



The Journal of Technology, Learning, and Assessment

Volume 5, Number 1 · August 2006

An Overview of Automated Scoring of Essays

Semire Dikli



www.jtla.org

A publication of the Technology and Assessment Study Collaborative
Caroline A. & Peter S. Lynch School of Education, Boston College

An Overview of Automated Scoring of Essays

Semire Dikli

Editor: Michael Russell

russelmh@bc.edu

Technology and Assessment Study Collaborative

Lynch School of Education, Boston College

Chestnut Hill, MA 02467

Copy Editor: Kevon R. Tucker-Seeley

Design: Thomas Hoffmann

Layout: Aimee Levy

JTLA is a free on-line journal, published by the Technology and Assessment Study Collaborative, Caroline A. & Peter S. Lynch School of Education, Boston College.

Copyright ©2006 by the Journal of Technology, Learning, and Assessment (ISSN 1540-2525).

Permission is hereby granted to copy any article provided that the Journal of Technology, Learning, and Assessment is credited and copies are not sold.

Preferred citation:

Dikli, S. (2006). An Overview of Automated Scoring of Essays. *Journal of Technology, Learning, and Assessment*, 5(1). Retrieved [date] from <http://www.jtla.org>.

Abstract:

Automated Essay Scoring (AES) is defined as the computer technology that evaluates and scores the written prose (Shermis & Barrera, 2002; Shermis & Burstein, 2003; Shermis, Raymat, & Barrera, 2003). AES systems are mainly used to overcome time, cost, reliability, and generalizability issues in writing assessment (Bereiter, 2003; Burstein, 2003; Chung & O'Neil, 1997; Hamp-Lyons, 2001; Myers, 2003; Page, 2003; Rudner & Gagne, 2001; Rudner & Liang, 2002; Sireci & Rizavi, 1999). AES continues attracting the attention of public schools, universities, testing companies, researchers and educators (Burstein, Kukich, Wolff, Lu, & Chodorow, 1998; Shermis & Burstein, 2003; Sireci & Rizavi, 1999). The main purpose of this article is to provide an overview of current approaches to AES. The article will describe the most widely used AES systems including Project Essay Grader™ (PEG), Intelligent Essay Assessor™ (IEA), E-rater® and Criterion™, IntelliMetric™ and MY Access!®, and Bayesian Essay Test Scoring System™ (BETSY). It will also discuss the main characteristics of these systems and current issues regarding the use of them both in low-stakes assessment (in classrooms) and high-stakes assessment (as standardized tests).

An Overview of Automated Scoring of Essays

Semire Dikli

Introduction

Automated Essay Scoring (AES) is defined as the computer technology that evaluates and scores the written prose (Shermis & Barrera, 2002; Shermis & Burstein, 2003; Shermis, Raymat, & Barrera, 2003). AES systems are developed to assist teachers in low-stakes classroom assessment as well as testing companies and states in large-scale high-stakes assessment. They are mainly used to help overcome time, cost, reliability, and generalizability issues in writing assessment (Bereiter, 2003; Burstein, 2003; Chung & O'Neil, 1997; Hamp-Lyons, 2001; Myers, 2003; Page, 2003; Rudner & Gagne, 2001; Rudner & Liang, 2002; Sireci & Rizavi, 1999).

A number of studies have been conducted to assess the accuracy and reliability of the AES systems with respect to writing assessment. The results of several AES studies reported high agreement rates between AES systems and human raters (Attali, 2004; Burstein & Chodorow, 1999; Landauer, Laham, & Foltz, 2003; Landauer, Laham, Rehder, & Schreiner, 1997; Nichols, 2004; Page, 2003; Vantage Learning, 2000a, 2000b, 2001b, 2002, 2003a, 2003b).

AES systems have been criticized for lacking human interaction (Hamp-Lyons, 2001), vulnerability to cheating (Chung & O'Neil, 1997; Kukich, 2000; Rudner & Gagne, 2001), and their need for a large corpus of sample text to train the system (Chung & O'Neil, 1997). Despite its weaknesses, AES continues attracting the attention of public schools, universities, testing companies, researchers and educators (Burstein, Kukich, Wolff, Lu, & Chodorow, 1998; Shermis & Burstein, 2003; Sireci & Rizavi, 1999).

