5 Implementation and Applications of the Intelligent Essay Assessor

Peter W. Foltz, Lynn A. Streeter, Karen E. Lochbaum, and Thomas K Landauer

INTRODUCTION

In 1998 Pearson Knowledge Technologies (formerly Knowledge Analysis Technologies) entered the business of automatically scoring text, such as essays. Prior to that, founders Landauer and Foltz had experimented with using automated essay grading in their large psychology lecture courses beginning in 1994. A typical prompt from that era was: "Describe the differences between classical and operant conditioning."

The innovation that Pearson Knowledge Technologies (PKT) brought to bear on scoring text was incorporating an automated, mathematical way of representing and assessing the content of text that corresponded to judgments that people make about the similarity in meaning between passages of text and individual words. This scoring technology is based on Latent Semantic Analysis (LSA), a machine-learning method that acquires and represents knowledge about the meaning of words and documents by analyzing large bodies of natural text. LSA is all of the following:

- A theory of how people learn the meanings of words.
- A mathematical system for computational modeling of thinking processes.
- A text analysis tool.

Thus, in addition to measures that traditionally are used to characterize text, incorporating natural language processing (NLP), readability measures, grammar and spelling errors, LSA measures gave a way to assess the semantics or content of what was written, not just its form.

LSA's ability to gauge the quality of a text's meaning at the level of human raters has produced over the last 25 years a cottage industry of new applications where content coverage and quality are the core metrics (e.g., there are about 20,000 references to LSA according to Google Scholar). Early demonstrations of this by PKT showed that LSA discriminated between high school students, undergraduates, and medical students in assessing the same essay prompt, "Describe the functioning of the human heart." Foltz and Landauer used LSA scoring in their undergraduate psychology courses for several years, with Foltz giving students the option of having him grade their psychology essay if they were dissatisfied with the machine score. Oddly enough, no one ever took him up on his offer.

Because the word "essay" connotes English Language Arts to most people, the initial market was scoring language arts essays. While LSA-based content scoring accounted for roughly 80% of the prediction variance, the English Language Arts community required stylistic and grammatical judgments, and feedback. Over the years the scoring algorithms evolved to support measurement and feedback on aspects of style and mechanics

such as grammar, punctuation, and spelling. Today approximately 60 variables have the potential to contribute to an overall essay score, as well as trait scores such as organization or conventions.

The Common Core State Standards (CCSS) change the landscape of how writing will be evaluated and what writing assignments will be. The CCSS reify the role of content in students' writing as an indication of mastery and higher order thinking skills. So, we have come full circle—PKT's shibboleth that substance matters more than form is now front and center of American curriculum reform.

In this chapter, first, far-reaching applications of PKT's Intelligent Essay Assessor (IEA) are described. Next, how the technology works for various types of scoring is described, and, finally, how well the technology works as compared to humans is described.

APPLICATIONS OF IEA

Automated Essay Scoring

The first major market for automated essay scoring was for English Language Arts (ELA). Educators, test publishers, and the public were reluctant to use automated scoring alone for summative assessment, but there was an appetite for giving students more practice writing essays in preparation for state writing exams. Teachers could grade on average three to four essay assignments a year; whereas, with computer-delivered automated scoring with instant feedback, much more writing practice could be given. For example, one teacher of PKT's WriteToLearn product has 120 students annually who produce 25,000 revisions to essays and summaries in the school year. Recognizing that writing is a contact sport that can be better played with technology, leads to students who markedly improve their writing skills.

Typical ELA curriculum includes writing in response to particular prompt types, such as narrative, expository, descriptive, and persuasive. Feedback in formative settings models the rubrics used to score state writing exams; usually a holistic score on a 4- or 6-point scale and trait scores, such as: ideas, organization, conventions, word choice, and sentence fluency. In addition, grammar and spelling errors are flagged. Figure 5.1 shows the system's scoring of a 12th grade persuasive prompt, "Should students upon graduating from high school be required to give a year of compulsory community service?"

WritetoLearn is implemented as a formative tool to continuously assess and improve skills rather than provide just an annual snapshot measure. For example, in South Dakota, it is incorporated as a formative writing assessment to replace the year-end summative state writing assessment for grades 5, 7, and 10. On average, students would revise an assignment four times: more revision practice than could or would occur in a conventional classroom with teacher grading. The results of using the system showed that student writing improved an average of one point on a six-point scale over those revisions (Foltz, Lochbaum, & Rosenstein, 2011).

Automated Scoring and Feedback of Paragraphs

Pearson's Writing Coach product is a grammar and writing curriculum program that includes automated evaluation of paragraphs as well as essays. With Writing Coach, students can write and receive feedback on each individual paragraph as they build toward a complete essay. Students receive paragraph feedback on the following features:

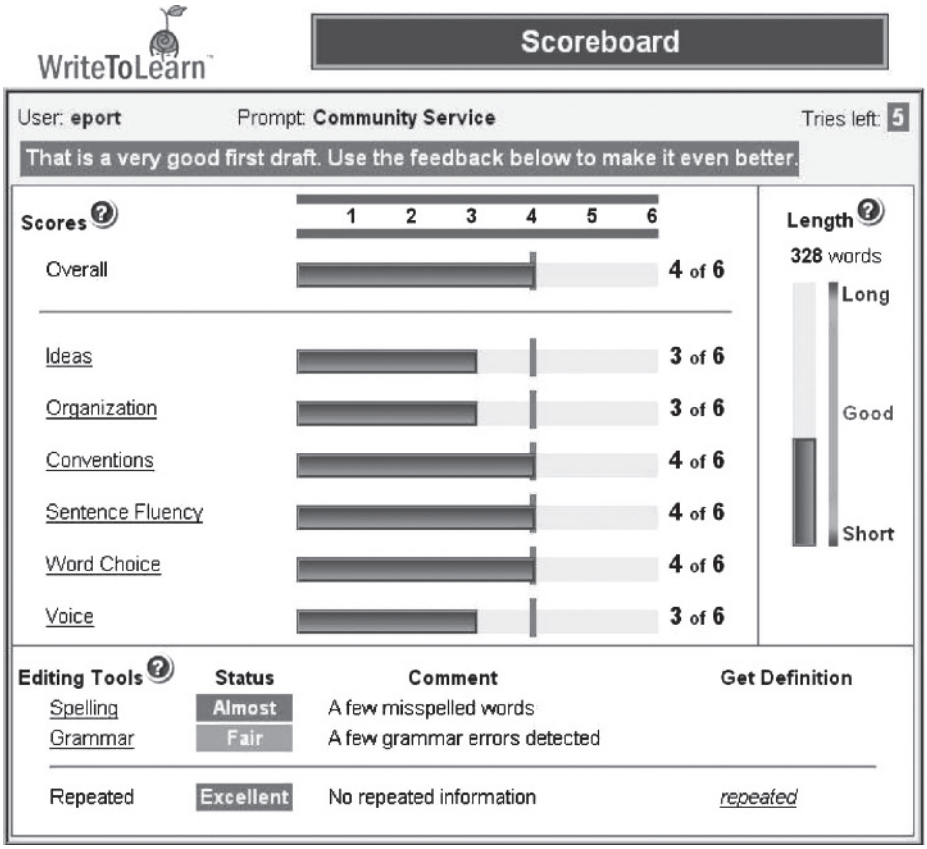

Figure 5.1 Essay feedback scoreboard. WriteToLearn provides students with an overall score as well as scores on six popular traits of writing. Passing scores are shown by the bars. Analysis of spelling, grammar, and redundancy (Repeated) is available by clicking on the links provided. Clicking on individual traits, such as Ideas or Organization provides more detailed explanations of how to improve those particular aspects of writing.

