Criterion[SM] Online Writing Evaluation Service

ETS's *Criterion*[SM] *Online Writing Evaluation Service* is an award-winning Web-based system that provides automated scoring and evaluation of student essays. The system uses *e-rater*®, an automated essay scoring engine, and a suite of programs that detect errors in grammar, usage, and mechanics, identify discourse elements in the essay, and recognize elements of undesirable style. Together, these evaluation capabilities provide students with specific feedback to help them improve their writing skills. *Criterion* is among ETS's *System 5*[TM] suite of K-12 products and services designed to improve school performance and student achievement.

Developed with Teachers and Students in Mind

Write, receive feedback, and revise. Repeating this process often is how students best improve their writing skills. Unfortunately, it also places an enormous load on the classroom teacher who is faced with reading and providing feedback for perhaps 30 or more essays for every topic assigned. As a result, teachers are not able to give writing assignments as often as they would wish.

Recognizing how it would benefit teachers and their students, educational researchers began in the early 1960s to look for ways to automate essay scoring (Page, 1966; Burstein et al., 1998; Foltz, Kintsch, & Landauer 1998; Larkey, 1998; Elliot, 2003). The 1980s then saw the initiation of pioneering work in automated feedback with the Writer's Workbench (MacDonald *et al*., 1982).

ETS's *Criterion*[SM] *Online Writing Evaluation Service*, which was first released in September 2001, brings together both efforts by offering automated essay scoring and diagnostic feedback that is specific to the student's essay and is based on the kinds of evaluations that teachers typically provide. *Criterion* is intended to be an aid, *not a replacement*, for classroom instruction. Its purpose is to ease the instructor's load, thereby enabling the instructor to give students more practice writing essays.

Who uses *Criterion*?

The *Criterion Online Writing Evaluation Service* won an Education Software Review (EDDIE) Award from *ComputEd Gazette* in August 2005, and two months later, it won an Award of Excellence from *Technology & Learning* magazine. More recently, *Criterion* was named a finalist in the Software & Information Industry Association Codie Award program. The coveted Codie awards showcase the software and information industry's finest products and services and to honor excellence in corporate achievement and philanthropic efforts. *Criterion* is a finalist in two categories: the Best Instructional Solution for Language/Arts English—Secondary, and the Best Instructional Solution for English Language Acquisition.

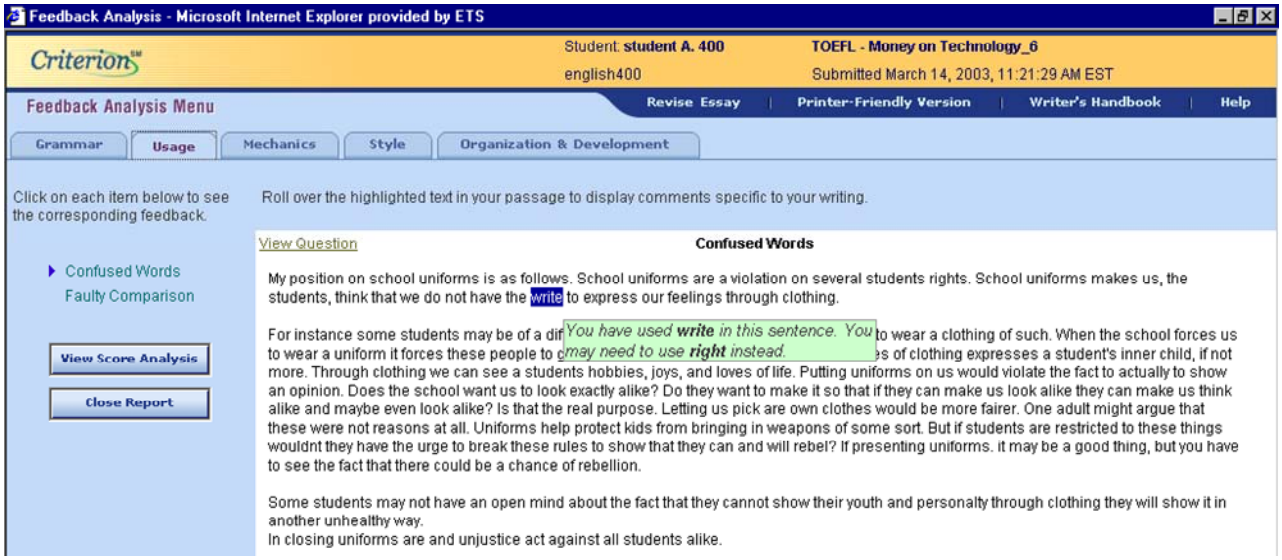
In 2006, the program has had more than 2 million essay submissions and has been purchased by elementary, middle and high schools, public charter schools, school districts, community colleges, and universities. Outside the United States, the system is used in many countries including Canada, England, India, Qatar, Vietnam, Taiwan, Singapore, and Japan.