The purpose of this article is to provide an overview of current approaches to automated essay scoring. The next section will describe the most widely used AES systems: Project Essay Grader™ (PEG), Intelligent Essay Assessor™ (IEA), E-rater® and Criterion™, IntelliMetric™ and MY Access!®, and Bayesian Essay Test Scoring System™ (BETSY). The final section will summarize the main characteristics of AES systems and will discuss current issues regarding the use of these systems both in low-stakes assessment (in classrooms) and high-stakes assessment (as standardized tests).

Automated Essay Scoring Systems

Project Essay Grader™ (PEG)

Project Essay Grader™ (PEG) was developed by Ellis Page in 1966 upon the request of the College Board, which wanted to make the large-scale essay scoring process more practical and effective (Rudner & Gagne, 2001; Page, 2003). PEG™ uses correlation to predict the intrinsic quality of the essays (Chung & O'Neil, 1997; Kukich, 2000; Rudner & Gagne, 2001).

Page and his colleagues use the terms *trins* and *proxes* while explaining the way PEG™ generates a score. While *trins* refer to the intrinsic variables such as fluency, diction, grammar, punctuation, etc., *proxes* denote the approximation (correlation) of the intrinsic variables. Thus, *proxes* refer to actual counts in an essay (e.g., establishing the correlation of fluency or *trin* with the amount of vocabulary or *prox*; Page, 1994).

The scoring methodology PEG™ employs is simple. The system contains a training stage and a scoring stage. PEG™ is trained on a sample of essays in the former stage. In the latter stage, proxy variables (*proxes*) are determined for each essay and these variables are entered into the prediction equation. Finally, a score is assigned by computing beta weights (coefficients) from the training stage (Chung & O'Neil, 1997). PEG™ needs 100 to 400 sample essays for training purposes (BETSY, n.d.).

One of the strengths of PEG™ is that the predicted scores are comparable to those of human raters. Furthermore, the system can computationally track the writing errors made by the users (Chung & O'Neil, 1997). However, PEG™ has been criticized for ignoring the semantic aspect of essays and focusing more on the surface structures (Chung & O'Neil, 1997; Kukich, 2000). By failing to detect the content related features of an essay (organization, style etc.), the system does not provide instructional feedback to students. An early version was found to be weak in terms of scoring accuracy. For example, since PEG™ used indirect measures of writing skill, it was possible to “trick” the system by writing longer essays (Kukich, 2000). PEG™ went through changes in 1990s (Kukich, 2000) and several aspects of PEG™ were modified including not only several parsers and various dictionaries, but also special collections and classification schemes (Page, 2003; Shermis & Barrera, 2002).

Intelligent Essay Assessor™ (IEA)

Another AES system, Intelligent Essay Assessor™ (IEA), analyzes and scores an essay using a semantic text-analysis method called *Latent Semantic Analysis* (LSA; Lemaire & Dessus, 2001), which is an approach created by psychologist Thomas Landauer with the assistance of Peter Foltz and Darrell Laham (Murray, 1998). IEA™ is produced by the Pearson Knowledge Analysis Technologies (PKT; Psotka & Streeter, 2004, p.2; PKT, n.d.). More detailed descriptions of LSA and IEA™ are provided below.

Latent Semantic Analysis

Latent Semantic Analysis (LSA) is defined as “a statistical model of word usage that permits comparisons of the semantic similarity between pieces of textual information” (Foltz, 1996, p. 2). LSA first processes a corpus of machine-readable language and then represents the words that are included in a sentence, paragraph, or essay through statistical computations (Landauer, Laham, & Foltz, 1998). LSA measures of similarity are considered highly correlated with human meaning similarities among words and texts. Moreover, it successfully imitates human word selection and category judgments (Landauer, Laham, & Foltz, 2003). The underlying idea is that the meaning of a passage is very much dependent on its words and changing even only one word can result in meaning differences in the passage. On the other hand, two passages with different words might have a very similar meaning (Landauer et al., 2003). The underlying idea can be summarized as:

“meaning of word₁ + meaning of word₂ + + meaning of word_k = meaning of passage” (Landauer et al., 2003, p. 88).