- Topic Focus: How well the sentences in the paragraph support the topic, as well a listing of those sentences that don't appear to support it.
- Topic Development: How well developed the ideas in the paragraph are.
- Variety of sentence length, sentence beginnings, and sentence structure.
- Transitions, vague adjectives, repeated words, pronouns, spelling, grammar, and redundancy (see Figure 5.2).

Summary Writing to Improve and Measure Reading Comprehension

Pearson's WriteToLearn tool expands the role of writing across the curriculum by automatically evaluating written summaries of informational texts in disciplines other than language arts. WriteToLearn's summary component evaluates writing across academic subjects, such as science, social studies, and history. Student feedback on a summary includes an assessment of how well the student covered the content in each major section of the reading, hints for how to improve content coverage in a particular section, and feedback on length, unimportant content, redundant content, and direct copying from

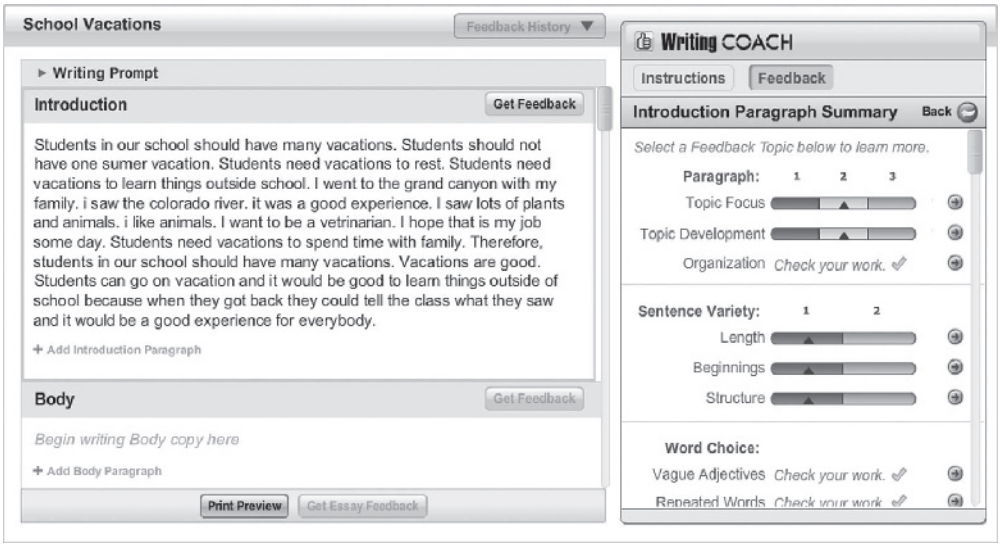

Figure 5.2 Writing Coach feedback for paragraphs. Writing Coach provides students with feedback on individual paragraphs including ratings for topic focus, topic development, and sentence variety, as well as feedback on word choice and spelling, grammar, and redundancy.

the original text. Scoring is accomplished by analyzing both the passage sections and summary for their holistic meanings, not by looking for particular key words.

Figures 5.3a through 5.3c show the summarization flow. First, a student reads a passage about penguins. Next, the student writes a summary of the passages that was just read. After submitting the passages for scoring, feedback is provided in the *scoreboard* which shows how well the student covered the content of each major section in the

Figure 5.3a Text for student to read.

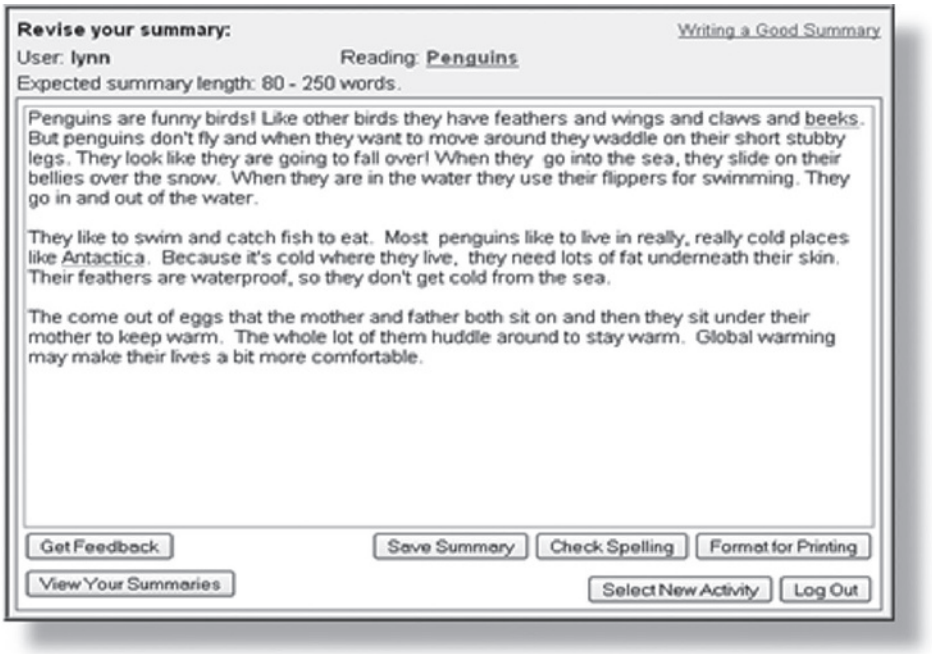

Figure 5.3b Student summary writing.

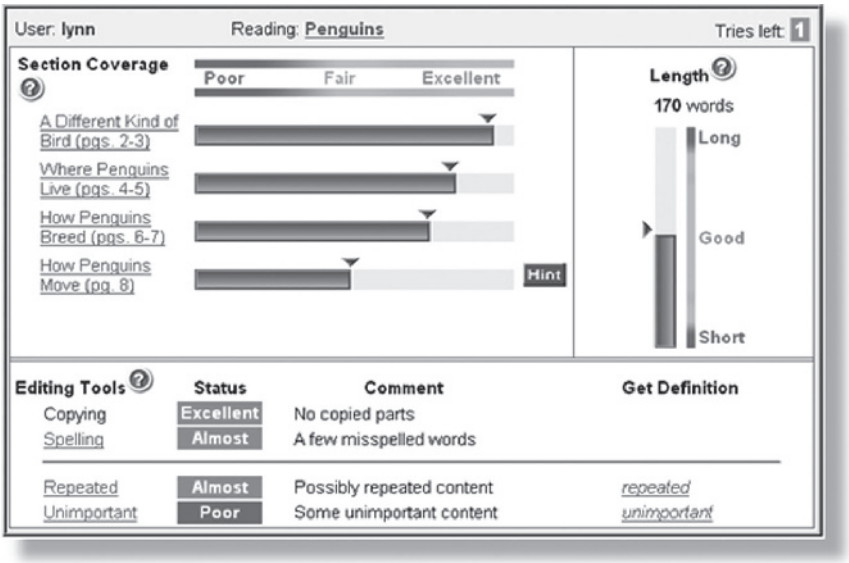

Figure 5.3c Summary Street feedback screen. This scoreboard is presented immediately after the student clicks "Get Feedback."

reading. The triangles above the content coverage bars show performance achieved on a previous submission. Students are encouraged to re-read the sections on which they are not doing well and revise their summary to push the score bars for each section over the passing threshold shown between Fair and Excellent.