How Does *Criterion* work?

Criterion uses two complementary applications that are based on natural language processing (NLP) methods. One is an application that is comprised of a suite of programs that evaluate and provide feedback for errors in grammar, usage, and mechanics; identify the essay's discourse structure; and recognize undesirable stylistic features. The companion scoring application extracts linguistically-based features from an essay and uses a statistical model of how these features are related to overall writing quality to assign a score to the essay. A new version of the *Criterion* software is scheduled for release with the start of each school year with possible interim releases as needed. Because the software is centrally hosted, updates are easily deployed and made immediately available to users. A group of ETS developers maintain the software.

Feedback

The writing analysis tools identify five main types of errors—agreement errors, verb formation errors, wrong word use, missing punctuation, and typographical errors. The system is trained on a large set or corpus of edited text, from which it



***Criterion* – Sample *Usage* Feedback**

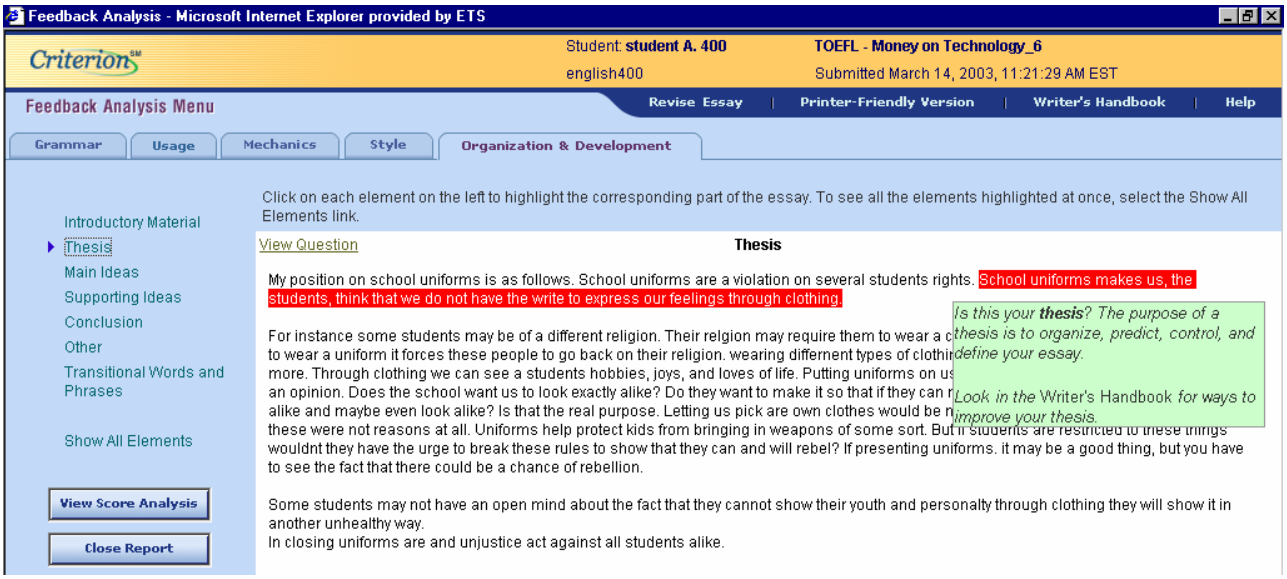
extracts and counts sequences of adjacent word and part-of-speech pairs called *bigrams*. For example, the noun phrase, *good job*, would be represented by the bigram – *ADJ NOUN*, since *good* is an adjective (ADJ) and *job* is a noun (NOUN). The system then searches student essays for bigrams that occur *much less often* than would be expected based on the corpus frequencies.