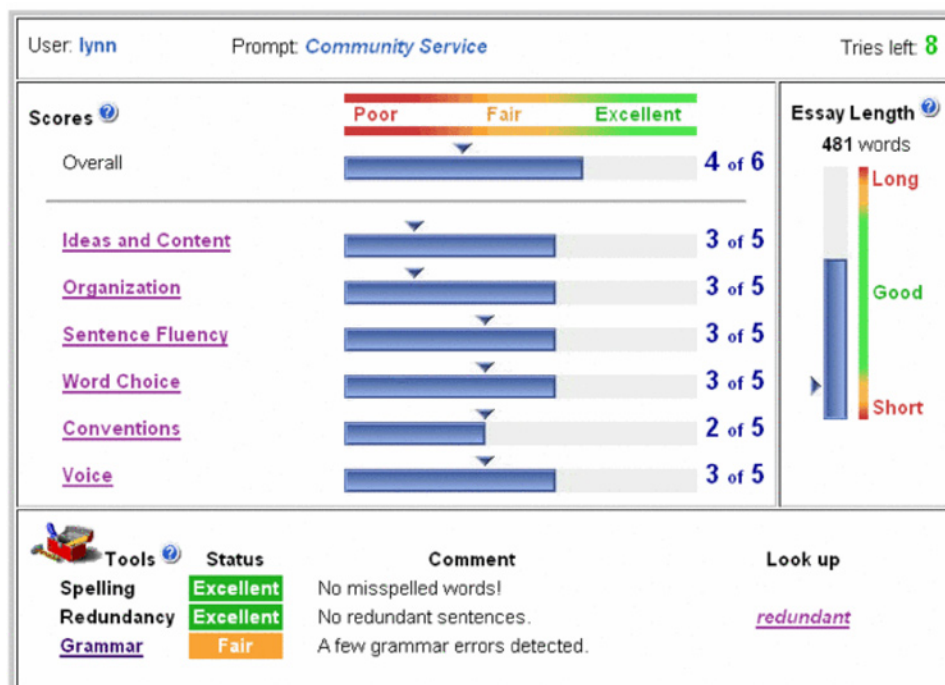
The educational applications of LSA include picking the most suitable text for students with different levels of background knowledge, automatic scoring of essay contents, and assisting students in summarizing texts successfully (Landauer et al., 1998). In order to evaluate the overall quality of an essay, LSA needs to be trained on domain-representative texts (texts that best represent the writing prompt). Then the essay needs to be characterized by LSA vectors (a mathematical representation of the essay). Finally, the conceptual relevance and the content of the essay are compared to other texts (Foltz, Laham, & Landauer, 1999; Landauer et al., 1998).

In the LSA based approach, the text is represented as a matrix. Each row in the matrix represents a unique word, while each column represents context. Each cell involves the frequency of the word. Then, each cell frequency is considered by a feature that denotes not only the importance of the word in that context but also the degree to which the word type carries information in the domain discourse (Landauer et al., 1998). The

semantics of a word are verified through all the contexts in which the word occurs. The number of occurrences of each word in a text determines its semantic space. For example, 300 paragraphs and 2000 words provide a 300x2000 matrix. Here, while each word is represented by a 300-dimensional vector, each paragraph is represented by a 2000-dimensional vector. By reducing these dimensions, LSA induces semantic similarities between words. This reduction is critical since it permits the representation of word meanings through the context in which they occur. The number of dimensions is also crucial. That is, if the number is too small, much of the information will be lost. On the contrary, if the number is too big, limited dependencies will be drawn between vectors. According to this method, the semantic information is determined only through the co-occurrence of words in a large corpus of texts (Lemaire & Dessus, 2001).

Intelligent Essay Assessor™ (IEA)

Unlike other AES systems, IEA™'s main focus is more on the content related features rather than the form related ones; however, this does not mean that IEA™ provides no feedback on formal aspects (e.g., grammar and punctuation) in an essay. In other words, even though the system uses an LSA-based approach to evaluate mainly the quality of the content of an essay, it also includes scoring and feedback on grammar, style and mechanics (Landauer, Laham, & Foltz, 2000; Landauer, Laham, & Foltz, 2003; Streeter, Psotka, Laham, & MacCuish, 2004). Figure 1 (next page) shows an example of the feedback feature provided by IEA™ (PKT, n.d.).

