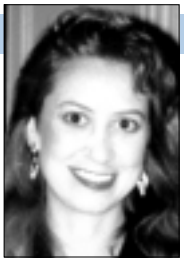


Marti A. Hearst
University of California, Berkeley
hearst@sims.berkeley.edu

The debate on automated essay grading

In this installment of Trends & Controversies, we look at a controversy: the use of computers for automated and semiautomated grading of exams. I am very pleased to have a well-rounded discussion of the topic, because in addition to three technical contributions, we have a commentary from Robert Calfee, the Dean of the School of Education at UC Riverside and an expert in the field of educational testing.

First, Karen Kukich, the director of the Natural Language Processing group at Educational Testing Service, provides us with an insider's view of the history of the field of automated essay grading and describes how ETS is currently using computer programs to supplement human judges in the grading process. Then, Tom Landauer, Darrell Laham, and Peter Foltz describe the use of Latent Semantic Analysis in a commercial essay-scoring system called IEA. They also address important ethical questions. Lynette Hirschman, Eric Breck, John Burger, and Lisa Ferro report on MITRE's current efforts towards automated grading of short-answer questions and discuss the ramifications for the design of general question-answering systems. Finally, Robert Calfee places these developments in the framework of current educational theory and practice.

After three years editing the Trends & Controversies feature, it is time for me to pass the column on to others. Thank you for reading, and I hope our trendy and controversial contributors' essays have enhanced your understanding of the shape of the field.

— Marti Hearst

Beyond Automated Essay Scoring

Karen Kukich, Educational Testing Service

The ability to communicate in natural language has long been considered a defining characteristic of human intelligence. Furthermore, we hold our ability to express ideas in writing as a pinnacle of this uniquely human language facility—it defies formulaic or algorithmic specification. So it comes as no surprise that attempts to devise computer programs that evaluate writing are often met with resounding skepticism. Nevertheless, automated writing-evaluation systems might provide precisely the platforms we need to elucidate many of the features that characterize good and bad writing, and many of the linguistic, cognitive, and other skills that underlie the human capacity for both reading and writing.

Using computers to increase our understanding of the textual features and cognitive skills involved in creating and comprehending written text will have clear benefits. It will help us develop more effective instructional materials for improving reading, writing, and other human communication abilities. It will also help us develop more effective technologies, such as search engines and question-answering systems, for providing universal access to electronic information.

A sketch of the brief history of automated writing-evaluation research and its future directions might lend some credence to this argument.

Pioneering research

Ellis Page set the stage for automated writing evaluation (see the timeline in Figure 1).[1] Recognizing the heavy demand placed on teachers and large-scale testing programs in evaluating student essays, Page developed an automated essay-grading system called Project Essay Grader. He started with a set of student essays that teachers had already graded. He then experimented with a variety of automatically extractable textual features and applied multiple linear regression to determine an optimal combination of weighted features that best predicted the teachers' grades. His system could then score other essays using the same set of weighted features. PEG's scores showed a multiple *R* correlation with teachers' scores of .78—almost as strong as the .85 correlation between two or more teachers.

In the 1960s, the kinds of features we could automatically extract from text were limited to surface features. Some of the most predictive features Page found included average word length, essay length in words, number of commas, number of prepositions, and number of uncommon words—the latter being negatively correlated with essay scores. Page called these features proxies for some intrinsic qualities of writing competence. He had to use indirect measures because of the computational difficulty of implementing more direct measures.

Despite its impressive success at predicting teachers' essay ratings, the early version of PEG received only limited acceptance in the writing and education community, precisely because it used indirect measures of writing skill. Critics argued that using indirect measures left the system vulnerable to cheating, because students could artificially enhance their scores using tricks—they could simply write a longer essay, for example. Another, more important, criticism was that because indirect measures did not capture important qualities of writing such as content, organization, and style, they couldn't provide instructional feedback to students. Although the general approach—identifying

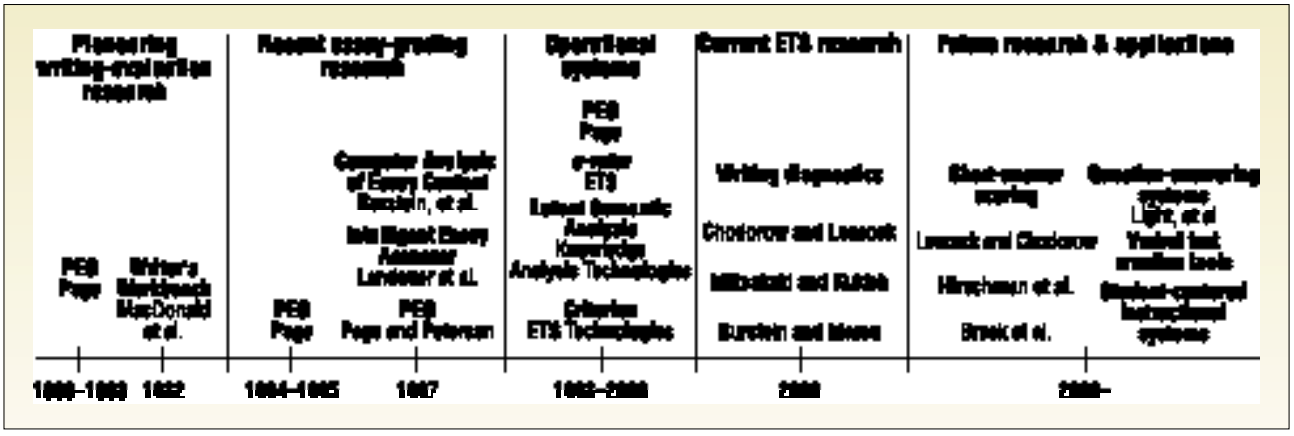


Figure 1. A timeline of research developments in writing evaluation. (This timeline is not comprehensive. This article focuses mainly on research and development at Educational Testing Service.)

textual features correlated with good writing—was sound, a significant research challenge remained: identifying and automatically extracting more direct measures of writing quality.

In the early 1980s, the Writer's Workbench tool set took a first step toward this goal.[2] WWB was not an essay-scoring system. Instead, it aimed to provide helpful feedback to writers about spelling, diction, and readability. In addition to its spelling program—one of the first spelling checkers—WWB included a diction program that automatically flagged commonly misused and pretentious words, such as *irregardless* and *utilize*. It also included programs for computing some standard readability measures based on word, syllable, and sentence counts, so in the process it flagged lengthy sentences as potentially problematic. Although WWB programs barely scratched the surface of text, they were a step in the right direction for the automated analysis of writing quality.

Recent research

By the 1990s, progress in the fields of natural-language processing and information retrieval encouraged researchers to apply new computational tools and techniques to the challenge of automatically extracting from essays more direct measures of writing quality.

Finding more direct measures. Essay-scoring guidelines for the Analytical Writing Assessment portion of the Graduate Management Admissions Test specify a set of general qualities of writing to evaluate (see www.gmat.org). Examples include *syntactic variety, topic content,* and *organization of ideas.* A team of ETS researchers, led by Jill Burstein, hypothesized a set of linguistic features that might more directly measure these general qualities—features they could automatically extract from essays using NLP and IR techniques.

For example, the ETS researchers could measure syntactic variety using features that quantify types of sentences and clauses found in essays, and they could approximate values for these features using syntactic processing tools available in the NLP community. They could measure topic content using vocabulary content analyses, deriving values for these features using vector space modeling techniques now common in IR. They used these techniques to compute similarity measures between documents based on weighted frequencies of vocabulary terms occurring in documents.

However, the researchers needed more sophisticated techniques to identify the essays' individual arguments and to evaluate their rhetorical structure. So, they devised a technique for approximating values for these features by first partitioning an essay into individual arguments using NLP techniques based on the identification of specific lexical and syntactic cues. They then applied vocabulary content analysis to each argument.

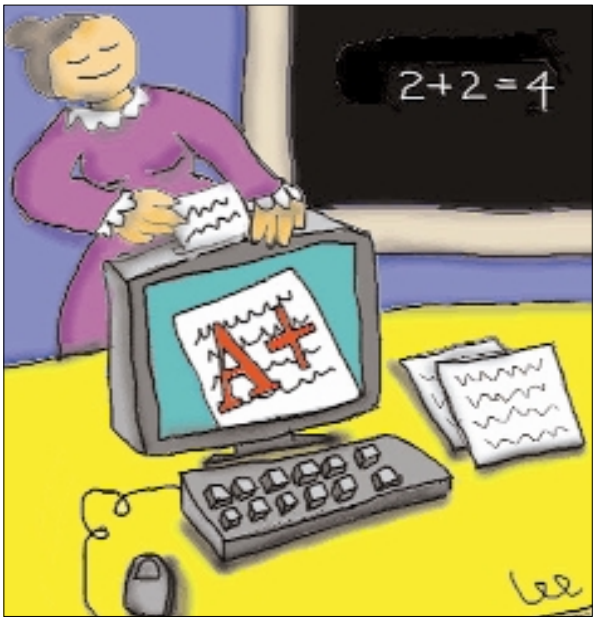


Illustration by Sally Lee

The *e-rater* prototype. A pilot version of the computerized GMAT Analytical Writing Assessment provided the data for a series of preliminary automated essay scoring studies. The AWA requires each student to write two essays, one to analyze an argument presented in a short text and another to express an opinion on a specific issue presented in a brief statement. Preliminary studies began with two essay sets, one for each essay type. Each set contained over 400 essays, and all the essays in each set addressed the same topic. Two writing experts using the GMAT guidelines scored each essay on a six-point scale. If the scores they assigned differed by more than one point, which happened in approximately 10% of the cases, a third expert reader resolved the discrepancy.

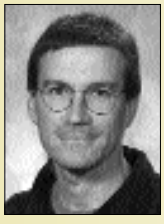
ETS researchers defined more

Karen Kukich is a principal research scientist and Director of the Natural Language Processing Research Group at Educational Testing Service in Princeton, N.J. Highlights of her career in NLP research and development include creating the first fully automated natural language report generation system, publishing an award-winning article on automated spelling correction, deploying a documentation generator for telephone network planning, and spearheading the operational *e-rater* team. She holds a BS in mathematics and philosophy and an MS and PhD in information science, all from the University of Pittsburgh. Contact her at kkukich@ets.org.