The expected frequencies come from a model of English that is based on 30-million words of newspaper text. Every word in the corpus is automatically labeled with a part-of-speech tagger that has been trained on student essays (Ratnaparkhi, 1996). Here are a few examples of what this part-of-speech tagger produces: "a" in the word sequence "a good job" would be labeled as a singular indefinite determiner (DET_SING), "good" is labeled as an adjective (ADJ), and "job" as a singular common noun (NOUN_SING). Frequencies are then collected for each function word (determiners, prepositions, etc.) and also for each adjacent pair of tags. For "a good job" there would be two bigrams: DET_SING-NOUN_SING and ADJ-NOUN_SING.

To detect violations of general English language rules, the system compares what it actually finds and what it expects to find, given the frequencies in the general corpus. Researchers interested in finding technical terms or collocations commonly use the statistical methods employed by the system to detect combinations of words that occur *more frequently* than would be expected based on the assumption that the words are independent. *Criterion* uses the measures for the opposite purpose—to find combinations that occur *less often* than expected, and therefore might be evidence of a grammatical error (Chodorow & Leacock, 2000). For example, the bigram that represents the phrase "this desks," and similarly tagged sequences that show number disagreement, occur much less often than expected in the newspaper corpus based on the frequencies of singular determiners and plural nouns.

The system uses two complementary methods to measure association: pointwise mutual information and the log likelihood ratio. The first gives the direction of association (whether a bigram occurs more often or less often than expected, based on the frequencies of its parts), but this measure is unreliable with sparse data (a data set that does not contain sufficient examples). The log likelihood ratio performs better with sparse data and gives the likelihood that the elements in a sequence are independent. Using both measures provides the direction *and* strength of association.

Of course, no simple model based on neighboring elements is adequate to capture all English grammar. This is especially true when we restrict ourselves to a small window of two elements, as we do with bigrams. For this reason, the application has filters to allow for low probability, but nonetheless grammatical, sequences. With bigrams that detect subject-verb agreement, filters check that the first element of the



***Criterion* – Sample Organization & Development Feedback**

bigram is not part of a prepositional phrase or relative clause (e.g., the bigram "college assume" in "My friends in college assume..." is not an error because the subject of "assume" is "friends").

Confused words. Some of the most common writing errors are due to the confusion of homophones, words that sound alike. E-*rater* detects errors among *their/there/they're*, *its/it's*, *affect/effect* and hundreds of other such sets. For the most common of these, the system uses 10,000 training examples of correct usage from newspaper text and builds a representation of the local context in which each word occurs by taking into account the two words and part-of-speech tags that appear before and after the confusable word. For example, a context for "effect" might be "a typical effect is found," and the local context would consist of the determiner "a" and adjective "typical" as well as a form of the verb "BE" in "is" and the past participle "found." For "affect," a local context might be "it can affect the outcome," where a pronoun and modal verb are on the left, and a determiner and noun are on the right.

Some easily confused words, such as *populace/populous*, are so rare that a large training set cannot easily be assembled from published text. In this case, generic representations are used. The generic local context for nouns consists of all the part-of-speech tags found in the two positions to the left of each noun and in the two positions to the right of each noun in a large corpus of text. In a similar manner, generic local contexts are created for verbs, adjectives, adverbs, etc. These serve the same role as the word-specific representations built for more common homophones. Thus, "populace" would be represented as a generic noun and "populous" as a generic adjective.

