Listening. Learning. Leading.®

ETS Automated Scoring and NLP Technologies

Using natural language processing (NLP) and psychometric methods to develop innovative scoring technologies

The growth of the use of constructed-response tasks (test questions that elicit open-ended responses, such as short written answers, essays, and recorded speech) — along with the ongoing need to report test results in a timely fashion — spurred the development of innovations in scoring.

ETS began conducting research on automated scoring of constructed-response tasks in the 1980s and expanded this research to incorporate NLP technologies in the mid-1990s. This line of research ultimately resulted in multiple types of automated scoring technologies for such fields and areas as architectural design, mathematics, essays, typed answers, computer science, and spoken responses. Today, the *e-rater* engine is used to assist human raters in scoring academic essays on the GRE® General Test and the TOEFL® test. The *e-rater* engine reliably predicts human scores, as indicated by more than 10 years of system evaluations.

Over the years, ETS researchers have extended their work in NLP research and developed other automated scoring technologies: the *c-rater*™ system, the *m-rater* engine, and the *SpeechRater*℠ engine. ETS has incorporated these technologies into many of its testing programs, products, and services, including the *Keeping Learning on Track*® program, the *Criterion*® Online Writing Evaluation service, and TOEFL® Practice Online. ETS also uses NLP to develop learning tools and test development applications, as well as the *Text Adaptor* tool.

Examples of ETS's NLP capability are available at **www.ets.org/research**.

What Is Automated Scoring?

At ETS, we have made a substantial investment in research on the automated scoring of open-ended tasks for more than a decade. Our goal is to improve the validity of the score results, while creating methods and computer applications that reduce the cost and effort involved in using human graders. We believe that scores should support the uses of assessment regardless of the role computers play in creating them.

We briefly describe here the automated scoring applications that we have developed. These applications — the *e-rater* engine, the *c-rater* system, the *m-rater* engine, and the *SpeechRater* engine — help evaluate responses to tasks that require test takers to write essays, fill in the blank, write math equations, or give oral responses. We also describe the *Text Adaptor* tool, which teachers can use to make tests and other classroom materials more understandable to English language learners in situations where knowledge of English is not an instructional goal.

We continuously refine each of these applications based on the best-available definitions of skill and proficiency, as well as state-of-the-art psychometric, NLP, and speech science. We are also able to provide quick results through the use of web-based services.

To learn more about how your program can benefit from the ETS automated scoring capabilities, contact **RDWeb@ets.org**.

The *c-rater*™ System

The *c-rater* system is ETS's technology for the automatic analytic-based content scoring of short free-text responses, ranging in length from a few to approximately 100 words.

Analytic-based content is the kind of content that is predefined by a test developer in terms of main ideas or concepts. These concepts form the evidence that a student needs to demonstrate as her/his knowledge in his/her response. The following shows an example of a test item with the expected analytic-based content in the response and one way of assigning score points.

Test Item (Full credit: 2 points)

Stimulus:

A reading passage

Prompt:

In the space provided, write the question that Alice was most likely trying to answer when she performed Step B.

Concepts or main/key points:

C_1: How does rain formation occur in winter?

C_2: How is rain formed?

C_3: How do temperature and altitude contribute to the formation of rain?

Scoring rules:

2 points for C_1

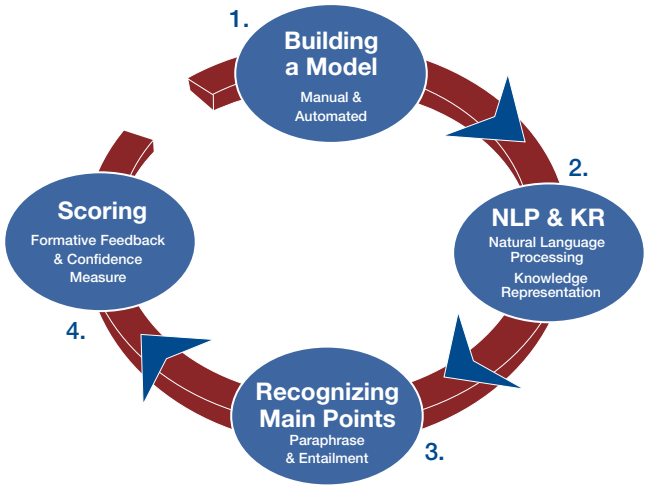
1 point for C_2 (only if C_1 is not present) or C_3 (only if C_1 and C_2 are not present)