Lynette Hirschman is the chief scientist at MITRE's Information Technology Division in Bedford, Mass. Her research interests include natural language processing for both written and spoken language, as well as conversation interaction. She has had a long-standing interest in the evaluation of natural-language systems and has written extensively on the subject. She is responsible for the Human Language Technology research at MITRE and is leading a long-term research project focused on building an automated system that can take and pass elementary school reading comprehension tests. She received her PhD from the University of Pennsylvania in formal linguistics. Contact her at the MITRE Corp., Bedford, MA; lynette@mitre.org.

Eric Breck is a senior artificial intelligence engineer at MITRE's Information Technology Center in Bedford, Mass. At MITRE, he has worked extensively on automated scoring methods for question answering and reading comprehension. He also is interested in reverse engineering. He received his BS in linguistics and mathematics from the University of Michigan at Ann Arbor. Contact him at the MITRE Corp., Bedford, MA; ebreck@mitre.org.

Marc Light is a lead scientist at the MITRE Corporation. He leads MITRE's research in question-answering, creating a search engine that can respond to factual questions with concise answers instead of documents. He also works on MITRE's reading comprehension research and led the Johns Hopkins Summer Workshop on Reading Comprehension during July and August 2000. His research focuses on empirically driven approaches to human-language processing, specifically those involving statistical methods. He received a PhD in computer science from the University of Rochester. Contact him at the MITRE Corp., Bedford, MA; light@mitre.org.



John Burger is a lead scientist at the MITRE Corporation. He has been involved in MITRE's natural language processing research since 1988. His research interests include the application of statistical approaches to language processing and information retrieval, as well as machine learning in general. He received a BS in mathematics and computer science from Carnegie-Mellon University. Contact him at the MITRE Corp., Bedford, MA; john@mitre.org.

Lisa Ferro is a senior artificial intelligence engineer at MITRE's Information Technology Center in Bedford, Mass. She has recently been focusing on producing resources for the evaluation of natural language systems and for the development of corpus-based systems. She received her PhD in linguistics from the University of Connecticut. Contact her at the MITRE Corp., Bedford, MA; lferro@mitre.org.

than 100 automatically extractable essay features—including the linguistic features mentioned earlier and a variety of proxy features. Then, they implemented computer algorithms to extract values for every feature from each essay. For both essay topics, they subjected various subsets of features to stepwise linear regression to determine optimal scoring models, or sets of weighted features, predictive of the scores the experts assigned.

They then tested each scoring model on an additional set of essays written on one of the same two topics. They extracted model-relevant features from the new essays and summed the weighted feature values for each essay to predict the score the writing experts assigned to that essay. Many models built in this manner achieved excellent results. The scores they assigned had the same level agreement as the two writing experts—that is, they agreed approximately 90% of the time. The most important result was that models consisting of mainly linguistic features worked as well as those containing only proxy features, thus providing evidence that we could automatically score essays using more direct measures of writing quality.

ETS patented the resulting automated essay scoring system, CAEC (Computer Analysis of Essay Content). Subsequent studies refined the linguistic features and their algorithms and tested the system on numerous other essay sets, each addressing a different topic. These studies demonstrated that the automated scoring technique, renamed *e-rater*, generalized across essay topics.[3] Since then, research has confirmed the psychometric validity of *e-rater* scores, in terms of external measures of students' writing abilities, cultural and second-language differences, and subject-specific applications such as advanced-placement tests in US history and English literature. (Visit www.ets.org/research/erater.html for more information on *e-rater* studies.)

PEG. Meanwhile, the PEG system was also undergoing transformations to include more direct measures of writing quality. In 1995, Page reported that PEG's "current programs explore complex and rich variables, such as searching each sentence for soundness of structure and weighing these ratings across the essay."[4] Other publications also discuss PEG's performance, although the specific features it now measures remain undisclosed.

The Intelligent Essay Assessor. At the same time, Tom Landauer and his colleagues were developing another approach to more direct measures of writing quality, called Latent Semantic Analysis. LSA aims at going beneath the essay's surface vocabulary to quantify its deeper semantic content.[5] Its main advantage is that it captures transitivity relations and collocation effects among vocabulary terms, thereby letting it accurately judge the semantic relatedness of two

Thomas K. Landauer is a professor of psychology and a fellow of the Institute of Cognitive Science at the University of Colorado. He is also the president of Knowledge Analysis Technologies, a small company doing R&D and applications of computational simulations of human cognition. His research is devoted to developing, testing, and applying mathematical models of learning and language. He received his PhD from Harvard. Contact him at landauer@psych.colorado.edu.

Darrell Laham is a cofounder and the CTO of Knowledge Analysis Technologies. He leads several ongoing LSA-based R&D projects on technologies for education, personnel management, and cooperative learning and decision-making. He holds an MA in educational measurement and a PhD in cognitive science from the University of Colorado. Contact him at Knowledge Analysis Technologies, 4001 Discovery Dr., Ste. 2110, Boulder, CO, 80303; dlaham@knowledge-technologies.com.

Peter W. Foltz is an associate professor at New Mexico State University and also a cofounder of Knowledge Analysis Technologies. His research interests are in computational models of text and discourse, human memory, and techniques for information retrieval and filtering. He has a PhD in cognitive psychology from the University of Colorado. Contact him at the Dept. of psychology, New Mexico State Univ., Box 30001, Las Cruces, NM, 88003; pfoltz@crl.nmsu.edu.

Robert Calfee is professor and the dean of the School of Education at the University of California, Riverside. He is a cognitive psychologist with research interests in the effects of schooling on the intellectual potential of individuals and groups. He earned his degrees at UCLA. Contact him at robert.calfee@ucr.edu.

documents regardless of their vocabulary overlap (see "The Intelligent Essay Assessor" on page 27).

Landauer and his colleagues recognized that LSA could provide a novel contribution to writing evaluation applications. They carried out research and development to test LSA's potential to score essays, evaluate summaries students wrote, and even evaluate college students' classroom writing assignments.[6,7] Their work culminated in the development of the Intelligent Essay Assessor system. They report essay-scoring accuracy similar to *e-rater* and PEG using IEA's measures of semantic quality and quantity, providing additional evidence that we can automatically derive more direct measures of writing quality.

Operational writing-evaluation systems

The availability of more direct and defensible measures of writing quality, along with a growing need for grading assistance for teachers and large-scale testing programs, ultimately opened minds and doors to the feasibility of automated writing evaluation. In the late 1990s, several automated writing-evaluation systems, including *e-rater*, PEG, and IEA, made the transition from research prototypes into fully operational systems.

In February 1999, *e-rater* became fully operational within ETS's Online Scoring Network for scoring GMAT essays. Each time ETS test developers introduce a new essay topic, OSN sends examinees' essays to two or more ETS writing experts to be scored in the usual manner. Once a sufficiently large sample of expertly scored essays accumulates, OSN invokes *e-rater*'s automated model builder to create and cross-validate a scoring model for that essay topic. Thus, new *e-rater* scoring models are certified in the same way new writing experts are certified. Once certified, a new *e-rater* scoring model automatically becomes one of the first two "experts" to score subsequent essays on that topic. None of OSN's mechanics change with the introduction of *e-rater*. All essays still receive at least two readings and require a third human expert to resolve scores that differ by more than one point. With almost half a million GMAT essays being scored each year, using *e-rater* clearly relieves a significant portion of the load on human scoring experts.

For high-stakes assessments, such as the GMAT exam, at least one human is always in the scoring loop. This safeguard helps prevent any radically creative or otherwise anomalous essays from slipping through the system unnoticed.

For low-stakes writing-evaluation applications, such as a Web-based practice essay system, a single reading by an automated system is often acceptable and economically preferable. For this purpose, ETS Technologies, a new subsidiary of ETS, has developed a fully automated service called *Criterion*. (See www.etstechnologies.com for more information about ETS Technologies products and services.) This service incorporates a set of safeguards for detecting off-topic and statistically anomalous essays. In addition, research is currently underway to enhance *Criterion* with additional writing-evaluation features that will provide students and teachers not only with holistic scores but also with diagnostic feedback about the specific strengths and weaknesses of the essays. Producing such feedback requires more basic research to identify even deeper, more direct measures of writing quality. Fortunately, ETS and other NLP researchers are now well positioned to pursue this challenge.

Current ETS writing research

Clearly, we've made progress toward identifying and automatically extracting more direct measures of writing quality. Research on *e-rater*, PEG, and IEA has identified automatically extractable semantic, syntactic, and rhetorical structure features that correlate with writing quality. These features are all measured holistically—that is, in terms of statistical averages over the whole essay text. But holistic scores do not tell the whole story.[8] A student who receives a low score wants to know precisely where specific problems occurred in the essay. To give students and teachers useful feedback, automated sys-

tems must identify and extract finer-grained features of writing. This is a much greater research challenge, but three recent ETS studies have already made progress toward this goal.

One study demonstrated the feasibility of a novel technique for detecting *lexical-grammatical errors* in essays,[9] including word-specific usage errors such as "pollutions" or "knowledge at math," as well as general grammar violations such as "I concentrates" or "this conclusions." This technique, called ALEK (*a*ssessment of *le*xical *k*nowledge), employs statistical models of probabilities of occurrences of word and part-of-speech bigrams and trigrams to detect unexpected words and word sequences such as the usage errors noted in the previous sentence (bigrams and trigrams are two-word and three-word sequences, such as "in the" and "in the beginning"). In a study that focused on 20 words, 79% of the usages that ALEK flagged were errors. However, a human reader detected many more errors than ALEK did (ETS researchers are working to address this problem). Furthermore, the total number of lexical-grammatical errors ALEK detected showed an inverse correlation with essay scores, indicating we could employ this feature as a more direct measure of writing quality and use it to provide explicit diagnostic feedback to the essay writer.

Another study demonstrated the feasibility of using a current linguistic theory called Centering Theory to detect *rough shifts* in topic within essays.[10] Centering Theory posits four types of transitions between sentences, ranging from easy to difficult (rough), based on the salience of entities referred to in succeeding sentences. The syntactic role an entity plays in a sentence—for example, subject, indirect object, direct object, and so forth—determines salience. A study of 100 GMAT essays showed that the ratio of rough shifts detected in essays was inversely correlated with essay scores. So, not only could we employ a rough shift feature as a component in scoring models to measure incoherence in essays, we could also use it to direct essay writers to specific sentences that need improvement.