The frequencies found in training are then used to estimate the probabilities that particular words and parts of speech will be found at each position in the local context. When an easily confused word is encountered in an essay, *e-rater* uses a maximum entropy statistical classifier (Ratnaparkhi, 1996) to select the more probable member of its homophone set, given the local context in which it occurs. If this is not the word that the student typed, then the system highlights it as an error and suggests the more probable homophone.

Undesirable style. The identification of good or bad writing style is subjective; what one person finds irritating another may not mind. E-*rater* highlights aspects of style that the writer may wish to revise, such as the use of passive sentences, as well as very long or very short sentences. Another feature of undesirable style that the system detects is the presence of overly repetitious words, a property that might affect an essay's overall quality rating. Burstein and Wolska (2003) provide a detailed description of how the tools in *Criterion* identify words that a student may overuse and which could interfere with the smooth reading of the essay.

Essay-based discourse elements. A well-written essay generally should contain discourse elements, which include introductory material, a thesis statement, main ideas, supporting ideas, and a conclusion. When grading essays, teachers commonly provide comments on these aspects of discourse structure, and the system makes decisions that exemplify how teachers perform this task.

For *e-rater* to learn how to identify discourse elements, human readers annotated a large sample of student essays with essay-based discourse elements. The annotation schema reflected the discourse structure of essay writing genres, such as persuasive writing in which a highly-structured discourse strategy is employed to convince the reader that the thesis or position stated in the essay is valid. The discourse analysis component uses a decision-based voting algorithm that takes into account the discourse labeling decisions of three independent discourse analysis systems. Details on the three systems are available in Burstein, Marcu, and Knight (2003).

How e-rater works

Earlier versions of e-rater had some 50 features, and a subset of these features would be selected to score the particular set of essays. The newer version of e-rater uses a fixed set of about 10 features in seven categories from which it derives the final score.

The seven score categories are:

- *Grammar score* – based on errors such as those in subject-verb agreement among others
- Mechanics score – derived from errors in spelling and other like errors
- *Usage score* – based on such errors as article errors and confused words (an example would be an instance in which the essay writer uses a word that although phonetically similar has a different meaning from the intended word; using "to" where it would have been proper to use "too")
- *Style score* – based on instances of overly repeated words and the number of very long or very short sentences as well as other such features
- *Lexical complexity score* – drawn from information such as the level of vocabulary the essay writer uses in the essay
- *Organization/development score* – based on the identification of sentences that correspond to the background, thesis, main idea, supporting idea, and conclusion
- *Prompt-specific vocabulary usage score* – derived from e-rater's evaluation of the word choice in an essay and the similarity to the word choice in samples of low- to high-quality essays written on the same topic

In addition to these seven score categories, essay length also may be considered and weighted in a controlled way.

How Do We Know *Criterion* Provides Useful Feedback?

Criterion's main use is to give students more opportunities to practice writing. Consequently, it is essential that the system provide students with accurate feedback with regard to errors, comments on potentially undesirable style, and information about discourse elements and organization of the essay. For students to improve their writing, the feedback needs to be similar to comments they would receive from their instructors. This is why developers assess the accuracy of *e-rater* scores and the writing analysis feedback by examining the agreement between people who perform these tasks. This inter-rater performance is the gold standard against which human-system agreement is compared. Additionally, where relevant, both inter-rater human agreement and human-system agreement are compared to baseline algorithms, when such algorithms exist. The performance of the baseline is considered a lower threshold. For a capability to be used in *Criterion* it must outperform the baseline measures and, in the best case, approach human performance.