Otherwise 0 points

There are four main processes in the *c-rater* system:

1. The first is Sample Responses (SR), in which a set of model responses are generated — either manually or automatically.
2. Second, the *c-rater* system automatically processes model responses and students' responses using a set of Natural Language Processing (NLP) tools and extracts the linguistic features.
3. Third, a matching algorithm uses the linguistic features culminated from both SR and NLP to automatically determine whether a student's response says the same thing or implies the expected concepts.
4. And fourth, the *c-rater* system applies the scoring rules to produce a score and individualized instructional feedback that justifies the score to the student.

4 Main Processes



The *c-rater* system has been used within many domains, including biology, English, mathematics, information technology literacy, business, psychology, and physics. Assessment and learning work in tandem, in a literal sense, in the *c-rater* system.

The *e-rater*® Engine

ETS first deployed the *e-rater* automated essay evaluation and scoring engine in 1999 to provide one of two scores for essays on the writing section of the Graduate Management Admissions Test® (GMAT®).*

The *e-rater* engine predicts essay scores based on features related to writing quality, including grammar, usage, mechanics, style, organization, and development. The computational methodology underlying the system is NLP, which identifies and extracts linguistic features from stored, electronic text or speech. The engine's score predictions have been shown to be comparable to human reader scores, and its additional capabilities can automatically flag or detect off-topic responses.

* ETS developed and administered the GMAT test on behalf of the Graduate Management Admission Council from 1999 to 2004.

The *e-rater* engine has gone through many changes since its first release. The most notable changes include increased coverage of the writing construct and enhancements to the set of linguistic features extracted for use in the *e-rater* model building and scoring procedures. The ability to develop more features that are relevant to the writing construct was a direct result of advances in the field of NLP.

Using NLP methods, the *e-rater* engine identifies and extracts the following features for model building and essay scoring:

- Grammatical, word usage, or mechanical errors
- Presence and development of essay-based discourse elements
- Style weaknesses
- Statistical analysis that examines use in user essays as compared to training essays at different score points
- Two measures of essay content

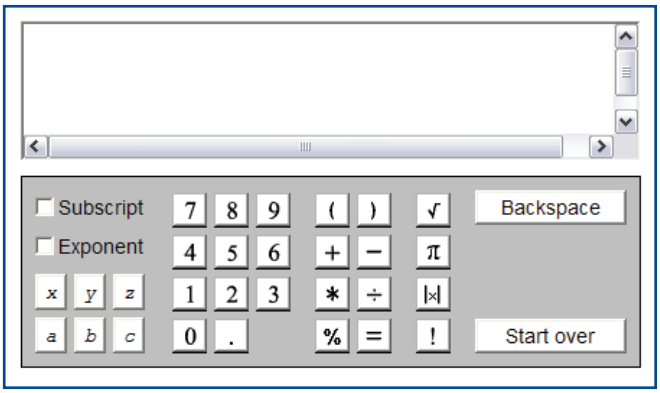
The *e-rater* engine is used in high- and low-stakes settings. In high-stakes settings, the engine is used operationally for both the Issue and Argument prompts of the Writing section of the GRE General Test, resulting in increased quality and faster score reporting. The engine is also used for the Independent prompt of the Writing section of the TOEFL iBT™ test.

In low-stakes applications, the engine is integrated into the *Criterion* Online Writing Evaluation service. This web-based, essay evaluation service is widely used as an instructional writing application in K – 12 and community college settings. Using the *e-rater* system, the *Criterion* service offers immediate, individualized feedback about errors in grammar, usage, and mechanics; the presence and absence of discourse structure elements (i.e., thesis statement, main points, supporting ideas, and conclusion statements); and style advice. The engine also provides advisories if an essay is irrelevant to the topic, has discourse structure problems, and contains disproportionately large numbers of grammatical errors (given essay length). All of the diagnostic feedback can be used by students to revise and resubmit an essay. Resubmissions receive additional feedback.