Yet another study focused on using automatically generated summaries to improve essay-scoring performance and to provide feedback.[11] This study generated summaries based on the essays' rhetorical relations—implicit relations between sentences or clauses such as cause, contrast, or elaboration—found in essays. Rhetorical relations are sometimes cued by transition words such as because, however, furthermore, and so forth. This study showed that using a rhetorical-structure-based summarizer to extract an essay's salient content could not only enhance a scoring model's performance but also point essay writers directly to their salient content—or inform them of the lack thereof.

Lexical-grammatical errors, rough shifts, and rhetorical relations are just three examples of finer-grained measures of writing quality that have proven to be statistically correlated with essay scores. However, we need further basic NLP research before any of these measures become operational. Although fully automated techniques are available for detecting lexical-grammatical errors and rhetorical relations, we need research to improve their accuracy. Rough-shift detection is partially automated; coreference resolution is the big challenge. We also need to identify other text features and cognitive skills correlated with writing quality. A parallel route to identifying these features and skills is through reading-comprehension research.

To the uninitiated, it might seem counter intuitive that scoring short answers poses a greater challenge than scoring essays.

Future research and applications

Many of the features that affect writing quality also affect ease of text comprehension. For example, lexical-grammatical errors, rough shifts, and inappropriate cue markers for rhetorical relations will likely increase the difficulty of understanding an essay or any text passage. Researchers studying reading comprehension have suggested additional features, such as ease of locating antecedents of pronouns, use of literal versus abstract or metaphorical language, use of infrequent word senses, and ease of identifying topic chains.

One way to determine whether features such as these play a role in writing quality is to determine the role they play in students' abilities to answer questions about written passages. Fortunately, large databases of question-difficulty statistics based on student responses on reading comprehension exams are available. Just as we can extract features from essays and submit them to statistical analysis to determine which ones are most predictive of essay scores, we can also extract features from reading-comprehension passages and submit them to statistical analysis to determine which ones are most predictive of question difficulty. Some research studies using linear-regression techniques[12] and tree-based-regression techniques[13,14] have already demonstrated that features such as those mentioned earlier are predictive of question difficulty. So, additional NLP research into developing tools for automatically evaluating question difficulty is likely to apply equally to evaluating writing quality.

Currently, the NLP research community has expressed much interest in the challenge of automated question answering.[15] People would like to be able to submit a question to a Web-based search engine and receive a short-answer instead of an extensive list of more or less relevant documents. Unfortunately, automated question answering poses at least as great a challenge as short-answer scoring. Furthermore, as MITRE NLP researchers point out in this issue (see the essay "Automated Grading of Short-Answer Tests" on page 31) and elsewhere,[16] we need short-answer scoring systems to manage the task of evaluating automated question-answering systems.

To the uninitiated, it might seem counter intuitive that scoring short answers poses a greater challenge than scoring essays. But it should be clear from the preceding sections that automated essay-scoring techniques can rely on statistical averages of general features such as overall vocabulary content and syntactic variety to derive evidence of writing quality. In contrast, short answers seem to provide little textual evidence of the writer's underlying meaning, hence the need for finer-grained measures and deeper analysis.

Researchers at ETS Technologies have been exploring techniques for scoring students' short-answer responses to end-of-chapter textbook questions. They have found this seemingly simple task requires a great deal of NLP power. Pronoun and other coreference resolution tools are essential because anaphora abounds in students' free-

form responses. In addition to general lexical databases, we need subject-specific thesauri. We also need special tokenizing and tagging techniques to derive parts of speech and partial parses from incomplete sentences and clauses. And all this computational machinery must derive some approximation of underlying predicate-argument or prepositional structure to ultimately make a judgment about how much a short answer's text represents the target concepts that constitute a "correct answer."

So, while the research challenges remain great, the benefits of using computers to increase our understanding of features and processes involved in creating and comprehending written text seem clearly worthwhile. The abilities to devise student-centered instructional systems for reading and writing, more effective search engines and question-answering systems, and universal access to electronic information will be just the beginning.

Acknowledgments

Contributors to this work from the ETS NLP Research Group include Magdalena Wolska, Daniel Zuckerman, Ramin Hemat, Dennis Quardt, Slava Andreyev and Lisa Hemat; consultant Krishna Jha; graduate students Eleni Miltsakaki, Krishna Prasad, Konrad Szczesniak, Hoa Trang Dang, and Tom Morton; and former group members Susanne Wolff and Probal Tahbildar.

Contributors to this work from the ETS Technologies NLP Research Group include Jill Burstein, Claudia Leacock, Chi Lu, and Eleanor Bolge; consultant Martin Chodorow; and summer student Irek Szczesniak.

References

1. E.B. Page, "The Use of the Computer in Analyzing Student Essays," *Int'l Rev. Education*, Vol. 14, 1968, pp. 210–225.
2. N. MacDonald et al., "The Writer's Workbench: Computer Aids for Text Analysis," *IEEE Trans. Comm.*, Vol. COM-30, No. 1, 1982, pp. 105–110.
3. J. Burstein et al., "Automated Scoring Using A Hybrid Feature Identification Technique," P*roc. Ann. Meeting Association of Computational Linguistics*, Montreal, Canada, 1998; www.ets.org/research/erater.html (current Nov. 2000).
4. E.B. Page and N.S. Petersen, "The Computer Moves into Essay Grading: Updating the Ancient Test," *Phi Delta Kappan*, Vol. 76, 1995, pp. 561–565.
5. T. Landauer, and S. Dumais, "A Solution to Plato's Problem: The Latent Semantic Analysis Theory of Acquisition, Induction, and Representation of Knowledge," *Psychological Rev.*, Vol. 104, 1997, pp. 211–240.
6. P. Foltz, W. Kintsch, and T. Landauer, "The Measurement of Textual Coherence with Latent Semantic Analysis," *Discourse Processes*, Vol. 25, 1998, pp. 285–308.
7. D. Laham, *Automated Content Assessment of Text using Latent Semantic Analysis to Simulate Human Cognition*, PhD Dissertation, University of Colorado, Boulder, Colo., 2000.
8. R.E. Bennett and I.I. Bejar, "Validity and Automated Scoring: It's Not Only the Scoring," *Educational Measurement: Issues and Practice*, Vol. 17, No. 4, 1998. pp. 9–17.
9. M. Chodorow and C. Leacock, "An Unsupervised Method for Detecting Grammatical Errors," P*roc. First Meeting of the North American Chapter of the Association for Computational Linguistics* (ANLP-NAACL-2000), Morgan Kaufmann, San Francisco, 2000, pp. 140–147.
10. E. Miltsakaki and K. Kukich, "Automated Evaluation of Coherence in Student Essays," *Proc. LREC-2000, Linguistic Resources in Education Conf.*, Athens, Greece, 2000; www.ling.upenn.edu/~elenimi/grad.html (current Nov. 2000).
11. J. Burstein and D. Marcu, "Toward Using Text Summarization for Essay-Based Feedback," P*roc. TALN 2000 Conf.*, Lausanne, Switzerland, 2000; www.isi.edu/~marcu/papers.html (current Nov. 2000).
12. R. Freedle and I. Kostin,"The Prediction of TOEFL Reading Item Difficulty: Implications for Construct Validity," *Language Testing*, Vol. 10, 1993, pp. 133–170.
13. K.M. Sheehan, "A Tree-Based Approach to Proficiency Scaling and Diagnostic Assessment," *J. Educational Measurement*, Vol. 34, 1997, pp. 333–352.
14. K. Sheehan, A. Ginther, and M. Schedl, *Development of a Proficiency Scale for the TOEFL Reading Comprehension Section, TOEFL Research Report*, ETS, Princeton, NJ, 1999.
15. M. Light (Ed.), *Proc. Workshop on Reading Comprehension Tests as Evaluation for Computer-Based Language Understanding Systems*, Association for Computational Linguistics, New Brunswick, 2000.
16. E.J. Breck et al., "How to Evaluate Your Question Answering System Every Day … and Still Get Real Work Done," *Proc. LREC-2000, Linguistic Resources in Education Conf.*, Athens, Greece, 2000.

The Intelligent Essay Assessor

Thomas K. Landauer and Darrell Laham, University of Colorado and Knowledge Analysis Technologies
Peter W. Foltz, New Mexico State University and Knowledge Analysis Technologies

There is a widespread belief that most students have inadequate studying, critical thinking, and writing skills. Two likely causes of these deficiencies are an overreliance on multiple-choice testing and too few opportunities to assess verbalized knowledge. Although textbooks have long supplemented face-to-face student–teacher interaction with independent learning for *acquiring* knowledge, students might also profit from tools for independent learning for *expressing* knowledge. One such tool might be an intelligent system that quickly and consistently gives useful feedback on freely expressed knowledge.

The Intelligent Essay Assessor's core technology

The Intelligent Essay Assessor (IEA), an essay-analysis, scoring, and tutorial-feedback system, is one of many current and potential applications of Latent Semantic Analysis. LSA is a machine-learning technology for simulating the meaning of words and passages.[1–4] The fundamental idea is that the aggregate of all the contexts in which words appear provides an enormous system of simultaneous equations that determines the similarity of meaning of words and passages to each other. LSA uses the matrix algebra technique of singular value decomposition to analyze a corpus of ordinary text of the same size and content as that from which students learn the vocabulary, concepts, and knowledge needed to write an expository essay. LSA represents every word and passage as a point in a high-dimensional *semantic space*. Relative position in the space estimates the similarity of meaning between any two words or passages. Simulations of many linguistic, psycholinguistic, and learning phenomena, as well as several other educational and personnel applications, show that LSA closely reflects corresponding similarities of meaning for humans.[5–7] As measured by simulations of human performance on standardized vocabulary and domain-knowledge multiple-choice tests, LSA is always significantly

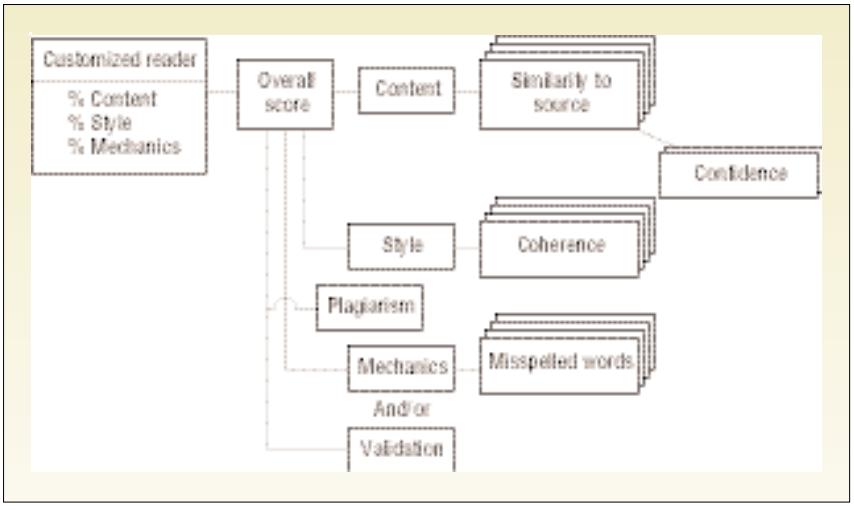


Figure 2. Schematic illustration of the computation of an Intelligent Essay Assessor score based on a customized combination of its three main components plus accessory measures.

more accurate—sometimes by factors of three or more times—than traditional key word approaches that rely on the occurrence of the same words or word stems in two passages.