For the different types of feedback, researchers evaluate performance using *precision* and *recall*. In identifying or labeling discourse elements (e.g. a thesis statement) or grammatical error, precision is the number of times the system and a human rater agree on the identifier or label divided by the total number the system identifies. For recall, it is the same except the number of system/human rater agreements is divided by the total number human raters have identified.

For grammar, usage, and mechanics or errors detected using bigrams and by the misuse of confusable words, ETS researchers decided to err on the side of precision over recall. The thought is that it is better to miss an error than to tell the

student that a well-formed construction is ill-formed. A minimum threshold of 90% precision was set for these kinds of errors to be included in *e-rater*. Recall varies for different error types and for particular confusable words. To estimate recall, 5,000 sentences were annotated to identify specific types of grammatical errors. For example, *e-rater* correctly identified 40% of the subject-verb agreement errors that the annotators identified and 70% of the possessive marker (apostrophe) errors. Precision for subject-verb agreement errors is 92% and for possessive marker errors is 95%. The confused word errors were detected 71% of the time.

To diagnose overly repetitious word use, ETS researchers had two judges evaluate 300 essays. The judges determined that 74 essays in the evaluation sample had overly repeated words, so the results are based on this subset. Using precision, recall, and the mean of precision and recall, the researchers found that the feature that measures a word's relative frequency in an essay best matched the findings of the judges. The researchers determined that words that the system flagged as appearing at a frequency of 5% or more were in agreement with what the judges cited for repetitious word use.

To evaluate system performance in identifying discourse structure, ETS researchers computed precision, recall, and the mean between the two for the system, the baseline algorithm, and also between the two judges. The baseline algorithm assigns a discourse label to each sentence in an essay based solely on the sentence position. An example of a baseline algorithm assignment would be that the system labels the first sentence of every paragraph in the body of the essay as a "Main Point." The results from a sample of 1,462 human-labeled essays indicate that the system outperforms the baseline measure for every discourse category. Overall, the precision, recall, and mean for the baseline algorithm are 0.71, 0.70, and 0.70, respectively, while for the discourse analysis system, precision, recall, and mean are uniformly 0.85 (Burstein, Marcu, & Knight, 2003). The average precision, recall, and mean are approximately 0.95 between two judges.

E-rater Scoring Evaluation

The performance of *e-rater* scoring is evaluated by comparing its scores to those of human judges. This is carried out in the same manner that ETS employs for its reader scoring sessions. If two judges' scores match exactly, or if they are within one point of each other on the 6-point scale, an additional reader is not required to resolve the score discrepancy. When judges disagree by more than a single point, a third judge resolves the score. In evaluating *e-rater*, its score is treated as if it were one of the two judges' scores. See Burstein, et al. (1998) for a detailed description of this procedure.

For a baseline, the agreement is computed based on the assignment of the modal (or most common) score to all essays in the cross-validation sample. Typical exact plus adjacent agreement between *e-rater* and the score assigned by a human rater is approximately 97%, which is comparable to that between two readers. Baseline agreement using the modal score is generally 75%-80%.

Criterion development

ETS developed *Criterion* with a group of 15 developers at a cost of more than a million dollars. The team had considerable experience in developing electronic scoring and assessment products and services having previously developed ETS's Online Scoring Network (OSN) and having implemented *e-rater* within OSN. One of the larger challenges the team faced was with the *Criterion* interface and how to present the potentially overwhelming amount of feedback information in a manageable format via browser-based software. They accomplished this by showing screen shots and prototypes to teachers and students and eliciting their comments and suggestions.

ETS researchers continue to hone the algorithms used in applications supporting Criterion, as well as to look into new features to add to the essay evaluation service. One current concentration has been on the detection of grammatical errors that are important to specific native language groups, such as identifying when a determiner is missing (a common error among native speakers of East Asian languages and of Russian) or when the wrong preposition is used. While the system identifies discourse elements, it does not evaluate their quality. Researchers are extending the analysis of discourse to be able to assess the expressive quality of each discourse element. This means, for example, not only telling the writer which sentence serves as the thesis

statement but also indicating how good that thesis statement is. The system's developers also are in regular contact with teachers who use *Criterion* and, wherever possible, use NLP technology to incorporate their suggestions into the system.

