represented in the texts (e.g., Graesser & McNamara, 2012). As such, this is the kind of situation in which mapped LSA has a clear advantage over multi-word pattern matching. Specifically, there is no guarantee that the patterns derived from one sample will generalize such that they can be used to analyze a new sample of essays. With the present application of LSA, the benchmarks were compared to the texts that all of the students had read. Thus, the issue of generalization was not of concern.

The specific-semantic-overlap hypothesis predicted that the SVM multiclass approach would have an advantage over mapped LSA. This hypothesis was not supported by our

results. One possible explanation is the relative frequencies of the types of sentences in the corpus. As is shown in Fig. 4, the performance of SVM multiclass on the CL and DET codes was close to that of mapped LSA. Figure 2 shows that these are the most frequent codes in the student essays. On the less frequent codes, the performance of SVM multiclass is compromised because it is attempting to achieve the highest overall classification of the entire set of sentences. To accommodate the more frequent codes, it may end up misclassifying many of the less frequent ones. Another possible explanation is student paraphrasing of the source text content. Although the overall semantic content of the student sentences may not vary much from the source texts, minor variations in wording may prove more of a challenge for the SVM multiclass approach, which was trained on our corpus of student essays, than it would be for LSA, which was (previously) trained on a much more extensive set of texts (the *tasaALL* corpus, with over 37,000 documents and over 92,000 terms, available at http://lsa.colorado.edu).

This brings up an additional disadvantage of the SVM multiclass approach. The LSA training was done once, and it can then be used for many different applications. To use an approach like SVMs for this type of task, human coders must annotate the sentences in a relatively large set of texts to use for training. This requires significant effort and significant expertise, and is thus a limiting factor on the generalizability of the SVM multiclass approach.

The results shown in Table 1 highlight other differences between the approaches. SVM multiclass was trained on the entire set of sentences and aimed to get the best overall performance on the set. Thus, it had relatively balanced recall and precision on the aggregate. For some of the categories, there was less balance, particularly on the DET category, which was biased toward recall, and the RC category, which was biased toward precision. This may mean that the representation learned for DET "casts a wide net" at the cost of picking up examples from other codes. Alternatively, it could mean that sentences from this category have a relatively high word overlap with sentences of other categories, so that it is more difficult to separate them.

There is no single standard of acceptability for *F* scores on text classification tasks; the acceptability level depends greatly on the task, and the levels that can be achieved depend greatly on the lengths and distinctiveness of the texts. In some classification tasks, *F* scores in the .90s have been reported. For sentences in student essays in a multiple-source situation, however, the classification task is more difficult. Although correlations cannot be directly compared with *F* scores, the results of Foltz et al. (1996) can give an indication of the difficulty of the task. For the task of identifying with LSA which source *document* (out of 21) was closest to a student sentence, they found correlations between .12 and .63 with human raters. Tellingly, the correlations between the four human raters ranged from .37 to .77. In our study, we had only three source texts, but we were matching with individual sentences (74 total) to identify 37 integrated-model codes. Although the level of performance was not as high as we might have hoped, it may well be completely adequate for the educational task that is our eventual goal: automatically calculating the levels of conceptual coverage and integration in student essays in order to provide feedback to both teachers and students.

We raised two significant challenges for the computer-based assessment of essays based on multiple texts—namely, the degree of semantic overlap between texts and the detection of inferences. However, the results suggest that a hybrid solution may be the best approach for addressing these challenges. Specifically, the present study suggested that mapped LSA is particularly useful when trying to detect the presence of semantically rich content, despite the fact that there was some semantic overlap across texts. On the other hand, multi-word pattern matching may be useful for detecting important relational inferences that are not semantically rich. These inferences are typically represented by a causal or logical connective (e.g., *because*, *therefore*, or *however*) and a few key semantic items from the text. As such, the successful detection of these inferences is improved by looking for patterns of the semantically depleted connectives (which LSA typically ignores) along with the semantically rich words that LSA is good at detecting.