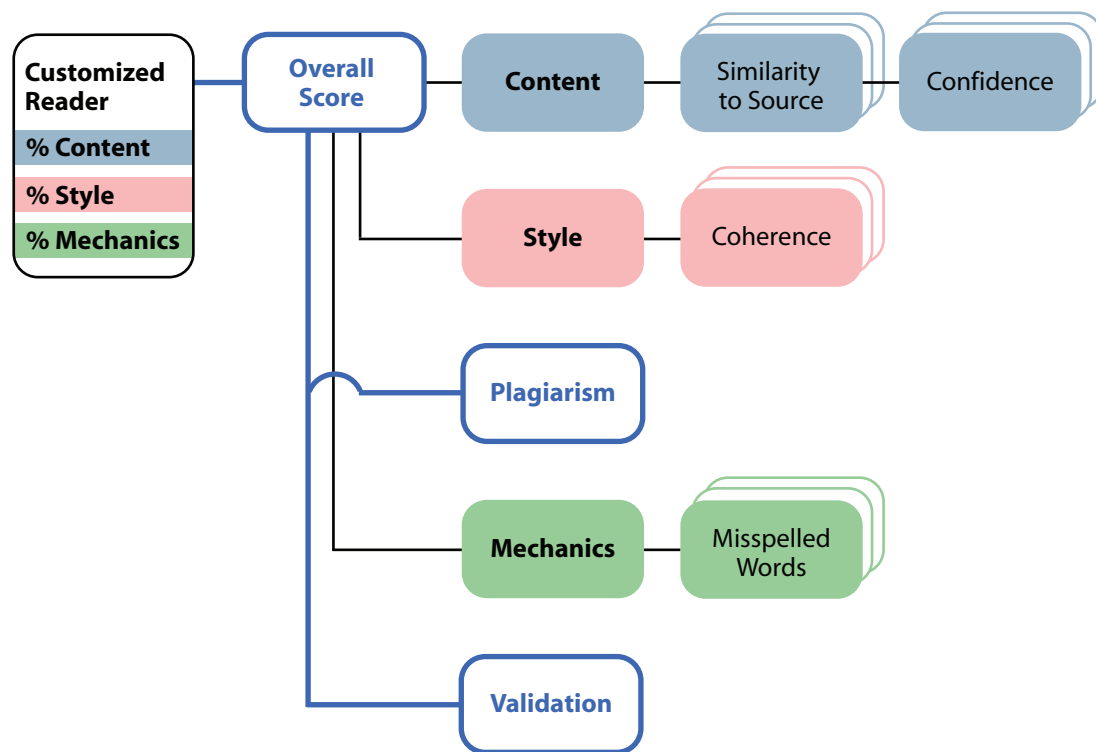
Figure 1: Sample Feedback Screen of IEA™

Landauer et al. (2003) claim that IEA™ can successfully analyze not only content-based essays, but also creative narratives. The system needs to be trained on a set of domain-representative texts in order to measure the overall quality of an essay. For example, a biology book can be used to evaluate a biology essay. IEA™ uses three sources to analyze an essay: “a) pre-scored essays of other students, b) expert model essays and knowledge source materials, c) internal comparison of an unscored set of essays” (Landauer et al., 2003, p. 90). This approach allows IEA™ to compare each essay with similar texts in terms of the content quality (Landauer et al., 2000; Landauer et al., 2003; Streeter et al., 2004). First, IEA™ compares content similarity between a student’s essay and other essays on the same topic scored by human raters to determine how closely they match (Landauer et al., 2000; Rudner & Gagne, 2001; Streeter et al., 2004). It then predicts the overall score by adding a “corpus-statistical writing-style” and mechanics (Landauer et al., 2000, p. 28). It also spots plagiarism and provides feedback (Landauer et al., 2000; Landauer et al., 2003).

As part of the usual procedure of IEA™, each essay is compared to every other one in a set. The essays that are extremely similar to each other are examined by LSA. Regardless of substitution of synonym, paraphrasing, or rearrangement of sentences, the two essays will be similar with LSA (Landauer et al., 2003). Detecting plagiarism is an essential feature since this type of academic dishonesty is quite hard to detect by human raters,

particularly when grading large number of essays (Shermis, Raymat, & Barrera, 2003). The structure of IEA™ is presented in Figure 2 (Landauer et al., 2003, p.90).