LSA is the basis of IEA's assessment and tutorial feedback concerning the knowledge content of essays. It deals effectively with the fact that there are an unlimited number of ways to express nearly the same meaning in different words. LSA also lets IEA base its scores primarily on the opinions of human experts about similar essays, rather than relying on specific key words and other index variables that correlate with human scores on other essays.

Using LSA, IEA always keeps knowledge content the dominant factor in its scores. Because expressing knowledge well requires good writing, graders cannot completely isolate the content of an essay from its stylistic and mechanical qualities. However, making content primary has favorable consequences for face validity, immunity to coaching and counterfeiting, utility for diagnosis and advice at a conceptual level, and the potential to encourage valuable study and thought. IEA measures the content, style, and mechanics components separately, and whenever possible computes each component in the same way, so that score interpretation is comparable across applications. Because LSA is based on machine learning from ordinary text rather than, for example, from coding of language-dependent rules, we can automatically apply IEA's content measures with nearly equal facility for any language, including ones that don't use the Latin alphabet.

How IEA works

The user first trains IEA on a corpus of domain-representative text (for example, when scoring biology essays, a biology textbook, or when scoring creative narrative essays, a sample representing the lifetime reading of a typical test-taker). LSA characterizes student essays by representing their meaning and compares them with highly similar texts of known quality. It adds corpus-statistical writing-style and mechanics measures to help determine overall scoring, validate an essay as appropriate English (or other language), detect plagiarism or attempts to fool the system, and provide tutorial feedback.

IEA computes and combines the three major components (content, style, and mechanics), plus two or more accessory measures, as illustrated in Figure 2. For customized application, it can adjust the rule for combining the components within defined limits. The default application is constrained multiple regression on human scores in a training sample. IEA always computes self-validation, confidence, and counterfeiting measures, such as the plagiarism detection measure depicted in Figure 2.

The main technical difference between IEA and other approaches is this: Other systems work primarily by finding essay features they can count and that correlate with ratings human graders have assigned. They determine a formula for choosing and combining the variables that produces the best results on the training data. They then apply this formula to every to-be-scored essay. What principally distinguishes IEA is its LSA-based direct use of evaluations by human experts of essays that are very similar in semantic content. This method, called *vicarious human scoring*, lets the implicit criteria for each individual essay differ. Thus, different students can focus on different aspects of a question, using different words and styles, and get the same score if expert opinions so dictate.

IEA in use. The Web-based version of IEA supplies instantaneous evaluations and, when implemented, tutorial advice. As reviewed later, IEA's overall scores are as reliable as that of a teacher or professional essay reader. The detail in its comments and suggestions is potentially unlimited, although not necessarily the same as what a human tutor would provide. For example, it can tell students what important content is missing from their papers and point them to relevant sources in their textbooks. It can also identify irrelevant and redundant sentences and report on conceptual coherence and other organizational qualities. However, as yet it cannot tell students whether they have made a specific point logically or persuasively, or that an independent clause should have had been set off by a comma, and so forth.

Scoring calibration. IEA's use of LSA-based training on background text lets it analyze and score essays with few or even no prescored examples. Our research shows that IEA can usually be optimally calibrated with 100 prescored essays, and sometimes with as few as 20. However, by extending the LSA technology, IEA can train itself to give accurate rankings without using any human grades. Here it uses the varying knowledge the essays express themselves to align them on a continuum of quality. To score in this way, IEA typically needs 200 or more student essays on the same well-defined topic.

IEA can also be calibrated by comparing essays either to ideal answers or to sections of a textbook that students should have studied. These are the least desirable approaches, because students don't usually write answers similar to those of professors or authors and often write equally good answers in many different ways.

In its tutorial feedback implementations, IEA is typically incorporated into online courseware accompanying a textbook. Students write essays or summaries of sections or chapters. IEA provides immediate evaluation and points the student back to pages or

sections of the text containing information that should have been used in the essay. The system also identifies irrelevant or redundant sentences and overall conceptual coherence. A well-controlled experiment has shown that the Summary Street version for middle schools produces significant improvement in relevant skills.[8,9] Another experimental application has used IEA as an embedded assessment measure for online discussion-based learning environments, where it cumulatively measures the knowledge contained in individual students' contributions.

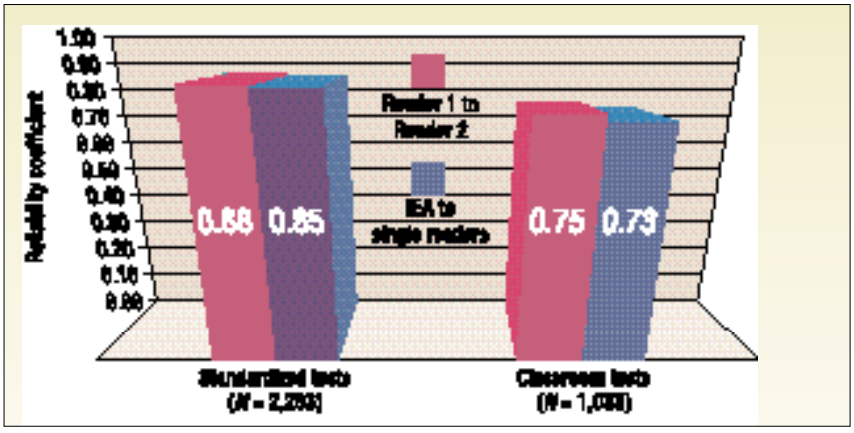


Figure 3. Average Latent Semantic Analysis-based Intelligent Essay Assessor results: Summary of reliability results for 3,296 essays on 15 diverse topics. The measure (reliability coefficient) is the correlation between two human experts (left bars) or between Intelligent Essay Assessor scores and one human grader, averaged over the two humans (right bars).

Reliability and validity of IEA

The standard way to evaluate the accuracy of a set of essay grades is to measure how well two independent scorings agree with each other—either scores two human judges gave or one by an automatic grader and others by humans. This criterion is not objective or absolute because human judges have legitimate differences of opinion. It is best interpreted simply as an estimate of how well the score in question will predict additional human opinions.

There is no reason why an automatic grader cannot predict a human score better than another human score does. One way is for the automatic grader to be more consistent in evaluating some factor. However, we need to make sure that the factor the machine is measuring better is one we want to stress. Otherwise, relying on this form of validation might lead students to focus on the wrong things—for example, simply using more rare, "trigger" or topic-specific words.

There are other possible criteria for essay-scoring accuracy, such as correlating scores with other measures of knowledge or better agreement with more expert judges. We will report on such measures later. First we review studies of IEA's reliability versus human graders as compared to the reliability between human graders. For validation, we prefer to deal with correlations between the continuous IEA scores and whatever scores the human graders use. This method gives an unbiased estimate of how well one predicts the other, while avoiding the complication of classifying the scores into discrete "grade" or "score" groups, a matter that involves instruction and policy issues largely irrelevant to validity. (However, when desired, IEA statistically predicts the discrete human classifications.)

In each case, we collected numerous essays students wrote to the same prompt in a real examination. Either large national or international professional testing organizations, such as ETS and CRESST, or professors at major universities provided these prompts. At least two graders independently graded each essay. These graders were knowledgeable in the test's content domain and quality criteria and trained in its scoring according to either holistic rubrics or analytic components. They were blind to the IEA scores and, in the case of professional scoring, uninformed that an automatic scoring system would be used. The student groups taking the tests included sixth graders, high school and college students, graduate students in psychology, medical school students, and applicants to graduate business management programs. The 15 different topics included heart and circulatory anatomy and physiology (the same prompt at all student levels in various studies), neural conduction, Pavlovian and operant conditioning, aphasia, attachment in children, Freudian concepts, the history of the Great Depression, the history of the Panama Canal, ancient American civilizations, alternative energy sources, business and marketing problems, and a creative narrative composition task in which students were given a two-sentence beginning of a story and asked to complete it. In all cases, the test essays differed from the essays used to train the system.

Figure 3 shows averaged results, and Figures 4 and 5 show scatter plots for two cases—a GMAT "argument" essay and a story-completion narrative.

IEA scores on average correlated with a human score as well as one human score correlated with another. There was some variation across student groups, tests, and human graders. As expected, the more reliable the human graders were, the better IEA predicted their scores.

Using the professional grades, we analyzed the contribution of the three components to overall scores. When combined by linear regression, they predicted human grades with a correlation of .85. Alone, the content, style, and mechanics scores predicted human grades with a correlation of .83, .68 and .66, respectively. When optimally weighted by linear regression, the relative contributions were .75, .13, and .11, respectively.

Other empirical validations of IEA accuracy. In three different ways, IEA essays scores have proven modestly more valid than human essay scores. First, in studies of essays on heart anatomy, LSA's scores predicted short-answer test scores over the same material better than did expert human scores on the same essays. Second, for a set of student essays on neural conduction, three sets of scores were obtained, one from undergraduate teaching assistants, one from graduate-student teaching assistants, and one from the professor. Using the same combined set for training, IEA agreed best with the professor, least with the undergraduates. The results of these two studies were statistically significant. Third, in a study involving GMAT essays, IEA was trained using all the pregraded scores of just one of the two graders for each essay. The IEA score predicted the second grader's score very slightly better than did the scores on which it was trained. This is attributable to IEA's comparison of a to-be-scored essay with all other essays, a kind of vicarious multiple human scoring.