Summary Writing to Improve Reading and Writing Skills

Summary Street was the result of joint development between researchers at the University of Colorado and scientists and software developers at Knowledge Technologies.[1] Several controlled studies were performed on the effects of Summary Street in the classroom. In one study 60 students in two sixth-grade classes each wrote two texts, one class using Summary Street and the other using a standard text editor. Results from this small study (see Figure 5.4) indicated that the students who used Summary Street:

- received higher grades on their summaries as assessed by teachers blind to the condition to which the student was assigned
- spent more time on the writing task
- retained the skills they learned even after they stopped using the tool.

In another study (Franzke, Kintsch, Caccamise, Johnson, & Dooley, 2005), 121 students used Summary Street for four weeks or were in a control group who received the same training but did their summary writing on word processors, which did not give the automated summary feedback. Students with Summary Street improved their content summary scores by an overall effect size of $d = 0.9$ compared to the control students. The results indicate that for a class of mixed-ability students, students scoring at the 50th percentile improved their writing performance with more difficult materials to the 82nd percentile. When the performance of low- and medium-ability students (the lower 75% of the distribution) was considered, the effect size increased to $d = 1.5$ for the most difficult materials. (An effect size of 1.0 corresponds to approximately a one-grade difference, e.g. from fifth to sixth grade.)

In a third study, University of Colorado researchers conducted a large two-year evaluation of 2,851 students in grades 5–9 in nine Colorado schools districts (Caccamise et al., forthcoming). Classes of students were assigned to either use Summary Street or

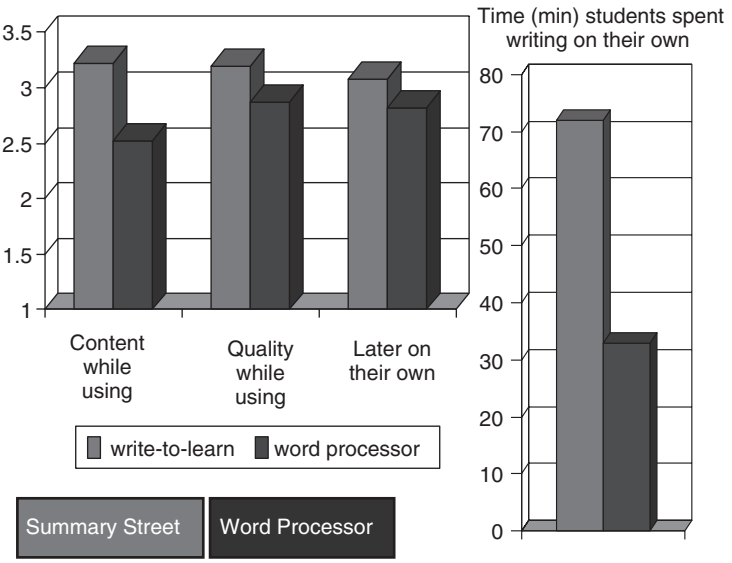

Figure 5.4 Summary Street produces better essays as judged by teachers in a two-week trial of sixth-grade students.

to receive traditional teacher-provided summarization instruction. Of the students who used Summary Street, most of them used it for an average of five to six different texts throughout the year. Students were given a summarization pretest at the beginning of the school year and at the end of the school year, as well as a standard short reading comprehension test (Test of Reading Comprehension, or TORC) at the beginning and end of the year. The experimental group was superior to the control group in summary writing for both years. Improvement in summarization was highly related to the number of texts a student studied during the year, as well as the amount of time students spent using the tool. Comprehension improvements on the TORC test were highly related ($p < .002$) to the amount of Summary Street use (see Figure 5.5).

Automated Essay Scoring in Postsecondary Environments

Use of automated essay scoring has been less prevalent in the postsecondary arena. The greatest impediment to adoption is that in college settings each professor creates his or her own assignment for a class of 10 to 500 students. When subject-area content needs to be evaluated, scoring models for unique (new) prompts typically need to be created. Thus, automated essay grading requires economies of scale in order to be cost-effective. Florida Gulf Coast University offers an example of IEA use where scale preconditions were met. "*Understanding the Visual and Performing Arts*" was a required freshman course with 800 students taught across 30 sections by adjunct professors. For the essay writing requirement, students analyzed a work of art—such as a painting, a sculpture, a piece of architecture, or a performing arts piece such as music, dance or theater. The essay prompts asked students to provide an objective analysis of particular elements of the art work, as well as to explain the meaning created by the particular work. While grading a "creative" essay might seem to be a particular challenge for automated assess-

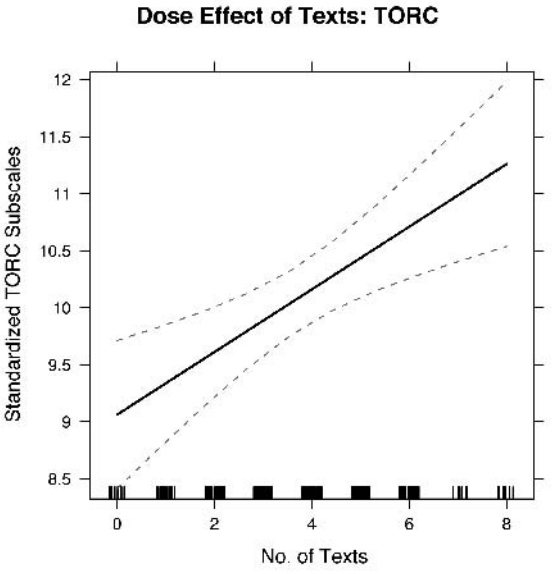

Figure 5.5 Performance on a standard reading comprehension test as a function of the number of texts studied with Summary Street during the school year with the 95% confidence interval shown as a solid line. The pretest scores for the same test were used as a covariate to control for student ability.

ment, professors were very pleased with the results. An analysis of IEA's reliability in grading showed it was more consistent than human scoring, matching human graders' scores 81% of the time vs. 64% when scored by two independent graders.

About 40% of students enrolling in community college need remediation in literacy, math or both. This is an ideal situation for automated essay grading. The size of this population is large, heterogeneous, and growing. To make progress in literacy skills such as writing requires dedicated hours of practice. And there are so many students in need that dedicated tutors are not an option. Pearson uses IEA in MyWritingLab, a web-based practice and assessment environment for the developmental writing market. As of the writing of this chapter, additional writing programs in the areas of science, history, social science, and business programs are implementing IEA to assess content knowledge in several MyLab learning environments for mainstream college students.