Figure 2: The Intelligent Essay Assessor™ Architecture



Landauer et al. (2000) point out the basic technical difference between IEA™ and other AES systems as follows:

Other systems work primarily by finding essay features they can count and correlate with ratings human graders 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 (p.28).

The producers of IEA™, Pearson Knowledge Technologies (PKT), report that the system needs smaller numbers of pre-scored essays to train. Unlike other AES systems, which require 300-500 training essays per prompt, IEA™ only requires 100 pre-scored essays (PKT, n.d.; Landauer et al., 2003). PKT claims that the system does not evaluate creativity and reflective thinking. It does, however, assess “expository essays on factual

topics” such as description of a psychological theory or function of the heart (Murray, 1998). IEA™’s future plans include moving from global assessment features, such as flow and coherence, to more specific ones such as the voice and audience (Landauer et al., 2003).

E-rater® and CriterionSM

The electronic essay rater (e-rater®) was developed by the Educational Testing Service (ETS) to evaluate the quality of an essay by identifying linguistic features in the text (Burstein, 2003; Burstein & Marcu, 2000). E-rater® uses natural-language processing (NLP) techniques, which identify specific lexical and syntactical cues in a text to analyze essays (Burstein, 2003; Kukich, 2000). A detailed description of NLP and information regarding the structure and functions of e-rater and CriterionSM are provided below.

Artificial Intelligence (AI) and Natural Language Processing (NLP)

Artificial intelligence (AI) is defined as the science of making intelligent machines. AI has several applications including game playing, speech recognition, understanding natural language processing, computer vision, and so on¹.

NLP is considered to be one of the most challenging areas of AI. The research in NLP comprises a variety of fields including corpus-based methods, discourse methods, formal models, machine translation, natural language generation, and spoken-language understanding (Salem, 2000). There have been several empirical methods used in NLP. Previous methods (e.g., rationalist methods) required manual encoding of linguistic knowledge, which has proven to be difficult due to the complex nature of human language. Recent methods (e.g., empirical methods), however, employ techniques that automatically extract linguistic knowledge from large-text corpora. In other words, empirical methods employ statistical or machine learning techniques to train the system on large amounts of authentic language data (Brill & Mooney, 1997).

NLP is claimed to be a complex task to comprehend because it contains several levels of processing and subtasks. It has four categories of language tasks including *speech recognition*, *syntactic analysis*, *discourse analysis*, *information extraction*, and *machine translation*. Speech recognition focuses on diagramming a continuous speech signal into a sequence of known words. Syntactic analysis, on the other hand, determines the ways words are clustered into components like noun- and verb-phrases. Semantic analysis involves diagramming a sentence to a type of meaning representation such as a logical expression. Whereas discourse analysis focuses on how context impacts sentence interpretation and information

extraction locates specific pieces of data from a natural language document. Finally, the task of machine translation is to translate text from one natural language to another such as English to German or vice versa (Brill & Mooney, 1997).

E-rater®

E-rater® was initially used by ETS for operational scoring of the Graduate Management Admissions Test Analytical Writing Assessment (GMAT AWA; Burstein, 2003; Burstein & Chodorow, 1999; Burstein & Marcu, 2000) and had been employed for scoring the AWA since February, 1999 (Burstein, 2003). However, as of January, 2006, ACT, Inc. started scoring GMAT essays using IntelliMetric™, which is Vantage Learning's automated essay scoring engine (Rudner, Garcia, & Welch, RR-05-08). The GMAT AWA is currently scored by two human raters on a 6-point holistic scale, with 6 being the highest score and 1 the lowest. If two raters differ by more than 1 point, a third rater is called for resolution (Burstein, 2003; Burstein & Chodorow, 1999). The test-taker's final score is determined through e-rater and one human-scorer. Similar to the prior practice with human raters, if there is a discrepancy between e-rater and the human rater by more than 1 point, a second human rater is included (Burstein, 2003). To date, ACT, Inc. continues to use to IntelliMetric™ scoring procedure (Rudner, Garcia, & Welch, 2005).