Although the idea of hybrid solutions is not new, and has been used in other systems that analyze student-constructed responses, such as AutoTutor (Graesser et al., 2000; Graesser et al., 2004), iSTART (McNamara et al., 2004), and SAIF (Britt et al., 2004), the present study has made a novel contribution. Specifically, the extant hybrid systems typically compute measures of semantic overlap using LSA and string matching, and they use statistical modeling to weight the indices derived by these techniques in order to classify the responses (e.g., Millis et al., 2007). As discussed above, the present study illustrates that these systems may be differentially useful for detecting different types of information. In other words, the best classification technique may depend on the type of sentence being classified.

Our ultimate goal is to develop an assessment system that could be used by teachers and researchers. The research presented here constitutes an initial step in that direction, but there is obviously considerable work left to accomplish. First, because of the importance of pattern matching for detecting inferences, an automated hybrid system that uses mapped LSA plus multi-word pattern matching needs to be developed and tested. This development would include algorithms that would detect the overall class of a sentence (e.g., is the presence of a connective sufficient to identify the sentence as a linking relationship?) and weight the relative importance of indices that could be derived from LSA and

pattern matching appropriately, given the nature of the assessment target (e.g., Millis et al., 2007). Specifically, LSA would be weighted more heavily when the assessment targets contain semantically rich information, whereas pattern matching would be weighted more heavily when the targets contain important function words, such as causal connectives. Alternatively, connectives could be used to segment sentences, and then LSA could be applied to the sentence segments. More research needs to be done to develop and test these algorithms.

The text set and the integrated model played key roles in the essay coding, both by the human coders and by the automated approaches. This presents an interesting challenge for developing a classroom-friendly system. One alternative is for researchers to develop text sets and integrated models for a range of topics, constituting a library from which teachers could select topics on which to assess multiple-source reading and writing. Alternatively, a system might be developed that scaffolds teachers through a process of text-set and integrated-model development, so that they themselves could generate the targets for an automated multiple-source assessment system.

Nonetheless, the research presented in this article illustrates the viability of developing automated systems for evaluating essays based on multiple texts (see also Britt et al., 2004). Given the adoption of the Common Core Standards, which emphasize that students need to be able to comprehend and use multiple documents, it stands to reason that this practice will increase. Developing computational systems that assist teachers and students in evaluating the quality of such essays could provide a vital tool for supporting this effort.

Author note The authors gratefully acknowledge the contributions of Rebecca Penzik and Kristen Rutkowski in the conduct of the experiment and analyses of the assessment data reported herein. The assessment project described in this article was supported by the Institute for Education Sciences, U.S. Department of Education, via Grants R305G050091 and R305F100007 to the University of Illinois at Chicago. The opinions expressed are those of the authors and do not represent the views of the Institute or the U.S. Department of Education. We also acknowledge the support of the Center for the Interdisciplinary Study of Language and Literacy at Northern Illinois University.

References

Attali, Y., & Burstein, J. (2006). Automated essay scoring with e-rater R V.2. Journal of Technology, Learning and Assessment, 4, 1–30.

Bråten, I., Strømsø, H. I., & Britt, M. A. (2009). Trust matters: Examining the role of source evaluation in students' construction of meaning within and across multiple texts. *Reading Research Quarterly*, *44*, 6–28.

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.

Burstein, J., Marcu, D., & Knight, K. (2003). Finding the WRITE stuff: Automatic identification of discourse structure in student essays. *IEEE Intelligent Systems*, *18*, 32–39.

Chen, J., Yeh, C.-H., & Chau, R. (2006). Identifying multi-word terms by text-segments. In *Proceedings of the Seventh International Conference on Web-Age Information Management Workshops* (pp. 10–19). Piscataway, NJ: IEEE Press. doi:10.1109/WAIMW.2006.16

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

Foltz, P. W., Gilliam, S., & Kendall, S. (2000). Supporting content-based feedback in on-line writing evaluation with LSA. *Interactive Learning Environments*, *8*, 111–127.