Performance Task Scoring

The ability to automatically evaluate content enables scoring of complex tasks, such as responding to scenarios, which require application and synthesis of complex knowledge either already possessed by the learner or gained experientially by taking a vocational or academic course. The first example of performance-based scoring comes from the Collegiate Learning Assessment, an assessment that is scored operationally by Pearson's automated scoring services. One task type involves presenting students with a scenario and a variety of information sources and asking the student to synthesize the information in a written response. An example of such an item is shown in Figure 5.6. Automated scoring performance on these types of items has an average Pearson correlation of 0.88 with the human consensus score, whereas the human–human correlation was 0.79.

Automated Assessment of Diagnostic Skills—National Board of Medical Examiners

A final example of automated assessment of performance tasks comes from a study done with the National Board of Medical Examiners where IEA was used to rate a physi-

You advise Pat Williams, the president of DynaTech, a company that makes precision electronic instruments and navigation equipment. Sally Evans, a member of DynaTech's sale force, recommends that DynaTech buy a small private plane (a SwiftAir 235) that she and other member of the sales force could use to visit customers. Pat was about to approve the purchase when there was an accident involving a SwiftAir 235.

Resources: Document Library
Newspaper article about the accident
Federal Accident Report on in-flight breakups in single engine planes
Internal Correspondence (Pat's email to you and Sally's e-mail to Pat)
Charts relating to SwiftAir's performance characteristics
Excerpt from magazine article comparing SwiftAir 235 to similar planes
Pictures and descriptions of Swiftair Models 180 and 235

Using the resources provided, the student writes a memorandum to the president presenting a reasoned decision of whether or not to purchase the Swiftair jet.

Figure 5.6 Sample performance task.

cian evaluation in simulations in which doctors examine and diagnose actors posing as patients feigning diseases. The patient notes produced by the doctors in this evaluation are illustrated in Figure 5.7. The notes are divided into text sections by patient's history, the findings of the doctor's physical examination of the patient, the potential diagnoses, and the additional diagnostic tests to be performed. IEA correlated more highly with the rating of the patient notes than did the expert physician ratings (Swygert et al., 2003).

The Collegiate Learning Assessment and medical example are just two examples in which IEA has been used in assessment other than standard language arts essays. Pearson has also done work with the military studying automated assessment of *Think Like a Commander* scenarios in which officers are presented with a scenario and asked to write a response detailing their approach to the scenario and the steps they would take. Automated scoring performance on such scenarios was shown to match that of the expert military evaluators (Lochbaum, Psotka, & Streeter, 2002).

Knowledge Technologies has used IEA to assess learning and performance in online collaborative work environments (see Foltz & Martin, 2008; Streeter, Lochbaum, LaVoie, & Psotka, 2007) to automatically monitor online discussion groups to alert the instructor to discussion drift; to assess relative contributions of participants; and to enhance the value of the discussion by automatically placing expert commentary into the discussion based on assessing the quality of the student discussion (LaVoie et al., 2010). IEA has also been used in psychiatric settings as a means of assessing clinical disorders to predict depression and schizophrenia from retelling familiar stories or from LSA analysis of transcribed psychiatric interviews in which the patient describes routine daily tasks (Elvevåg, Foltz, Weinberger, & Goldberg, 2007).

Short-Answer Scoring for Science

Pearson and the Maryland State Department of Education have worked together since 2007 on evaluating automated scoring for the Maryland Science Assessment (MSA). Since 2010, Pearson's automated scoring system has been used as a second scorer for short-answer science items that are best suited to automated scoring techniques. Short-

History	L-upper arm dull pain upon exertion (walking x2 weeks, each episode lasting <5 minutes, one episode at rest last night. 2. No associated chest pain, shortness of breath, numbness, parasthesia, weakness/paralysis, dizziness, syncopal episodes. 3. Past medical history of hypertension. 4. Post-menopausal, no hormone replacement therapy. Occupational history, social history, social history negative for activities that may contribute to arm strain.
Physical Examination	1. No focal tenderness, erythema, warmth. 2. L-upper extremity exam with normal pulse, capillary refill, motor/sensory function, reflexes.
Differential Diagnosis	Tendonitis Bursitis Angina Pectoris
Diagnostic Workup	EKG CBC Plain film X-ray L-upper extremity

Figure 5.7 Sample text from patient diagnostic notes.

Use the technical passage 'Green Ocean Machine' to answer the following.

The passage states that "the new green partner [alga] seems to provide Hatena with most of its energy needs."

Describe the process that enables organisms to use energy from light to make food. In your description, be sure to include:
the specialized features needed to produce food
the substances needed to produce food
the substances produced during this process

Figure 5.8 Sample Maryland science short answer item.

answer science questions on the Maryland assessment, which average about 50 words, are scored automatically on a 4-point holistic scale. For some types of prompts, students read a passage on a scientific question and write a response after reading the relevant text. For others, they simply respond to a given scientific question. One example item scored successfully using automated scoring is as follows.

HOW IEA SCORES

The IEA uses machine-learning techniques to learn how to score based on the collective wisdom of trained human scorers. Training IEA involves first collecting a representative sample of essays that have been scored by human raters. IEA extracts features from the essays that measure aspects of student performance such as the student's expression of knowledge and command of vocabulary and linguistic resources. Then, using machine-learning methods, IEA examines the relationships between both the scores provided by the human scorers and the extracted features in order to learn how the human scorers weigh and combine the different features to produce a score. The resulting representation is referred to as a "scoring model." This section provides details of the scoring features used in IEA, how the features are combined to score different traits of writing, and considerations for building and evaluating the performance of scoring models.

IEA Scoring Features

The quality of a student's essay can be characterized by a range of features that measure the student's expression and organization of words and sentences, the student's knowledge of the content of the domain, the quality of the student's reasoning, and the student's skills in language use, grammar and the mechanics of writing. In developing analyses of such features, the computational measures extract aspects of student performance that are relevant to the constructs for the competencies of interest (e.g., Hearst, 2000; Williamson et al., 2010). For example, a measure of the type and quality of words used by a student provides an effective and valid measure of a student's lexical sophistication. However, a measure that counts the number of words in an essay, although it will likely be highly correlated with human scores for essays, does not provide a valid measure of sophistication of writing. Because a student's performance on an essay typically requires showing combined skills across language expression and knowledge, it is critical that the scoring features used in the analysis cover the construct of writing that is being

scored. Thus, multiple language features are typically measured and combined to provide a score. IEA uses a combination of features that measure aspects of the content, lexical sophistication, grammar, mechanics, style, organization, and development within essays. Figure 5.9 illustrates some of the features used in IEA and how they relate to specific constructs of student writing performance.