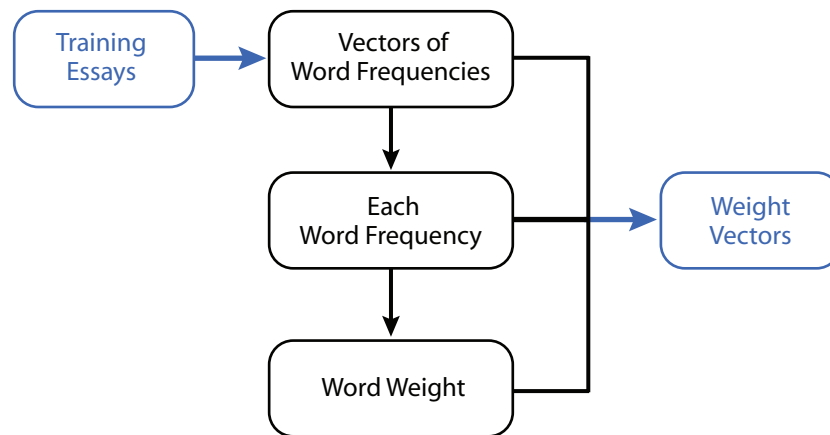
E-rater® employs a corpus-based approach to model building, in which actual essay data are used to examine sample essays. A corpus-based approach of building NLP-based tools requires researchers to usually use copy-edited text sources like newspapers. However, e-rater®'s feature analysis and model building require unedited text corpora that represent the particular genre of first-draft student essays (Burstein, 2003; Burstein, Leacock, & Swartz, 2001).

The features of e-rater® include a syntactic module, a discourse module, and a topical-analysis module. These modules provide outputs for model building and scoring. E-rater® has been trained on a set of essays scored by at least two human raters on a 6-point holistic scale to build models (Burstein, 2003; Burstein & Chodorow, 1999; Burstein et al., 2003; Burstein & Marcu, 2000). The origin of the syntactic module is parsing. In order to capture syntactic variety in an essay, "a parser identifies syntactic structures, such as subjunctive auxiliary verbs and a variety of clausal structures, such as complement, infinitive, and subordinate clauses" (Burstein, Chodorow, & Leacock, 2003, p. 1). The discourse module uses a conceptual framework of conjunctive relations including cue words (e.g., using words like "perhaps" or "possibly" to express a belief), terms (e.g., using conjuncts such as "in summary" and "in conclusion" for summarizing), and syntactic structures (e.g., using complement clauses to

identify the beginning of a new argument) to identify discourse-based relationship and organization in essays (Burstein, 2003; Burstein & Chodorow, 1999; Burstein et al., 2003; Burstein & Marcu, 2000; Burstein, Kukich, Wolff, Lu, & Chodorow, 1998). Finally, the topical analysis module identifies vocabulary usage and topical content (Burstein, 2003; Burstein et al., 2003; Burstein & Marcu, 2000). Unlike a poor essay, a good essay needs to be relevant to the topic assigned. Moreover, the variety and type of vocabulary used in good essays differ from that of poor essays. The assumptions behind this module are that good essays resemble other good essays. A similar assumption is valid for poor essays, as well (Burstein & Chodorow, 1999; Burstein et al., 1998). The general procedure for a vector-spec model (Salton, as cited in Burstein & Marcu, 2000), which is used to capture the topic or vocabulary usage (Burstein & Chodorow, 1999; Burstein et al., 2003; Burstein et al., 1998; Burstein & Marcu, 2000), is described as follows:

...training essays are converted into vectors of word frequencies, and the frequencies are then transformed into word weights. These weight vectors populate the training space. To score a test essay, it is converted into a weight vector, and a search is conducted to find the training vectors most similar to it, as measured by the cosine between the test and training vectors. The closest matches among the training set are used to assign a score to the test essay (Burstein, 2003, p. 117).