References

Attali, Y., & Burstein, J. (2004, June). *Automated essay scoring with e-rater V.2.0.* Paper presented at the Annual Meeting of the International Association for Educational Assessment, Philadelphia, PA.

Breland, H. M. (1996). Word frequency and word difficulty: A comparison of counts in four corpora. *Psychological Science 7*(2): 96-99.

Breland, H. M., Jones, R. L., & Jenkins, L. (1994). *The College Board vocabulary study* (ETS Research Report No. 94-26). Princeton, N.J.: ETS.

Burstein, J., & Wolska, M. (2003). Toward evaluation of writing style: Overly repetitious word use in student writing. In proceedings of the *Tenth Conference of the European Chapter of the Association for Computational Linguistics.* East Stroudsburg, Penn.: Association for Computational Linguistics.

Burstein, J., Marcu, D., & Knight, K. (2003). Finding the WRITE stuff: Automatic identification of discourse structure in student essays. *IEEE Intelligent Systems: Special Issue on Natural Language Processing 18*(1): 32-39.

Burstein, J., Kukich, K., Wolff, S., Lu, C., Chodorow, M., Braden-Harder, L., & Harris M. D. (1998). Automated scoring using a hybrid feature identification technique. In proceedings of the *Thirty-Sixth Annual Meeting of the Association for Computational Linguistics,* 206-210. East Stroudsburg, Penn.: Association for Computational Linguistics.

Chodorow, M., & Leacock, C. (2000). An unsupervised method for detecting grammatical errors. In proceedings of the *First Annual Meeting of the North American Chapter of the Association for Computational Linguistics,* 140-147. East Stroudsburg, Penn.: Association for Computational Linguistics.

Elliott, S. (2003). Intellimetric: From here to validity. In M. Shennis & J. Burstein (eds.), *Automated Essay Scoring: A Cross-Disciplinary Perspective,* Hillsdale, N. J.: Lawrence Erlbaum Associates.

Foltz, P. W., Kintsch, W., & Landauer, T. K. (1998). Analysis of text coherence using latent semantic analysis. *Discourse Processes 25*(2-3): 285-307.

Golding, A. (1995, June). A Bayesian hybrid for context-sensitive spelling correction. Paper presented at the *Third Workshop on Very Large Corpora,* 39-53. 30 Cambridge, MA.

Larkey, L. (1998). Automatic essay grading using text categorization techniques. In proceedings of the *Twenty-First ACM-SIGIR Conference on Research and Development in Information Retrieval,* 90-95. New York: Association for Computing Machinery Special Interest Group on Information Retrieval.

MacDonald, N. H., Frase, L. T., Gingrich P. S., & Keenan, S. A. (1982). The Writer's Workbench: Computer aids for text analysis. *IEEE Transactions on Communications 30*(1): 105-110.

Page, E. B. (1966). The imminence of grading essays by computer. *Phi Delta Kappan,* 48:238-243.

Ratnaparkhi, A. (1996). A maximum entropy part-of-speech tagger. In proceedings of the *Empirical Methods in Natural Language Processing Conference,* University of Pennsylvania. East Stroudsburg, Penn.: Association for Computational Linguistics.

Rudner, L. M., & Liang, T. (2002). Automated essay scoring using Bayes' theorem. *Journal of Technology, Learning, and Assessment, 1*(2).

Salton, G., Wong, A., & Yang, C. S. (1975). A vector space model for automatic indexing. *Communications of the ACM 18*(11): 613-620.

Shermis, M., & Burstein, J. eds. (2003). *Automated essay scoring: A cross-disciplinary perspective.* Hillsdale, N.J.: Lawrence Erlbaum Associates.

Listening. Learning. Leading.