Content-Based Features in IEA

One of the hallmarks of IEA has been its ability to score essays in content domains. IEA uses LSA, a statistical semantic model (Deerwester et al., 1990; Landauer & Dumais, 1997) as the basis for scoring content features. LSA derives semantic models of English (or any other language) from an analysis of large volumes of text. For essay scoring applications, we typically use a collection of texts that is equivalent to the reading a student is likely to have done over their academic career (about 12 million words). LSA builds a co-occurrence matrix of words and their usage in paragraphs and then reduces the matrix by Singular Value Decomposition (SVD), a technique similar to factor analysis. The output of this analysis is a several hundred dimensional semantic space in which every word, paragraph, essay, or document is represented by a vector of real numbers to represent its meaning. The semantic similarity between words, paragraphs, and essays can be determined by computing the cosine between the vectors of two units of text. For example, the sentence "Surgery is often performed by a team of doctors" has a high semantic similarity to "On many occasions, several physicians are involved in an opera-

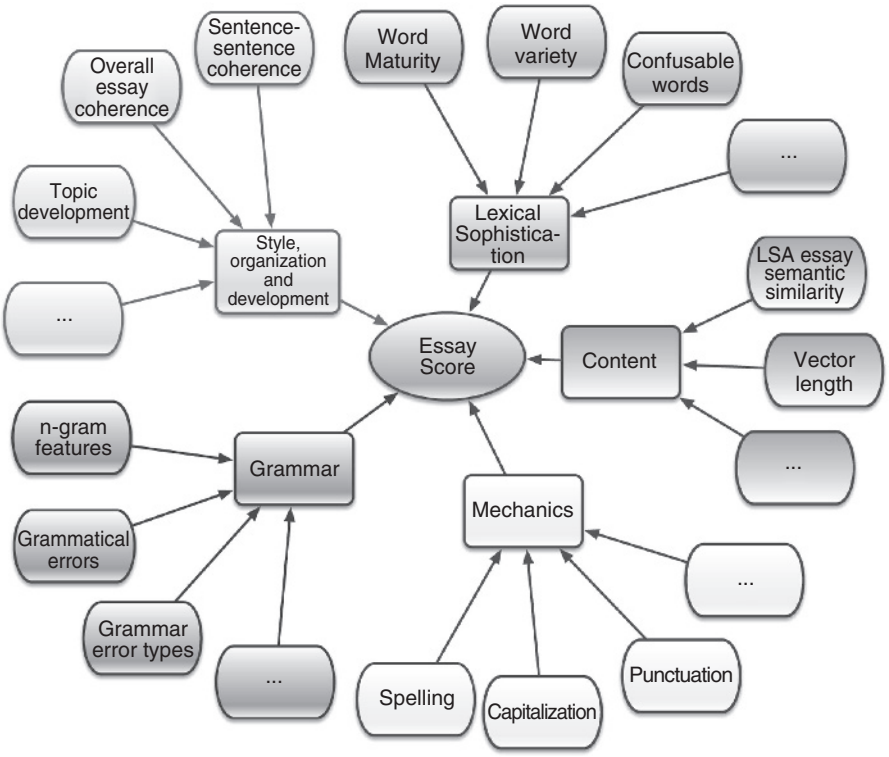

Figure 5.9 Features used in the IEA.

tion" even though they share no words in common. Although the technique is based on the statistics of how words are used in ordinary language, its analysis is much deeper and more powerful than the simple frequency, co-occurrence, or keyword counting and matching techniques that have sometimes been used in traditional NLP techniques. For an overview of NLP methods, see Chapter 4.

LSA is now in wide use around the world in many applications in many languages, including Internet search, psychological diagnosis, signals intelligence, educational and occupational assessment, intelligent tutoring systems, and in basic studies of collaborative communication and problem solving. The accuracy of the LSA meaning representation has been empirically tested in many ways. For example, LSA improves recall in information retrieval, usually achieving 10–30% better performance *cetera paribus* by standard metrics (Dumais, 1994; Landauer & Dumais, 1997) by matching documents with similar meanings, but utilizing different words. After training on domain corpora from which humans learned or might have learned, LSA-based simulations have passed multiple choice vocabulary tests and textbook-based final exams at student levels (Landauer, Foltz, & Laham, 1998). In rating the similarity of meaning between pairs of paragraphs and the similarity of meaning between pairs of words, LSA measures the similarity of meaning 90% as well as two human raters do when agreeing with each other about word and paragraph meanings (Landauer et al., 1998). LSA has been found to measure coherence of text in such a way as to predict human comprehension as well as sophisticated psycholinguistic analysis, while measures of surface word overlap fail badly (Foltz, 2007; Foltz, Kintsch, & Landauer, 1998).

Within IEA, LSA is used to derive measures of content, organization, and development-based features of writing. For example, the "LSA essay semantic similarity" measure compares the semantic similarity of a student essay against a set of training essays of known quality. A content score is assigned to the essay based on the scores of the most similar essays, weighted by their semantic similarity. This correlates highly with human scores of essays (see Landauer, Laham, & Foltz, 2001, 2003; Rehder et al., 1998). LSA-based measures are also used to compare the content of individual sentences to each other to compute measures of coherence (see Foltz et al., 1998) as well as to computing semantic similarity of the content of sentences or paragraphs against gold standard samples (see Foltz, 1996; Foltz, Gilliam, & Kendall, 2000). Finally, measures based on the LSA-based vector length of an essay in the semantic space are used. The vector length in an LSA-based semantic space provides an index of the preciseness of content within an essay (Rehder et al., 1998).

Other Language Features in IEA

Along with content-based measures, a range of other automatically computed measures are also used to score the lexical sophistication, grammatical, mechanical, stylistic, and organizational aspects of essays. Measures of lexical sophistication include measuring the developmental maturity of the words used (see Landauer, Kireyev, & Panaccione, 2011) as well as the variety of types of words used. Grammar and mechanics measures use NLP-based approaches to analyze the specific linguistic features of the writing. For grammar, such measures detect run-on sentences, subject–verb agreement, and sentence fragments use of possessives, among others. For assessing mechanics, measures are used that examine appropriate spelling, punctuation, and capitalization.

The assessment of stylistic and organizational aspects of essays are evaluated using a combination of LSA-based measures to analyze coherence in the essay, as well as

NLP-based measures that assess aspects of the organization, flow, and development across the essays. In addition, for specific essay types, additional features are incorporated which assess aspects of topic development, such as the strength of an introduction, use of supporting arguments, and the quality of the conclusion. Unless explicitly called for by a test design and documented for users, measures based on raw counts of words, sentences, or paragraphs are not included (e.g., counting words, adjectives, number of occurrences of "therefore"). While these measures can be predictive, students can be too easily coached to exploit such count-based measures.

Building a Scoring Model

IEA is trained to associate the extracted features in each essay to scores that are assigned by human scorers. A machine learning-based approach is used to determine the optimal set of features and the weights for each of the features to best model the scores for each essay. From these comparisons, a prompt and trait-specific scoring model is derived to predict the scores that the same scorers would assign to any new responses. Based on this scoring model, new essays can be immediately scored by analysis of the features weighted according to the scoring model.