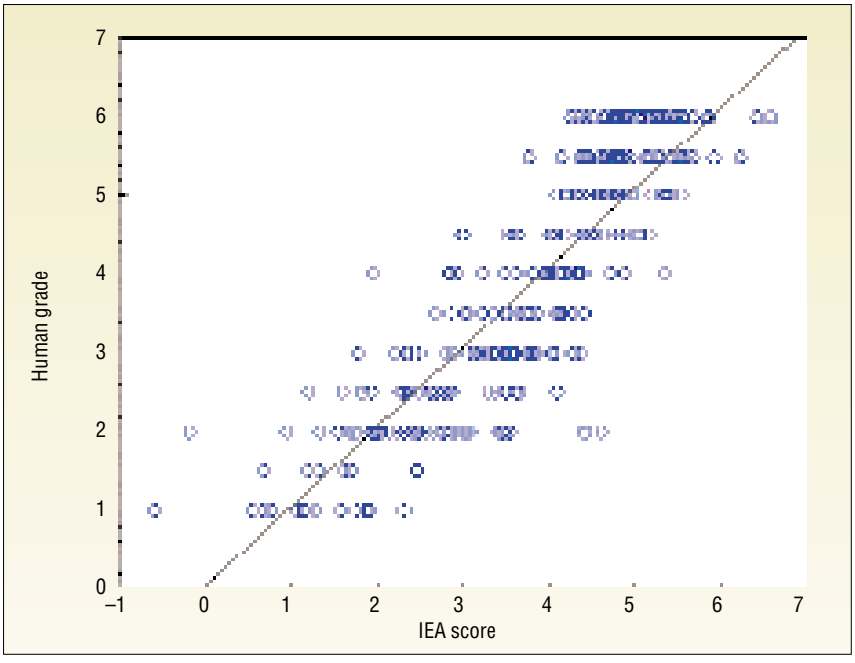


Figure 4. Scatter plot for Intelligent Essay Assessor scoring of two GMAT topics versus independent professional scores from ETS on a sample provided by ETS. On this sample, IEA and the ETS e-rater obtained the same reliabilities versus human readers. All comparisons were blind, and IEA was trained on different essays from those used for the test results shown.

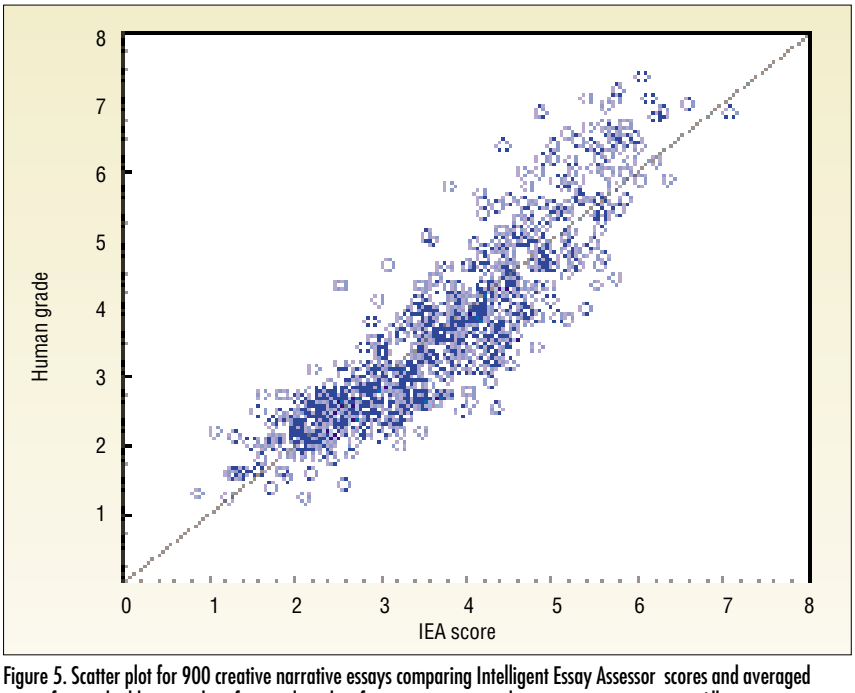


Figure 5. Scatter plot for 900 creative narrative essays comparing Intelligent Essay Assessor scores and averaged scores for two highly trained professional readers from an international assessment organization. All comparisons were blind, and IEA was trained on different essays from the test results shown. The corresponding correlation coefficient is .90, identical to that between two expert human raters.

IEA's internal validity and oddity checks. IEA includes a battery of programs that estimate confidence in the accuracy of the system's score for a particular essay and check that the student wrote the essay in normal English word order. They also ascertain that the student is not trying to fool the system by larding it with rare or topical words; that an essay is not highly unusual, either by being very original or off topic; and that it is not a copy, paraphrase, or rearrangement of another essay. In all such cases, IEA flags the essay for special attention. For example, if comparison essays are insufficiently similar to the to-be-scored essay, or are too variable in their implications for content quality, the essay is passed to a human grader.

We know from empirical user testing that it is very difficult to trick the system into an incorrect grade. As in other systems, it is possible to compose a good essay and then do something to make it abnormal (for example, reordering some of its words or sentences), without incurring much penalty. However, we know of no way to get a high grade from IEA without knowing the material well. As an added precaution, we periodically add or substitute new counterfeit-detection routines that we do not reveal.

IEA responses to some social and philosophical issues

People sometimes worry that computer-based essay grading will fail to credit novel creative answers or answers that reflect greater knowledge than the system was taught. IEA is typically aimed at factual knowledge; we usually don't want highly creative essays on anatomy or jet engine repair. Nonetheless, even with topics in which creativity can be desirable, our experience with IEA has been favorable. For example, in opinion essays on the GMAT, there is ample opportunity for crediting creativity, yet IEA was as reliable as the professional readers. Similarly, as shown in Figure 5, on creative narratives, IEA scores agreed with highly trained expert grader scores as well as the latter agreed with each other. How could this happen? One hypothesis is that a constant setting permits only a limited variety of story themes, plots, and characters—ones that draw upon common knowledge and experience. LSA can capture the similarity of texts that differ only in irrelevant details. For example, IEA treats the theme of "a boy searching for his horse" as very similar to that of "a girl looking for her pony." If good and bad themes are spelled out well or poorly in typical ways, IEA will measure their quality in the same way that it assesses more obviously focused expression of knowledge.

There are other reasons for IEA's success with novel essays. Because it is based on

human judgments of similar essays, the range of performance that the system can measure is unlimited. For example, essays on heart anatomy and function were reliably scored, whether written by sixth graders, college undergraduates, or medical students. Essays better (or worse) than any seen during system training can be scored higher (or lower) than any in the training set. Assume that humanly scored "6" essays are on average highly similar to seven others scored "6" and three scored "5." If IEA encounters a new essay that is highly similar to ten essays scored "6," it might give it a "6.3." In self-calibrated scoring, IEA could, in principle, determine that a particular essay was better than any seen before, by as much as three standard deviations or more. (Of course, any essay that unique is more likely to be way off topic or psychopathic. In any case, it would be flagged.)

We discussed this issue because critics of machine essay grading often assume that computer systems must be slavishly measuring overlap with a finite model and are alarmed by an imagined lack of sensitivity to creativity and genius. On the contrary, it now appears possible that automatic methods can be superior to humans in this regard.

Unique Contributions of IEA

To our knowledge, the Intelligent Essay Assessor is unique among commercially available systems in the following ways:

- It is always based primarily on semantic content, which it measures at a conceptual level rather than by the occurrence of selected words.
- It explicitly embodies holistic human judgments.
- It can be automatically applied to new topics and in new languages without manual construction of new rule sets or the like.
- It can provide useful tutorial commentary on missing content, semantic coherence, redundancy, and irrelevance.
- It detects plagiarism and "system gaming" and computes validity self-checks.
- It has been validated against double expert human scores across a wide variety of different topics and student populations.

This is not to imply that other systems are not capable of equal scoring accuracy—at least some are—or that none are using similar accessory measures. Nor would we claim that IEA is better than others for all purposes. For example, for a composition assignment in which each essay is on a different topic, or in which content is secondary to form, or for evaluating sentence structure, grammar, spelling, or the logic of an argument independent of what it is about, some other methods have greater face validity, and probably greater accuracy and utility.

The future of automatic scoring methods

All present technologies for automatic scoring of essays, including IEA, leave considerable room for improvement. More specifically, methods must evaluate and give critical feedback and suggestions for improvement in detailed matters of logic, syntax, and expression at the sentence level, and of clarity, comprehensibility, and affective qualities (such as humor, suspense, and evocativeness) at sentence, paragraph, and organizational levels. Doing those things will take much better articulated understanding and modeling of human language than we now have.

References

1. M.W. Berry, S.T. Dumais, and G.W. O'Brien, "Using Linear Algebra for Intelligent Information Retrieval," *SIAM Rev.*, Vol. 37, No. 4, 1995, pp. 573–595.
2. S. Deerwester et al., "Indexing By Latent Semantic Analysis," *J. Am. Soc. for Information Science*, Vol. 41, No. 6, 1990, pp. 391–407.
3. T.K. Landauer and S.T. Dumais, "A Solution to Plato's Problem: The Latent Semantic Analysis Theory of the Acquisition, Induction, and Representation of Knowledge," *Psychological Rev.*, No. 104, 1997, pp. 211–240.
4. T.K. Landauer, P.W. Foltz, and D. Laham, "An Introduction to Latent Semantic Analysis," *Discourse Processes*, Vol. 25, Nos. 2–3, 1998, pp. 259–284.
5. B. Rehder et al., "Using Latent Semantic Analysis to Assess Knowledge: Some Technical Considerations, *Discourse Processes*, Vol. 25, Nos. 2–3, 1998, pp. 337–354.
6. M.B. Wolfe et al., "Learning from Text: Matching Readers and Text by Latent Semantic Analysis," *Discourse Processes*, No. 25, 1998, pp. 309–336.
7. P.W. Foltz, W. Kintsch, and T.K. Landauer, "The Measurement of Textual Coherence with Latent Semantic Analysis," *Discourse Processes*, Vol. 25, Nos. 2–3, 1998, pp. 285–308.
8. E. Kintsch, D. Steinhart, and G. Stahl, Interactive Learning Environments, 8XX, 2000.
9. D. Steinhart, *Summary Street: an LSA-Based Intelligent Tutoring System for Writing and Revising Summaries*, unpublished doctoral dissertation, Univ. of Colorado, Boulder, Colo., 2000.