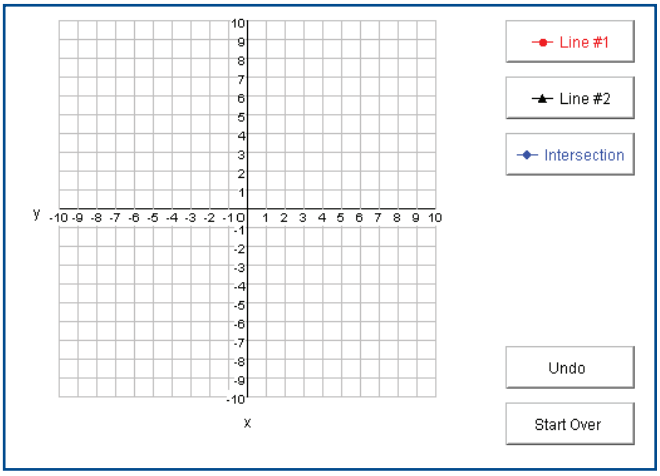
The *e-rater* engine is also the scoring engine behind test-preparation products. These include TOEFL® Practice Online and practice tests for high-stakes writing tasks, such as those that appear on the GRE® and TOEFL exams.

The *m-rater* Engine

The *m-rater* automated scoring engine scores computer-delivered constructed-response mathematics items for which a response is either a mathematical expression or equation, or a graph. When the response is an expression or equation, the *m-rater* engine is used in conjunction with ETS's Equation Editor, which allows a student to enter an equation or inequality or other expression in a standard format, with exponents, radical signs, etc. When the response is a graph, the *m-rater* engine is used in conjunction with ETS's Graph Editor, which allows the student to enter a graph consisting of one or more points, lines, broken lines, or curves. The *m-rater* engine can score responses to items in a set of items conditional on the student's responses to previous items in the set.



Equation Editor



Graph Editor

When the response is an expression or equation, the *m-rater* engine determines if the student's response is mathematically equivalent to the correct response. It is therefore not necessary, when writing an *m-rater* engine scoring model, to list all acceptable versions of the correct response. The *m-rater* engine determines mathematical equivalence by numerically evaluating the two expressions or equations at many points to be sufficiently confident that they are equivalent (or to find a counterexample that shows they are not equivalent). The *m-rater* engine randomly selects the points to be evaluated. In addition, the content specialist writing the scoring model can specify additional points to be evaluated. Research has shown that numerical evaluation has roughly the same level of accuracy as symbolic manipulation.

When the response is a graph, the student enters the graph in the Graph Editor by selecting points in the coordinate plane and selecting a button to indicate how the points are to be connected — with a straight line, with a curve, with broken line segments, or not connected at all, but left as points. The *m-rater* engine then scores the response based on the points the student selected.

Both the Equation and Graph Editors can be configured by content specialists for individual items. For example, specialists can specify the letters a student can enter as variables in an equation, or the scale and grid interval of the axes in a graph.

The *SpeechRater*℠ Engine

The *SpeechRater* engine provides automated scoring of spoken English proficiency, as demonstrated through spontaneous speaking tasks like those found on the TOEFL test. It has been used to score the TOEFL Practice Online Speaking test since 2006.

Most other automated capabilities for assessing English-learners' spoken responses are limited to tasks for which the responses are predictable, such as reading a passage aloud or repeating a sentence. The *SpeechRater* engine is not limited in this way. It can be used to score spontaneous responses, in which the range of valid responses is very broad. The engine allows the advantages of automated scoring (reliability, flexibility, reduced cost, and speed) to be applied, even to very naturalistic tasks.

How the *SpeechRater* engine works

The *SpeechRater* engine processes each response with an automated speech recognition system specially adapted for use with non-native English. Based on the output of this system, natural language processing is used to calculate a set of features that define a "profile" of the speech on a number of linguistic dimensions, including fluency, pronunciation, vocabulary usage, and prosody. A model of speaking proficiency is then applied to these features in order to assign a final score to the response. While the structure of this model is informed by content experts, it is also trained on a database of previously observed responses scored by human raters in order to ensure that the engine's scoring emulates human scoring as closely as possible.

Furthermore, if the response is found to be unscoreable due to audio quality or other issues, the *SpeechRater* engine can set it aside for special processing.