Training on Human-Scored Data

The sample of student responses used for training and evaluating the scoring engine should represent the full range of student responses and scores. Typically the set of essays should represent a normal distribution, while ensuring that there are sufficient (e.g., at least a minimum of 10–20) examples at each score point. During training of the system, the responses should be 100% double-scored by human scorers and also receive resolution scores for non-adjacent agreement. By having scores from multiple human scorers, IEA can be trained on something closer to the true score (e.g., the average of multiple human raters) rather than the scores of an individual rater. The goal is to have as much and as accurate information as possible about the range of possible responses and how those responses should be evaluated. Generally, essay sets that are not as accurately scored by human raters will result in less accurate automated scoring models (e.g., Foltz, Lochbaum, Rosenstein, & Davis, 2012).

The number of responses typically required to train the scoring engine varies depending on the type of prompt and expected use of the response. For general formative and content-based scoring, 200–300 essays are required to train the scoring engine. For an essay prompt in a high-stakes assessment, a sample of about 500 student responses is preferred, while for a short-answer prompt 1000 responses are recommended for best performance (see Foltz et al., 2012) These numbers allow for using part of the data to train the scoring engine while holding out the other part for testing and validation. If such numbers of responses do not exist, a smaller testing and validation set can be used, or, alternatively, techniques such as jack-knife methods can also be used for evaluating expected performance based on the training set results alone.

Types of Scorable Traits

Human scorers are able to score essays for different traits within essays by focusing on different features of the essays in their evaluation. For example, to score an essay on conventions, a human scorer would focus on a student's grammar, spelling, and punctua-

tion. Similarly, IEA can generalize to scoring different traits by choosing and weighting different combinations of features. A subset of the features can be used in the training, for example just choosing features related to conventions if scoring a convention trait. By then training IEA on human scores, it learns to associate the features within the IEA set that best model judgment on a specific trait. IEA has been used to accurately score a range of traits including:

- overall quality
- content
- development
- response to the prompt
- effective sentences
- focus and organization
- grammar, usage, and mechanics
- word choice
- development and details
- conventions
- focus
- coherence
- reading comprehension
- progression of ideas
- style point of view
- critical thinking
- appropriate examples, reasons and other evidence to support a position.
- sentence structure
- skilled use of language and accurate
- apt vocabulary.

Evaluating Responses for Scorability

Before scoring a student response, IEA analyzes the response to determine the confidence with which it can score it accurately. IEA uses a variety of statistical and probabilistic checks to make this determination based on characteristics of the response on which it was trained and experience with a variety of both good- and bad-faith responses. Responses that appear to be off topic, not English, or highly unusual or creative will be directed to a human for scoring.

Variants on IEA for Scoring Different Types of Student Responses

Short-answer scoring

Short-answer responses (e.g., responses on the order of 5 to 50 words) pose somewhat different scoring problems than longer essays. A sample student response is shown in Figure 5.10 that illustrates some of these problems.

A first problem is that responses of a sentence or two can be challenging because they contain very little information with which to evaluate a student's knowledge and ability. Second, spelling has a critical effect on short responses. If the majority of the words in a response are misspelled, it is very difficult to evaluate anything but the student's spelling ability. Third, short-answer items can often be very open-ended and so the range of

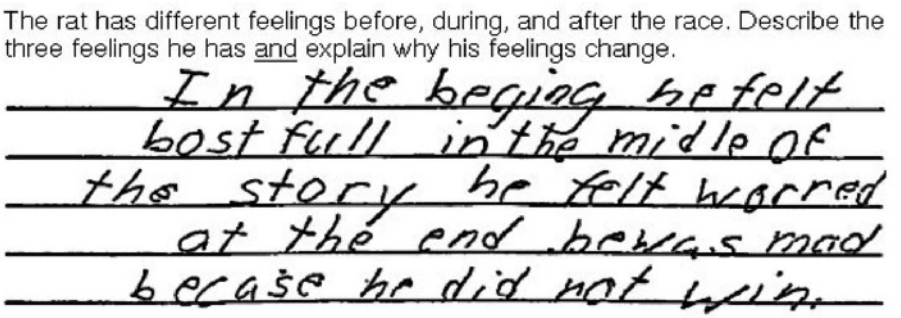

Figure 5.10 Short answer student response.

acceptable possible responses very broad. In contrast with essays, the quality of short-constructed responses is also characterized more by word choice and the usage of specific terminology. To address these differences, a variant of IEA is used for scoring short-answers. In addition to the features from IEA, the short-answer variant uses statistical classifiers and assessment-specific heuristics for treating ordering of events in a process or explanation to model each short answer. In addition, compared to essay scoring, the development of short-answer-response scoring requires more student data to reach the accuracy required for high-stakes use. Based on research with the State of Maryland over five years, we have found that about one half to two-thirds of the short-answer science items can be scored automatically with similar accuracies to human scorers (see Thurlow, Hermann, & Foltz, 2010; Thurlow, Hermann, & Lochbaum, 2011). In these cases, the automated scoring system operates as a second scorer on those questions. For the remaining items, double human scoring is used exclusively.

Summary scoring

Summary writing allows students to practice both reading comprehension and writing across content areas. Automatic evaluation of summaries enables students to participate in a read, write, and revise cycle that encourages them to re-read, rethink and re-express those parts of the text that they have not yet fully understood. Our automatic summary evaluation process measures how well a student's summary covers the content of each major section of a reading by calculating the semantic similarity between the summary and each section of the text. Studies of its use in classrooms have shown that it produces improved reading comprehension and improved content writing when compared to students who did not receive automated feedback (Franzke et al., 2005).

Prompt-independent scoring models

In the examples of scoring models described above, IEA is trained specifically on each prompt for scoring traits associated with the particular assessment. When educators are mainly interested in gauging the stylistic and mechanical aspects of writing, a variant of IEA has been developed that provides a generalized grading (prompt-independent) scoring method. The prompt-independent method was calibrated on thousands of essays across multiple topics and prompts. Prompt-independent scoring is somewhat less reliable (self-consistent) than prompt specific scoring (generally about 10% lower reliability).

However, for formative instructional use, prompt-independent scoring enables teachers to author essay prompts that are tied to their own lesson plans and curriculum.

EVALUATION OF SCORING ENGINE PERFORMANCE

Evaluation of the performance of a scoring engine is critical throughout the test development process. In the pilot testing phase of item development, evaluation is performed to determine how amenable items are for automated scoring. Before deployment, finalized scoring models are evaluated on held-out tests sets to determine generalizability and robustness of scoring. During deployment, evaluation of the scoring engine is often performed to ensure that the scoring remains consistent with the goals of the testing. In the case of IEA being used as the sole scorer, random samples of essays can be chosen for backreads by human scorers as a check on the automated scoring. When IEA is being used as a second scorer, agreement rates with the other human scorer as well as with resolution scorers can be constantly monitored for performance. In addition, when used as a second scorer, evaluation of the agreement with human scorers can be used to detect drift in the human scorers and scorer consistency.