Automated Grading of Short-Answer Tests

Lynette Hirschman, Eric Breck, Marc Light, John D. Burger, and Lisa Ferro
The MITRE Corporation

The educational community and the public have accepted the idea of having a computer grade tests—provided that the tests are structured in such a way that there is no subjective judgment involved, as in grading tests with multiple-choice or yes-or-no answers. It is far more controversial to have computers grading essays, as in the *e-rater* system from Educational Testing Service[1] or—as we discuss here—short-answer tests.

Why short-answer tests? Why automatic evaluation?

Why use short-answer tests instead of multiple choice? First of all, they are more "authentic." Answering real-world questions is more like taking a short-answer test than taking a multiple-choice test. Another motivation is economic; constructing high-quality multiple-choice test items is expensive. Finally, multiple-choice tests, unlike short-answer tests, lend themselves to test-taking strategies, which do not evaluate the student's understanding of the question.

Why automatic evaluation for short-answer tests? For the educational-testing community, one motivation is economic: if you can replace two human graders with one human and one system, you can reduce the cost of grading the examination. This substitution seems acceptable, as long as you can demonstrate that it won't affect the final grade and that human judges make the final decision, should the human and system disagree. A second, more important, motivation is that automated grading of short-answer questions provides students with much more immediate feedback—there is no need to wait for an instructor to provide a "ruling"

Mars Polar Lander—Where Are You?

(January 18, 2000) After more than a month of searching for a signal from NASA's Mars Polar Lander, mission controllers have lost hope of finding it. The Mars Polar Lander was on a mission to Mars to study its atmosphere and search for water, something that could help scientists determine whether life ever existed on Mars. Polar Lander was to have touched down December 3 for a 90-day mission. It was to land near Mars' south pole. The lander was last heard from minutes before beginning its descent. The last effort to communicate with the three-legged lander ended with frustration at 8 a.m Monday.

"We didn't see anything," said Richard Cook, the spacecraft's project manager at NASA's Jet Propulsion Laboratory. The failed mission to the Red Planet cost the American government more than $200 million dollars. Now, space agency scientists and engineers will try to find out what could have gone wrong. They do not want to make the same mistakes in the next mission. Controllers have been testing dozens of different scenarios to try and explain what might have happened to the lander. (Sources: Associated Press, CBC News Online, CBC Radio news, NASA)

A. Who is the Polar Lander's project manager?
B. What was the mission of the Mars Polar Lander?
C. When did the controllers lose hope of communicating with the lander?
D. Where on Mars was the spacecraft supposed to touch down?
E. Why did NASA want the Polar Lander to look for water?

Figure 6. Sample reading comprehension passage with questions. News story courtesy of the Canadian Broadcasting Corporation 4 Kids site, http://cbc4kids.com/general/whats-new/daily-news.

Question:
B. What was the mission of the Mars Polar Lander?

Correct sentence key:
Sentence 3: The Mars Polar Lander was on a mission to Mars to study its atmosphere and search for water, something that could help scientists determine whether life ever existed on Mars.

Answer key:
to study Mars' atmosphere and to search for water |
to help scientists determine whether life ever existed on Mars

Sample system answer:
to study its atmosphere

Answer-word recall:

key 1: (alternative 1):	[study, Mars, atmosphere, search, water]
system:	[study, atmosphere]
Recall:	2/5 = 40%
Precision:	2/2 = 100%

Figure 7. Answer keys and answer-word recall and precision calculations.

on the correctness of the answer. This immediacy supports interactive drills and testing, including diagnostic feedback for intelligent tutoring. However, an automated grading system's success ultimately depends on its ability to closely approximate the kinds of judgments a human grader would make.

Natural language research

As natural language system developers, our perspective on this issue differs from that of educators. Our long-term research goal is to develop systems that can read and understand common types of articles, stories, or news reports, to help people gather and digest vast amounts of information. For the past several years, we have been working specifically on developing systems that can take and pass reading-comprehension examinations—both multiple-choice and short-answer tests.[2] Figure 6 shows a sample reading-comprehension story and the related test questions.

How can we measure how well a system understands what it reads? One way is to have it answer questions about an article or story that it has read—that is, to have it take (and pass) the same kinds of tests we give to people, namely reading-comprehension tests. Our hypothesis is that if we can build systems that can pass general reading-comprehension tests, we can build commercially useful systems—systems that will provide factual answers to users' informational queries. We can track the progress of our research by "grading" these automated systems using the same tests that we use on people.

System development for any automated language-processing system requires constant testing with feedback: did the latest change make the system perform better or worse? This cycle becomes even more important when the modules of a reading-comprehension system rely on statistical methods, such as hidden Markov Models, or various kinds of machine-learning techniques. In fact, the cycle of testing, feedback, and improvement isn't much different from what a student needs when learning new subject matter. As we noted earlier, students also benefit from a tight loop of studying, doing drills, receiving diagnostic feedback, and taking tests.

To support our research, we have created several reading-comprehension test corpora and an evaluation infrastructure. We also hope to engage other groups in building and evaluating reading-comprehension systems[3] (see www.clsp.jhu.edu/ws2000/groups reading for information on the Johns Hopkins Summer Workshop devoted to reading comprehension). There is a related research effort, namely the Text Retrieval Conference's open evaluation of question-answering systems, sponsored by the National Institute of Standards and Technology. The TREC question-answering track evaluates systems that answer factual questions from information in a multi-gigabyte document collection. For the 1999 evaluation, the systems were presented with 200 questions to answer; for the 2000 evaluation, they were given 700 questions. The systems must provide a short (50 characters) or long (250 characters) answer to each

question.[4,5] Human judges then review the question and each system's response to determine whether the system response constitutes a correct answer.

Between the multi-site work on reading comprehension and the TREC conference on question answering, there is now an active natural language research community that needs a method for rapid automated grading of short answer questions. In each development cycle of an automated question-answering system, it is necessary to run the system and test its performance, to make sure it is improving. For rapid development, this needs to be done many times a day. It simply isn't possible to wait for an expert human to grade each output the system produces, before doing more development.

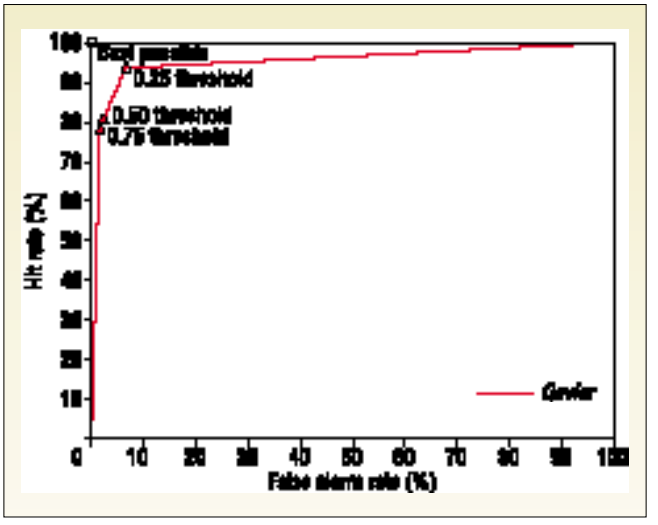


Figure 8. ROC curve for answer-word recall versus the human answer.

Strategies for automated short-answer grading

If we want to build an automated evaluation for short-answer questions, the first question is, what constitutes a correct answer? An easy ad hoc answer might be, whatever the human judge says is correct. One approach would be to examine many answers judged correct and build a classifier that we could train to produce similar judgments. Unfortunately, we don't have millions of "judged answers" available—although we do have the answer judgments from some 10 to 15 systems for the 200 TREC questions used in the 1999 evaluation.

A second approach—the one we are currently taking—is to compare the system answer to an answer key. The answer key might just come from the answers in the back of the book—if the book provides such answers. Otherwise, we have a human expert create a set of appropriate answers.[6] An answer key consists of one or more acceptable answers for each question. For example, Figure 7 shows question B from Figure 6 together with its answer key. We see that there are two alternate answers listed, separated by a vertical bar:

> to study Mars' atmosphere and to search for water |
> to help scientists determine whether life ever existed on Mars

Once we have the answer key, we can develop an automated comparison technique that measures the closeness of the system answer to the answer key.

Sentence correctness. Initially, we experimented with two measures: sentence correctness and answer-word recall. For sentence correctness, a human expert creates a correct sentence key, consisting of the sentence(s) from the passage that best answered the question. Determining correctness then simply requires comparing the sentence chosen by the system to the sentence in the answer key—or, if the sentences are numbered, comparing sentence numbers. Thus, in Figure 6, the correct answer sentence for question B ("What was the mission of the Mars Polar Lander?") is the third sentence: "The Mars Polar Lander was on a mission to Mars to study its atmosphere and search for water, something that could help scientists determine whether life ever existed on Mars." If the system returns that sentence, it is correct. If it returns another sentence, it is incorrect. However, there can be cases (a little over 10% in our simple initial corpus) where there is no single sentence that provides the answer. In addition, a sentence often contains more information than what is needed to answer the question: compare the length of the correct sentence (30 words) to the length of the answer key (9 or 10 words) in Figure 7. Clearly, we need some other measures of answer correctness.

Answer-word recall (and precision). Our long-term goal is to build a system that returns a concise answer, not just a sentence from the text. So we have also developed more fine-grained measures of answer correctness, based on overlap of stemmed content words between the answer key and the system answer. Answer-word recall measures coverage: it is the overlap between the system answer and answer key, divided by the number of words in the answer key. Perfect (100%) recall means all the answer key words appear in the system answer (possibly along with other other words); 0% recall means none of the answer key words are in the system answer. Answer-word precision measures conciseness: it is the overlapping words divided by the number of words in the system answer. Perfect (100%) precision means that all words in the system answer appear in the answer key; 0% precision means that none of the system answer-words appear in the answer key. In Figure 7, if the system returned as its answer "to study its atmosphere," there are two stemmed content words in the system answer (study, atmosphere), five content words in the first key (study, Mars, atmosphere, search, water), and eight in the second (help, scientist, determine, whether, life, ever, exist, Mars). The first answer key alternative provides a better score, based on the two overlapping content words: it yields a recall of 2/5 (40%) and a precision of 2/2 or 100%.