Franzke, M., Kintsch, E., Caccamise, D., Johnson, N., & Dooley, S. (2005). Summary street: Computer support for comprehension and writing. Journal of Educational Computing Research, 33, 53–80.

Goldman, S. R. (in press). Reading and the Web: Broadening the need for complex comprehension. To appear in R. J. Spiro, M. DeSchryver, M. S. Hagerman, P. Morsink, & P. Thompson (Eds.), *Reading at a crossroads? Disjunctures and continuities in current conceptions and practices*. New York, NY: Routledge.

Goldman, S. R., Lawless, K. A., Gomez, K. W., Braasch, J. L. G., MacLeod, S., & Manning, F. (2010). Literacy in the digital world: Comprehending and learning from multiple sources. In M. C. McKeown & L. Kucan (Eds.), Bringing reading research to life (pp. 257–284). New York, NY: Guilford Press.

Goldman, S. R., Lawless, K. A., Pellegrino, J. W., Braasch, J. L. G., Manning, F. H., & Gomez, K. (in press). A technology for assessing multiple source comprehension: An essential skill of the 21st century. To appear in M. Mayrath, J. Clarke-Midura, & D. H. Robinson (Eds.), *Technology-based assessments for 21st century skills: Theoretical and practical implications from modern research*. Charlotte, NC: Information Age Publishing.

Goldman, S. R., Ozuru, Y., Braasch, J., Manning, F., Lawless, K., Gomez, K., & Slanovits, M. (2011). Literacies for learning: A multiple source comprehension illustration. In N. L. Stein & S. W. Raudenbush (Eds.), Developmental science goes to school: Implications for policy and practice (pp. 30–44). New York, NY: Routledge.

Graesser, A. C., Lu, S., Jackson, G. T., Mitchell, H. H., Ventura, M., Olney, A., & Louwerse, M. M. (2004). AutoTutor: A tutor with dialogue in natural language. Behavior Research Methods, Instruments, and Computers, 36, 180–192. doi:10.3758/BF03195563

Graesser, A. C., & McNamara, D. S. (2012). Automated analysis of essays and open-ended verbal responses. In H. Cooper et al. (Eds.), *APA handbook of research methods in psychology*. Washington, DC: American Psychological Association.

Graesser, A. C., Wiemer-Hastings, P., Wiemer-Hastings, K., Harter, D., Person, N., & the Tutoring Research Group. (2000). Using latent semantic analysis to evaluate the contributions of students in AutoTutor. Interactive Learning Environments, 8, 129–147.

Hastie, T., Tibshirani, R., & Friedman, J. (2009). The elements of statistical learning: Data mining, inference, and prediction (2nd ed.). Berlin, Germany: Springer.

Hastings, P., Hughes, S., Magliano, J., Goldman, S., & Lawless, K. (2011). Text categorization for assessing multiple documents integration, or John Henry visits a data mine. In G. Biswas, S. Bull, J. Kay, & A. Mitrovic (Eds.), Artificial intelligence in education: Proceedings of the 15th International Conference, AIED 2011. Berlin, Germany: Springer.

Joachims, T. (2002). *Learning to classify text using support vector machines*. Unpublished PhD thesis, Cornell University.

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, 211–240. doi:10.1037/0033-295X.104.2.211

Landauer, T. K., Laham, D., & Foltz, P. W. (2003). Automatic essay assessment. Assessment in Education: Principles, Policy & Practice, 10, 295–308.

Lawless, K. A., Goldman, S. R., Gomez, K., Manning, F., & Braasch, J. (in press). Assessing multiple source comprehension through evidence centered design. In J. P. Sabatini & E. R. Albro (Eds.), *Assessing reading in the 21st century: Aligning and applying advances in the reading and measurement sciences.* Lanham, MD: Rowman & Littlefield.

Magliano, J. P., & Graesser, A. C. (2012). The computer-based analysis of student constructed responses. *Behavioral Research Methods*.