The performance of a scoring model can be evaluated both in how well the scores match human scoring, but also how well the scores align with the constructs of interest (see also Williamson, Chapter 10). The most common benchmark is to compute the reliability of the scoring engine by examining the agreement of IEA's predicted scores to human scorers, as compared to the agreement between human scorers. Metrics for computing the reliability include correlation, kappa, weighted kappa, and exact and adjacent agreement. Using "true scores" (e.g., the average of multiple scorers or the consensus score) for the comparison can provide more accurate measures of IEA's accuracy. However, human agreement is seldom sufficient as a means to evaluate performance. IEA performance can be compared against external variables that provide a measure of the validity of the scoring, including comparison of IEA scores with scores from concurrent administrations of tests with a similar construct, agreement with scores from subsequent tests, predicting student age or grade level, agreement to scorers with different levels of skill, and tests of scoring across different population subgroups. It should be noted though that when used in a formative context, evaluation should also be considered within a framework of measuring learning gains. As students receive formative feedback, revise their essays and resubmit, automated scoring can be evaluated in how well it improves students' content knowledge, reading skills (e.g., Franzke et al., 2005) and writing abilities (e.g., Foltz et al., 2011).

IEA has been evaluated across a range of different types of essays at different levels of student ability. Table 5.1 presents correlation coefficients between automated scores and consensus human scores for a sample of written constructed responses. For example, the third row shows score accuracy indicators for a set of five information-integration items, the Collegiate Learning Assessment items described above. For each item, students were asked to write memos that synthesized information from multiple sources, including letters, memos, summaries of research reports, newspaper articles, maps, photographs, diagrams, tables, charts, and interview notes or transcripts. The resulting set of 1239 written responses was then scored by machine and by independent scorers. One can compare the average Pearson correlation between pairs of human scorers, shown as the human–human correlation of 0.79, with machine–human correlation of 0.88. Thus, for this example, automatic scores are closer to a stable consensus human score than one expert score is to another.

Table 5.1 Indicators of Scoring Quality for Four Operational Item Sets. N is the average the number of test-takers per test or item used in the calculations. Machine–human correlations are between one fully automatic score and a consensus human score (rating the same material). For comparison, the human–human column shows correlations between scores from two human scorers for the same materials.

Assessment Prompt Material	*N*	*Machine–Human Pearson Correlation*	*Human–Human Pearson Correlation*	*Source*
81 published essay prompts (grades 6–12)	200	0.89	0.86	Prentice Hall
18 research-leveled essay prompts (grades 4–12)	635	0.91	0.91	MetaMetrics
5 performance tasks using multiple sources	1239	0.88	0.79	Council for Aid to Education
10 essay prompts used for placement	4858	0.90	0.90	ACCUPLACER

Note that the correlations shown in Table 5.1 are item-level correlations. Assessments typically include many different types of items designed to get as accurate a measure as possible of a student's knowledge, ability, and skill level. At the assessment level, taking all items into account, correlations between human scorers and between human and automated scorers are typically higher, approaching 0.95 or above.

Trait Scoring

While Table 5.1 shows scores for the overall quality of the essays, IEA can provide scores for individual traits of writing as well. The performance for scoring six writing traits based on six prompts is shown in Table 5.2. For each of the prompts, students were directed to read a particular text and respond in the context of the text. The prompts asked students to compare and contrast components of the reading, identify and synthesize particular aspects of the reading, and use important and specific details from the reading to support their response.

Table 5.2 Performance on Essay Prompts. Correlation and exact agreement between automated and human scores for the six traits based on six different prompts.

	Machine–Human Pearson Correlation	*Machine–Human Exact Agreement*	
Traits		*Mean (%)*	*Range (%)*
Ideas	0.93	73	58–87
Organization	0.93	69	60–74
Word Choice	0.93	68	63–74
Sentence Fluency	0.90	64	58–70
Conventions	0.85	60	52–66
Voice	0.92	65	62–68

IEA Applied to Content Scoring

Because IEA can score content-based essays, it has been applied to a range of content areas including history and social science topics for students ranging from grade school to college level. Because the automated scoring system is trained on the semantics of the domain (subject area), it is able to provide reliable scores of the content knowledge of students. For example, tests were performed on high school and entry-level undergraduate writing prompts on the history of the Great Depression, the history of the Panama Canal, ancient American civilizations, alternative energy sources, business and marketing problems, psychology of attachment in children, aphasia, and Pavlovian and operant conditioning (Foltz, Laham, & Landauer, 1999; Landauer et al., 2001). The average inter-rater correlation of two human scorers was 0.75 and the average correlation of the automated scoring to each single rater was 0.73—which was not significantly different from each other. The results further showed that the greater the expertise of the human rater, the greater the correlation to the automated scorer, thereby providing a measure of the validity of the scoring. In addition to providing scores on content, the methods are able to provide feedback about different aspects of the content where students have stronger or weaker knowledge. Feedback from this automated scoring has shown to significantly improve student content learning (see Foltz et al., 2000).

Short-Answer Scoring

As described above, short-answer scoring uses a modified version of IEA to account for different language and content features found in short responses. This version of IEA is being used operationally for scoring the State of Maryland's science assessment. Maryland's approach to item development is to create items independently of automated scoring considerations. The items are then evaluated for how well they can be scored automatically. Those responses that can be scored reliably by automated scoring techniques are scored by one human scorer and the automated system. Those responses that cannot be scored automatically continue to be scored by two human scorers. Since 2010, Pearson's automated scoring system has participated in operational scoring, acting as the second scorer for roughly two-thirds of the items on the Maryland assessment (see Thurlow et al., 2010, 2011). Table 5.3 summarizes scoring performance for ten

Table 5.3 Short Answer Scoring Performance

	Human–Human				*IEA–Human*			
Prompt	*N*	*R*	*Exact*	*Adj*	*N*	*R*	*Exact*	*Adj*
1	1507	0.71	79	100	1471	0.76	75	100
2	695	0.52	70	100	674	0.66	71	100
3	675	0.76	75	99	642	0.85	75	99
4	661	0.68	77	100	638	0.79	78	100
5	680	0.64	71	100	669	0.73	73	100
6	885	0.70	81	100	843	0.75	82	100
7	702	0.80	76	99	672	0.85	76	100
8	1674	0.57	70	99	1624	0.68	71	99
9	1666	0.81	78	100	1610	0.89	83	100
10	500	0.87	81	100	500	0.86	78	100

operational prompts showing the correlation, exact, and adjacent levels of agreement for human–human and IEA–human scoring. Overall, IEA–human performance is at an equivalent level as human–human performance.

CONCLUSION

Every year millions of elementary, secondary, and college and career essays and summaries are evaluated by IEA. It is used as a backend scoring system for summative tests, publishers' textbooks, for test preparation, as well as for products like WriteToLearn and Writing Coach. IEA provides a means to incorporate accurate scoring for a wide range of written responses including language arts, content, reading comprehension summaries, and short-answer essays, as well as responses in performance items, situation judgment tasks and clinical assessments.