Our initial experiments used answer-word recall only, because our early systems returned an entire sentence. Using precision was less informative, because all the responses were sentences of roughly equal length. Also, because the sentences contained many extraneous words, the precision would have been very low. In these experiments, we found that the correct answer sentence and answer-word recall were reasonably well correlated. We compared performance with versions of the system containing different modules and found that if a module (for example, a proper name tagger) helped under one metric, it generally helped under the other. However, we also wanted to see how well these methods (particularly answer-word recall) corresponded to human judges.

To compare our answer-word recall measure to human judgments, we used the answers returned from various automated systems in the TREC evaluation, together with the associated human judgments of answer correctness. We asked our expert, Lisa Ferro of the MITRE Corporation, to

Table 1. Human judgments compared to thresholded answer word recall.

Recall	Human judged as incorrect Frequency	Human judged as incorrect Incorrect (%)	Human judged as correct Frequency	Human judged as correct Correct (%)
0.00	29,709	92.4	336	5.8
0.01–0.25	325	1.0	36	0.6
0.26–0.50	1,399	4.4	747	13.0
0.51–0.75	173	0.5	109	1.9
0.76–0.99	5	0.0	61	1.1
1.00	548	1.7	4,479	77.7
Total	32,159	100.0	5,768	100.0

Table 2. Error analysis comparing human to automated judgments for 72 discrepancies.

SOURCE of DISCREPANCY	#	%
TREC assessor (arguably) wrong	7	10%
Responses seemed relevant ("tough call")	27	37%
Thresholded word recall score wrong	38	53%
TOTAL	**72**	**100%**

create an answer key for the TREC data, and we plotted the answer-word recall versus the recorded human judgments, giving the Receiver Operating Condition curve shown in Figure 8. Specifically, we wanted to see the effect of selecting different thresholds for answer-word recall: to call an answer correct, should we require 25% answer-word recall or 100%? The ROC curve and the detailed figures in Table 1 show the trade-offs in selecting this threshold.[7] For example, if we call an answer correct if it has a word recall of over 25%, we get a hit rate of 93.6% (the automated system scores the answer as correct, given that the human assessor judges the system correct) and a false alarm rate of 6.6% (where the automated scorer judges the answer correct, although the human judge has called it incorrect). In general, there is good correlation between the positive answer-word recall and the human assessors judging the answer correct.

Based on this correlation, do we have a satisfactory automated method to grade short answers? Not yet. First, there are some obvious problems with our methodology. Word overlap is far too simple. For example, in this experiment, wrong answers predominate. If we had an automated scoring method that said *all* answers were wrong, it would agree about 85% of the time with the human judges; at a word recall threshold of 25%, the automated system and the human judge agree 93.6% of the time. Furthermore, there are limitations with word recall as a metric, regardless of threshold. Almost 6% of correct answers have 0% word recall (no word overlap between the answer key and a correct answer), and 1.7% of the incorrect answers have 100% word recall (all the answer key content words are found in the answer, but it's still wrong).

To understand the discrepancies between the human judges and automated comparison, we randomly selected 990 responses; out of these, we examined the 72 responses for which the automated system differed from the human judge. The results are shown in Table 2. In 7 cases, the human grader appears to have made a mistake; in 27 cases, it is unclear whether the human or the automated grading system was correct —so in almost half the cases (47%), it wasn't even clear if the automated system was wrong. For the remaining 38 cases, the automated decision based on thresholded word recall was clearly wrong, but many of these are easily fixed. Half of these errors (19/38) could be fixed by normalizing comparisons across different kinds of numerical expressions. This would fix the mismatch between an answer of "three" versus the answer key "3," or "Tuesday" and "April 3." Other discrepancies (7/38) were due to differences in answer granularity or answer phrasing—for example, if the answer key says "George Washington" and the system returns "Washington." And the remaining 12 discrepancies were due to other problems. So, with some additional work, answer-word recall might approximate human judgment reasonably well—or at least well enough to support rapid system development.

But aside from these detailed concerns, answer-word recall is a very limited measure. It ignores many important dimensions of what makes an answer correct. These include intelligibility (the coherence of the answer, to make sure it isn't just a "bag of words"); conciseness, or absence of extraneous material (measured perhaps by answer precision); and justifiability (providing some evidence from the relevant passage that the entity has the appropriate characteristics and the system didn't just guess correctly). Moreover, appropriate measures must be sensitive to the specific instructions that the student receives. Is the student asked to provide a minimal answer, the best phrase, or a sentence from the text? A conciseness measure, such as precision, is appropriate if the student is instructed to provide a short answer—but it is irrelevant if student is asked to identify the best sentence from the text as an answer, where the student has no control over conciseness. We need to develop task-appropriate measures that capture these additional dimensions of answer correctness.

We suspect that, in the long run, building a system that can grade a short-answer test might be almost as hard as building a system that can take (and pass) a short-answer test. If we succeed, we will have created a useful tool that will help both developers of natural-language understanding systems and educational-test developers. Ultimately, these methods will benefit both students and teachers—making drill and self-test materials more readily available to students and providing them with better feedback, while removing some of the drudgery from teaching by providing help with routine grading.

References

1. J. Burstein et al., "Computer Analysis of Essays," *Proc. NCME Symp. Automated Scoring*, 1998; www.ets.org/research/

ncmefinal.pdf (current Nov. 2000).

2. L. Hirschman et al., "Deep Read: A Reading Comprehension System," *Proc. 37th Ann. Meeting Assoc. Computational Linguistics*, Assoc. for Computational Linguistics, College Park, Md., 1999, pp. 325–332.

3. M. Light et al., *Proc. Workshop on Reading Comprehension Tests as Evaluation for Computer-Based Language Understanding Systems*, Assoc. for Computational Linguistics, New Brunswick, 2000.

4. E.M Vorhees and D.K. Harman, *The Eighth Text Retrieval Conf.*, NIST Special Publications, Feb. 2000.

5. E.J. Breck et al., "A Sys Called Qanda," *Proc. Eighth Text Retrieval Conf.*, Nat'l Inst. Standards and Technology, Gaithersburg, Md., 1999, pp. 499–506.

6. L. Ferro, "Reading Comprehension Tests: Guidelines for Question and Answer Writing," internal working paper, The MITRE Corp., Bedford, Mass, 2000.

7. E.J. Breck et al., "How to Evaluate Your Question Answering System Every Day…and Still Get Real Work Done," *Proc. Second Int'l Conf. Language Resources and Evaluation* (LREC-2000), European Language Resources Association, Paris, 2000.

To Grade or Not to Grade

Robert Calfee, University of California, Riverside

If, during this new millennium, we hope to encourage more groups in our society to participate—in education, work, and politics—then we need to better develop our capacity to communicate effectively. This outcome will require working on not only oral and written communication but also critical-thinking skills—and knowledge and skill in technical writing are touchstones for effective communication skills.

Student compositions reveal the mind in remarkable ways; unlike spoken discourse, the reader can retrace the author's reasoning and rethink his or her argument. Creative writing and casual communication certainly have a place in society, but success in business and industry depends on clarity, which is best learned and assessed through writing. Unfortunately, instructors don't always have sufficient time or resources to effectively grade student compositions or provide feedback on their reading-comprehension skills. This is where automated essay-grading systems such as those described in this issue can help.

For instance, the Intelligent Essay Assessor described by Thomas Landauer and his colleagues provides the novice writer with instant feedback about the match between his composition and a "semantic core" created from a related set of readings and compositions prepared by other writers. IEA also provides structural information: "You captured idea *x* quite well, but need to look more carefully at *y* and *z*."

The current system

Research suggests that becoming an effective writer requires at least these three elements: guided practice, effective instruction, and informed feedback.[1] Today's classrooms, from elementary grades through graduate studies, rely mostly on the guided-practice element—more specifically, instructors give writing assignments but offer only limited guidance.

The tempting topic of effective instruction must wait for another time; suffice it to say that high school English classes tend to emphasize literature and grammar in roughly equal proportions, neither of which contribute to the knowledge and skill needed to design the technical reports that I emphasize in this essay. The conditions remain the same during the college years (especially for students identified as requiring remediation), with the exception of the occasional technical-writing course found in engineering and business schools. The typical research paper needs to define a problem, lay out a few analytic points, and reach a conclusion—something akin to the five-paragraph essay. In high school and college, the emphasis is more often on creativity.

Regarding informed feedback, ideally the student receives a close and careful reading of one or more drafts, with attention to several elements of effective composition—usually organization and coherence, style and usage, and mechanics (grammar, spelling, and so on). Of these principles, textual integrity clearly matters most for the reader (until the mechanics interfere with understanding, micro-level idiosyncrasies annoy but do not detract). Granted, most writing instructors spend a good deal of energy on editorial corrections of surface nonconventions such as incorrect punctuation and grammatical errors. This approach makes sense when you consider that the instructors must grade hundreds of papers each week. They don't have time to think about individual compositions. In addition, students might become argumentative when challenged on a composition's high-level features: "What do you mean it's not coherent? It includes all the facts!" Grammar and spelling, in contrast, are either right or wrong, assuming acceptance of certain conventions. Transforming this state of affairs so that teachers can change their approach would require more teachers (which is unlikely) and shifting the viewpoint of what matters from an emphasis on style to an emphasis on substance (which would be difficult).

The last few decades have seen numerous efforts to automate the intricacies of assessing written material, but few advances deal with scoring. For hunt-and-peckers like me, the appearance of spelling and grammar checkers was a blessing. They don't solve every problem or guarantee an "A," but at least you don't have to worry about minutiae (such as how to spell minutiae) while composing, or even when revising and polishing. But what about automating the "big stuff"?

Automating the big stuff

The three other essays in this installment of Trends & Controversies also address this issue of increased automation capabilities, and they all share certain conclusions.

First, virtually every automated system generates scores that correlate with the ratings of human judges as closely as human judges agree with one another. The high correlations might reflect the interrelatedness of different elements in naturally occurring compositions; writers who produce well-organized passages also use a rich vocabulary and carefully revise mechanics. Experiments with test passages in which the various elements are independently varied (that is, well organized but with poor mechanics, or strong vocabulary but with lots of misspellings) would show how the different systems in this issue respond to different elements.[2]

Second, all three essays take for granted the reading–writing connection.[3] In most nonacademic settings, this connection undergirds writing tasks. For example, when an engineer prepares an evaluation report on a new widget, she first learns about the widget, then outlines the report's main points (often using a model), studies other documents for background, and finally prepares a draft. Reading and writ-

ing intertwine continuously. In school, however, traditions separate reading and writing in all but a few settings, mostly in the later grades and college-bound tracks.