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

Magliano, J. P., Millis, K. K., the RSAT Development Team, Levinstein, I., & Boonthum, C. (2011). Assessing comprehension during reading with the Reading Strategy Assessment Tool (RSAT). Metacognition and Learning, 6, 131–154. doi:10.1007/s11409-010-9064-2

McNamara, D. S., & Magliano, J. P. (2009). Self-explanation and metacognition: The dynamics of reading. In D. J. Hacker, J. Dunlosky, & A. C. Graesser (Eds.), Handbook of metacognition in education (pp. 60–81). Mahwah, NJ: Erlbaum.

McNamara, D. S. et al., (2004). iSTART: Interactive Strategy Trainer for Active Reading and Thinking. Behavioral Research Methods, Instruments, and Computers, Vol. 36, pp. 222–233.

Medlock, B. (2008). *Investigating classification for natural language processing tasks* (Technical Report No. UCAM-CL-TR-721). Cambridge, U.K.: University of Cambridge.

Millis, K. K., Magliano, J. P., Todaro, S., & McNamara, D. S. (2007). Assessing and improving comprehension with latent semantic analysis. In T. Landauer, D. S. McNamara, S. Dennis, & W. Kintsch (Eds.), LSA: A road to meaning (pp. 207–225). Mahwah, NJ: Erlbaum.

Mitchell, T. (1997). Machine learning. New York, NY: McGraw-Hill.

Organization for Economic Co-operation and Development. (n.d.). *Reading literacy*. Retrieved September 17, 2010, from www.pisa.oecd.org/pages/0,3417,en_32252351_32235979_1_1_1_1_1,00.html

Papka, R., & Allan, J. (1998). Document classification using multiword features. In *Proceedings of the 7th International Conference on Information and Knowledge Management (CIKM)* (pp. 124–131). New York, NY: ACM. doi:10.1145/288627.288648

Rouet, J. F. (2006). The skills of document use. Mahwah, NJ: Erlbaum.

Rouet, J. F., Britt, M. A. (2011). Relevance processes in multiple document comprehension. In M. T. McCrudden, J. P. Magliano, & G. Schraw (Eds.), Text relevance and learning from text (pp. 19–52). Charlotte, NC: Information Age Publishing.

Snow, C. (2002). Reading for understanding: Toward an R&D program in reading comprehension. Santa Monica, CA: RAND Corporation.

van Rijsbergen, C. V. (1979). Information retrieval (2nd ed.). London, U.K.: Butterworth.

Weiss, S., Indurkhya, Zhang, T. T., & Damerau, F. (2005). Text mining: Predictive methods for analyzing unstructured information. New York, NY: Springer.

Wiley, J., & Voss, J. F. (1999). Constructing arguments from multiple sources: Tasks that promote understanding and not just memory for text. Journal of Educationl Psychology, 91(2), 301–311.

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

Wolfe, M. B. W., & Goldman, S. R. (2003). Use of latent semantic analysis for predicting psychological phenomena: Two issues and proposed solutions. Behavior Research Methods, Instruments, and Computers, 35, 22–31. doi:10.3758/BF03195494

Zhang, W., Yoshida, T., & Tang, X. J. (2007). Text classification with multi-words. In *Proceedings of 2007 IEEE International Conference on System, Man, and Cybernetics* (pp. 3519–3524). Piscataway, NJ: IEEE Press.

Zhang, W., Yoshida, T., & Tang, X. J. (2008). Text classification based on multi-word with support vector machine. Knowledge-Based Systems, 21, 879–886.

Zhang, W., Yoshida, T., & Tang, X. (2011). A comparative study of TF*IDF, LSI and multi-words for text classification. Expert Systems with Applications, 38, 2758–2765.

Zhang, W., Yoshida, T., Tang, X. J., & Ho, T. B. (2009). Augmented mutual information for multi-word extraction. International Journal of Innovative Computing, Information and Control, 5, 543–554