As a formative tool, IEA provides more revision practice than could occur in a conventional classroom with teacher grading. As such, it helps support the trend to replace annual summative assessments with formative tools to improve skills rather than use an annual snapshot measure. This permits more personalized learning for the student and allows the teacher to focus on students that need help by monitoring the learning of individual students and the class as a whole. Because there is a great deal of commonality among state rubrics for evaluating essays, the secondary market for automated scoring is very large and the tools are widely applicable. Across America there is tremendous appetite to improve students' writing and reading skills as the situation is dire and is rightfully labeled a crisis by the educational establishment, employers, parents, and students. Tools that make practice simple and enjoyable and provide meaningful feedback can be keys to remediating literacy.

NOTE

1. WriteToLearn's Summary Street component is based on ten years of research and evaluation, as part of an Interagency Educational Research Initiative (IERI) research and effectiveness trial project, combined with seven years of professional educational software development and both software and educational effectiveness testing at Knowledge Analysis Technologies (since 2004, Pearson's Knowledge Technologies group). At University of Colorado, the research was performed under the direction of Professors Walter Kintsch and Tom Landauer, and at New Mexico State University under Professor Peter Foltz. Landauer and Foltz currently direct research at Pearson's Knowledge Technologies group. Professor Louis Gomez at UCLA Policy and Dr. Jack Stenner of MetaMetrics, Inc have also collaborated in the WriteToLearn research.

REFERENCES

Caccamise, D., Snyder, L., Allen, C., DeHart, M., Kintsch, E., Kintsch, W., & Oliver, W. (forthcoming). *Summary street: Scale-up and evaluation.*

Deerwester, S., Dumais, S., Furnas, G. W., Landauer, T. K, & Harshman, R. (1990). Indexing by Latent Semantic Analysis. *Journal of the American Society for Information Science, 41*(6), 391–407.

Dumais, S. T. (1994). Latent Semantic Indexing (LSI) and TREC-2. In D. Harman (Ed.), *The Second Text REtrieval Conference (TREC2)* (pp. 105–116). National Institute of Standards and Technology Special Publication 500–215.

Elvevåg, B., Foltz, P. W., Weinberger, D. R., & Goldberg, T. E. (2007). Quantifying incoherence in speech: An automated methodology and novel application to schizophrenia. *Schizophrenia Research*. doi: doi:10.1016/j.schres.2007.03.001.

Foltz, P. W. (1996). Latent Semantic Analysis for text-based research. *Behavior Research Methods, Instruments and Computers, 28*(2), 197–202.

Foltz, P. W. (2007). Discourse coherence and LSA. In W. Kintsch, T. K. Landauer, D. McNamara, & S. Dennis (Eds.), *LSA: A road to meaning*. Mahwah, NJ: Lawrence Erlbaum Publishing.

Foltz, P. W., Gilliam, S., & Kendall, S. (2000). Supporting content-based feedback in online writing evaluation with LSA. *Interactive Learning Environments, 8*(2), 111–129.

Foltz, P. W., Kintsch, W., & Landauer, T. K (1998). The measurement of textual coherence with Latent Semantic Analysis. *Organizational Process, 25*(2–3), 285–307.

Foltz, P. W., Laham, D., & Landauer, T. K. (1999). The intelligence essay assessor: Applications to educational technology. *Interactive Multimedia Electronic Journal of Computer-Enhanced Learning, 1*(2). Available at http://imej.wfu.edu/articles/1999/2/04/

Foltz, P. W., Lochbaum, K. E., & Rosenstein, M. B. (2011). *Analysis of student ELA writing performance for a large scale implementation of formative assessment*. Paper presented at the the Annual Meeting of the National Council for Measurement in Education (NCME).

Foltz, P. W., Lochbaum, K. E., Rosenstein, M. B., & Davis, L. E. (2012). *Increasing reliability throughout the automated scoring development process*. Paper presented at the Annual Meeting of the National Council for Measurement in Education, Vancouver, CA.

Foltz, P. W., & Martin, M. J. (2008). Automated communication analysis of teams. In G. F. Goodwin, E. Salas, & S. Burke (Eds.), *Team effectiveness in complex organizations and systems: Cross-disciplinary perspectives and approaches*. New York: Routledge.

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.

Hearst, M. A. (2000). The debate on automated essay grading. *IEEE Intelligent Systems & Their Applications, 15*(5), 22–27.

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.

Landauer, T. K, Foltz, P. W., & Laham, D. (1998). Introduction to Latent Semantic Analysis. *Discourse Processes, 25*(2–3), 259–284.

Landauer, T. K, Kireyev, K., & Panaccione, C. (2011). Word maturity: A new metric for word knowledge. *Scientific Studies of Reading, 15*(1), 92–108.

Landauer, T. K, Laham, D., & Foltz, P. W. (2001). Automated essay scoring. *IEEE Intelligent Systems*, September/October, 27–31.

Landauer, T. K, Laham, D., & Foltz, P.W. (2003). Automated scoring and annotation of essays with the Intelligent Essay Assessor. In M. D. Shermis & J. Burstein (Eds.), *Automated essay scoring: A cross-disciplinary perspective* (pp. 87–112). Mahwah, NJ: Lawrence Erlbaum Associates, Inc.

LaVoie, N., Streeter, L., Lochbaum, K., Wroblewski, D., Boyce, L.A., Krupnick, C., & Psotka, J. (2010). Automating expertise in collaborative learning environments. *Journal of Asynchronous Learning Networks, 14*(4), 97–119.

Lochbaum, K., Psotka, J., & Streeter, L. (2002). *Harnessing the power of peers*. Paper presented at the Interservice/Industry, Simulation and Education Conference (I/ITSEC), Orlando, FL, December.

Rehder, B., Schreiner, M. E., Wolfe, M. B, Laham, D., Landauer, T. K, & Kintsch, W. (1998). Using Latent Semantic Analysis to assess knowledge: Some technical considerations. *Discourse Processes, 25*, 337–354.

Streeter, L., Lochbaum, K., LaVoie, N., & Psotka, J. (2007). Automated tools for collaborative learning environments. In D. McNamara, T. Landauer, S. Dennis, & W. Kintsch (Eds.), *Latent Semantic Analysis: A road to meaning*. Mahwah, NJ: Lawrence Erlbaum.

Swygert, K., Margolis, M., King, A., Siftar, T., Clyman, S., Hawkins, R., and Clauser, B. (2003). Evaluation of an automated procedure for scoring patient notes as part of a clinical skills examination. *Academic Medicine, 78*, 10, S75–S77.

Thurlow, M. M., Hermann, A., & Foltz, P. W. (2010). *Preparing MSA science items for artificial*

intelligence scoring. Paper presented at the Maryland Assessment Group Conference, Ocean City, MD, November.

Thurlow, M. M., Hermann, A., & Lochbaum, K. E. (2011). *Preparing MSA science items for artificial intelligence scoring*. Paper presented at the Assessment Group Conference, Ocean City, MD, November.

Williamson, D. M., Bennett, R., Lazer, S., Bernstein, J., Foltz, P. W., Landauer, T. K., Sweeney, K. (2010). Automated scoring for the assessment of Common Core Standards. doi: http://www.ets.org/s/commonassessments/pdf/AutomatedScoringAssessCommonCoreStandards.pdf