Currently, the *SpeechRater* engine uses a subset of the information used by trained human raters to score spoken responses. Because of the challenging nature of automated analysis of speech from non-native English speakers, at varying proficiency levels, many of the engine's features focus on speech delivery, rather than the higher-level aspects of language use or topic development. However, ongoing research is gradually reducing the differences between the criteria human raters use and those applied by the *SpeechRater* engine.

"The SpeechRater *engine processes each response with an automated speech recognition system specially adapted for use with non-native English."*

Using the *SpeechRater* engine in low-stakes settings

The most recent version of the *SpeechRater* engine (2.0, 2009) shows a correlation of 0.73 with human scores from the operational TOEFL test. Based on the engine's current construct coverage and agreement with human scores, it is suitable for use on assessments used to make low-stakes decisions.

The features the *SpeechRater* engine uses to establish its profile for a response are not limited to the type of items found on the TOEFL test. Since they target the underlying speaking proficiency construct, they could also be applied to other item types that address a similar construct using speaking tasks.

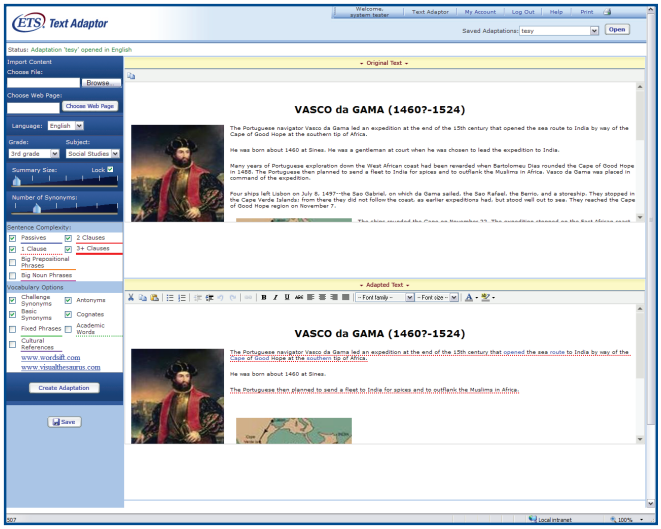
The *Text Adaptor* Tool

The *Text Adaptor* tool is a web-based, instructional support technology designed to ensure that all students, including English language learners (ELLs), have access to the content of written text. The tool's design takes into account the literature about best practices as it relates to text scaffolding and modification for non-native, English speaking (NNES) instruction. New linguistic feature development continues to be informed by recent literature and pilot studies with classroom teachers.

The tool supports the process of instructional scaffolding that ultimately facilitates student content learning, with special attention focused on NNES learners. In its current use, the *Text Adaptor* tool is part of a teacher professional development package that includes:

1) Guided teacher professional development and support related to linguistic awareness and linguistically targeted instruction

2) Lessons, activities, material/text, and assessment authoring tools to support the application of the professional development



Teachers who use the *Text Adaptor* tool are instructed about effective text modification strategies. This instruction is designed to support teachers' ability to develop text-based curriculum materials that take into account the linguistic challenges ELLs may face when working with academic texts. Once teachers are trained, the *Text Adaptor* tool can be used to support authoring and modification of any text related to instruction or assessment in the classroom.

The tool has been piloted in two university-based, teacher professional development and certification programs. Another completed pilot integrates the *Text Adaptor* tool with ETS's *Keeping Learning On Track* program. The tool continues to be integrated into other ETS programs and piloted in new programs.

In addition, the tool is currently being repurposed for test developer training, especially for those who develop content-area tests and lack training about construct-irrelevant linguistic features that often interfere with test-taker understanding among ELLs.

"Once teachers are trained, the Text Adaptor *tool can be used to support authoring and modification of any text related to instruction or assessment in the classroom."*

About ETS

At nonprofit ETS, we advance quality and equity in education for people worldwide by creating assessments based on rigorous research. ETS serves individuals, educational institutions and government agencies by providing customized solutions for teacher certification, English-language learning, and elementary, secondary and post-secondary education, as well as conducting education research, analysis and policy studies. Founded in 1947, ETS develops, administers and scores more than 50 million tests annually — including the TOEFL® and TOEIC® tests, the GRE® test and *The Praxis Series*™ assessments — in more than 180 countries, at over 9,000 locations worldwide.

Listening. Learning. Leading.®

www.ets.org