Third, all the essays focus on scoring but also mention the need for more detailed feedback. In "Automated Grading of Short-Answer Tests," the authors suggest the value of prompt and detailed feedback, but only for microlevel questions ("You don't understand the point of this specific question"). The emphasis on scoring links to summative evaluation—to feedback at the end of the writing process. Automated systems can potentially convey formative information along the way, but the other three essays offer little information on this possibility.

I contend that automation's value in helping novices become expert writers rests more on the beginning rather than the end of the learning process. Let me offer a metaphor. The area north of Monterey, California, is famed for its artichokes, and one of my favorite road stops is an old barn displaying a "grader," an inclined plane with crossbars narrow at the top and wide at the bottom. The tiny artichokes fall through first, and the giant artichokes don't fall through until the bottom. These farmers evidently know how to both develop and grade their products. The question that comes to mind is, how might we use the power and cost-effectiveness of automated text-evaluation systems to provide inexperienced writers with feedback and support for improving performance, thus developing or growing—and not just grading—our students?

The Intelligent Essay Assessor

In the effort to understand ways of better developing students' reading and writing skills, I rely on the IEA system as the touchstone, mostly because I am more familiar with it than with the other systems, both conceptually and operationally. In "The Intelligent Essay Assessor," Thomas Landauer, Darrell Laham, and Peter Foltz introduce the IEA and discuss its effectiveness. However, several IEA elements not highlighted in their essay offer considerable promise for supporting student growth. Specifically, IEA can employ automated strategies that support both students and teachers by providing informed feedback through an interactive assessment process that intertwines curriculum, instruction, and assessment. This ability to offer informed feedback assumes greater importance than inter-rater consistency.

To develop this point, I first expand on Landauer, Laham, and Foltz's essay. What outcomes (other than scoring) does IEA deal with, and how well does it handle these matters? To what degree is it cost-effective and time-efficient in these ancillary domains? What about practical matters of acceptability and feasibility; will teachers find it helpful?

What first attracted me to IEA is the reading–writing connection. School success from the middle-school grades onward requires students to read a body of material, analyze its content, and write a response. Unfortunately, today's curricula do not help students learn this rather demanding task. Reading and writing are separate parts of the curriculum, with different times, textbooks, and tests. Often, in science and social studies, teachers are poorly prepared to handle either reading or writing. These lacunae have several consequences for curriculum, instruction, and assessment. If the teacher's curricular goal is to help students appreciate the concept of energy (different forms of energy, the conservation principle), and the teacher doesn't know how to help the students with vocabulary and comprehension, then the curricular goal falls through the cracks. The teacher will also be at a loss as to how to provide adequate instruction and conduct proper assessment.

IEA goes to the crux of the reading–writing connection. To what degree can a student read about a topic, analyze it from a defined perspective, and write a response?

Imagine a middle-school science teacher who presents to the class a project-based assignment on deserts:

> This month we are going to study the desert. I've brought in many books and found some related Web sites. Next week, we'll visit the Jurupa Science Museum, where you can see plants and animals, along with rocks and other items typically found in a desert environment. Earlier this year, a builder asked the Town Council to build 500 new homes in Canyon Crest, up the hill from where most of you live. It's just desert, and many people think we need more homes. Our job is to study what's going on and write letters to the Town Council about what we think they should do.

This assignment goes well beyond the typical middle-school curriculum and requires extraordinary instructional capabilities. But suppose that the project moves ahead; the letters to the Town Council begin to take shape. How can the teacher judge the formative adequacy of the initial drafts and the progress toward the final summative goal? The teacher and his young charges should eventually know whether their project produced a collection of exemplary arguments or a set of mundane and unconvincing sputterings. Furthermore, along the way, the teacher should know how to provide assessments, as well as suggestions for enhancing the works in progress. The provision of this along-the-way input—the essence of all levels of professional development—is surely important for students.

If applied to such an assignment, IEA would perform some tasks very well, others indifferently, and a few not at all. Of course, IEA (along with the other systems described in this issue) is still evolving, so its full potential remains to be seen.

IEA's strong points. IEA goes to the crux of the reading–writing connection. To what degree can a student read about a topic, analyze it from a defined perspective, and write a response? The IEA strategy—conceptual and technical—is to construct a canonical template around the focal topic (the desert, for example). Latent Semantic Analysis, a theoretical technique with origins in the cognitive revolution of the 1970s, starts with word–concept networks, adds a dash of mathematical technology to crunch the information producing the template, and then presents an interface for translating theory and technology into practical outcomes. The process begins with input and output texts. Input texts include various readings on the topic (books and other artifacts); student essays—which experienced judges rate as more or less adequate—are the output texts. LSA begins by getting to know the territory, but then learns the difference between "knowing well" and "knowing not so well"—or "expressing well" and "not so well."

Using LSA, IEA grades responses particularly well. The data that Landauer and his colleagues present in "The Intelligent Essay Assessor" show that this system does an excellent job of scoring essays in which students read a text and respond to a related

prompt. The technique is both time-efficient and cost-effective; students and teachers receive information quickly and precisely, and the cost promises to be reasonable. IEA also provides informed feedback by examining student work as it emerges, offering suggestions about overall quality (the likely grade) and identifying places for improvement. For example, IEA's Summary Street application (my favorite) checks a reader's capacity to extract the key information in a text by means of a written summary. Unlike the MITRE system, IEA assesses comprehension by asking the reader to decide what is important. Feedback immediately tells the summarizer which text segments have been neglected and identifies material that is irrelevant or redundant. The program's potential for enhancing comprehension of expository passages merits special attention.

An average performance. However, even Summary Street could be enhanced in a couple of ways. First, it could include teaching texts that pose increasingly defined and difficult challenges to the reader. Science textbooks, for instance, often contain *seductive distracters*—information included to spark interest but irrelevant to the main points. Second, the program could inform the student about missing elements in a chain of reasoning. Feedback from the main IEA program tends to be rather generic, although some specific indicators are quite valuable. The internal coherence and essay validation measures, for instance, point the student to text-level problems but do not offer much help about how to address the issues.

In addition, IEA provides various ancillary supports, including the usual grammatical and spelling backups, with the usual pros and cons. The developers should leave these matters to others, remaining aware that some audiences might value information about writing mechanics. The weighted contributions in Figure 2 (see the first essay) seem about right: 75% content, 15% style, and 10% mechanics. These statistics provide reasonable guidelines for further system development.

Also, although the current system offers students (and teachers) more detailed and instructive information than other automated systems with which I am familiar, it still lacks the overarching framework that can respond to the question, "What does the novice reader and writer need to know to master this task, and when during the writing process does he or she need to know it?" This article is not the place to fully explore this question, but I offer a tentative answer in the next section.

Room for improvement. IEA does not perform some pedagogical tasks at all—for example, assessing text structure, which is important in both reading and writing.[4] LSA depends on an inductive strategy to handle content. The program inputs a huge amount of information and calculates dimensional cosines. It's amazing how well this strategy works for grading. But so do other, less sophisticated methods, perhaps because of the interrelatedness mentioned earlier. Text structures do more than replicate content organization, however; they shape, amplify, and transform the content. The Greeks produced not just geometry but also rhetoric, the structural principles that undergird modern communications. As the authors of "Beyond Automated Essay Scoring" explain, adding rhetorical structures to existing systems will be nontrivial (and, for some purposes, perhaps unnecessary). But for IEA and LSA to serve as an instructional support system that offers efficient guidance, as well as informed feedback, incorporating text frameworks might be significant for future developments.

Vocabulary also matters, and IEA does not measure this skill. To avoid dings for misspelling, the novice writer often relies on everyday words, but the sophisticated rater sees a composition replete with plebeian words and clichéd phrases. IEA provides a readability index that reflects unusual usages but does not indicate to the student (or teacher) the level of skill in expressive vocabulary. The importance of this matter recently captured my attention when colleagues and I discovered that today's middle-school students tend to eschew a risky lexicon, staying with terms that are tried, true, high frequency, and easy to spell.

Practical concerns

Many of today's classrooms, especially those serving poor communities, lack the resources needed to implement these programs. A discussion of the disparities in human capital would take me beyond this essay's scope, but it is clear that schools in underserved neighborhoods do not have the equipment needed for effective access to net-based systems. Pay a visit to such a school and you will discover a handful of computers from the early 1990s, maybe with a phone-line modem. This situation has evoked well-publicized but generally scattershot responses. The harm from inadequate resources might be more substantial than is generally recognized. Mark Russell and Walt Haney, for instance, found that students preparing their written assignments on computers rather than with paper and pencil improved on average from "Needs Improvement" (a delicate way of saying "failed") to "Passing."[5] This is because the students write more with the first method and spend more time revising their drafts.

Imagine the impact of combining the best of what we know with what we can do—imagine if we could do everything, from giving immediate and informative feedback to ensuring that every student spends time tinkering with a keyboard. Computer technologies—including better equipment and the kinds of response systems described in this issue, will not put teachers out of business, as some seem to fear. Rather, they will provide them with tools that amplify their professional skill and knowledge. ■

Acknowledgment

The Office of Educational Research and Improvement (IERT-9979834) supported this essay's preparation.

References

1. C.R. Cooper and L. Odell, eds., *Research on Composing: Points of Departure*, Nat'l Council of Teachers of English, Urbana, Ill., 1978.

2. S.W. Freedman, "Student Characteristics and Essay Test Writing Performance," *Research in the Teaching of English*, Vol. 17, 1983, pp. 313–324.

3. N.N. Nelson and R.C. Calfee, eds., *The Reading-Writing Connection: The Yearbook of the Nat'l Soc. for the Study of Education*, Univ. of Chicago Press, Chicago, 1998.

4. M.J. Chambliss and R.C. Calfee, *Textbooks for Learning: Nurturing Children's Minds*, Blackwell, Oxford, UK, 1998.

5. M. Russell and W. Haney, "Testing Writing on Computers: An Experiment Comparing Student Performance in Tests Conducted via Computer and via Paper-and-Pencil," *Education Policy Analysis Archives*, Vol. 5, No. 3, May/June, 1997; http://epaa.asu.edu/epaa/v5n3.html (current Nov. 2000).