Here, a vector can be described as the mathematical representation of an essay. Moreover, word frequencies can be calculated by counting the words in a paragraph and dividing by the number of their occurrence (each time a word appeared in a paragraph). While word weights refer to the frequency divided by the number of words in an essay, training space refers to the entire set of vectors that were generated from the training essays. Finally, in this context cosine is the distance between test- and training-vectors. Figure 3 provides the graphic representation regarding the transformation of training essays into vectors of word frequencies, then into each word frequency, and finally into word weight. In short, training essays are converted into vectors of word frequencies and then into weight vectors.

Figure 3: Transformation of Training Essays into Vectors

To summarize, e-rater uses NLP to identify the features of the faculty-scored essays in its sample collection and store them-with their associated weights-in a database. When e-rater evaluates a new essay, it compares its features to those in the database in order to assign a score. Because e-rater is not doing any actual reading, the validity of its scoring depends on the scoring of the sample essays from which e-rater's database is created (Educational Testing Service, n.d.).

CriterionSM

CriterionSM is a web-based essay scoring and evaluating system, which relies on other ETS technologies called *e-rater*[®] and *Critique* writing analysis tools. As discussed in detail above, e-rater is an automated essay scoring system. As a writing analysis tool, Critique includes a group of programs that identify errors in grammar, usage, and mechanics and that recognize discourse elements and elements of undesirable style in an essay. Besides providing instant holistic scoring, CriterionSM also gives individualized diagnostic feedback based on the types of evaluations that teachers give when responding to student writing (Burstein et al., 2003). The feedback component of CriterionSM is called an *advisory component*. The advisory component functions as a supplement to the e-rater score, but does not determine the score (Burstein, 2003). The feedback types that the advisory component contains are as follows:

- a. The text is too brief to be a complete essay (suggesting that student write more).
- b. The essay text does not resemble other essays written about the topic (implying that perhaps the essay is off-topic).
- c. The essay response is overly repetitive (suggesting that the student use more synonyms) (Burstein, 2003, p. 119).

Along with holistic scoring, CriterionSM provides diagnostic feedback on grammar, usage, and mechanics; style and diction; and organization and development. CriterionSM covers a number of writing genres including persuasive, descriptive, narrative, expository, cause and effect, comparison and contrast, problem and solution, argumentative, issue, response to literature, workplace writing, and writing for assessment. It provides writing topics at various levels including elementary school (4th and 5th grades), middle school (6th, 7th, and 8th grades), high school (9th, 10th, 11th, and 12th grades), college (1st year/placement and 2nd year), upper division or graduate school (Graduate Record Examination® (GRE)), and non-native speakers of English (Test of English as a Foreign Language® (TOEFL)). The topics are taken from authentic retired ETS essay topics. They are obtained from various ETS testing instruments such as NAEPTM (National Assessment of Educational Progress)², English Placement Test designed for California State University³, Praxis^{TM4}, and TOEFL^{®5}. CriterionSM is capable of analyzing essays on the topics for which it has been “trained.” A minimum of 465 essays scored by expert raters are required to train the system on a topic. However, teachers are not limited to use the topics in the CriterionSM library and they can create and assign their own topics. While holistic scoring can not be reported for teacher-created topics, it is possible to obtain feedback of every dimension of writing (ETS, n.d.).

The electronic portfolio and writer’s handbook features aim to facilitate the writing process for the students. The electronic portfolio allows students to store their first and subsequent drafts online. Writer’s handbook, on the other hand, provides students with opportunities to view feedback definitions, examples of correct and incorrect use, and an explanation of every error reported. Teachers have power over several features of CriterionSM. They can manage student access to the program by activating/inactivating the website or setting start/finish dates. Teachers can also control the student access to spell check, diagnostic feedback, or holistic scoring by turning on/off these features. Finally, teachers have an option to insert their own feedback within the student essay (ETS, n.d.).

Besides its instructional use in classrooms, CriterionSM can also be used for remediation and placement purposes by the schools. Some schools use CriterionSM for benchmark testing. Some schools use the CriterionSM program for exit testing. In this case, both CriterionSM and a faculty reader assign a score to the given essay. If the difference between two scores is more than one point a third rater is included in the scoring process (ETS, n.d.).

IntelliMetric™ and MY Access!®

IntelliMetric™, an AES system developed by Vantage Learning, is known as the first essay-scoring tool that was based on artificial intelligence (AI) (Elliott, 2003; Shermis & Barrera, 2002; Shermis, Raymat, & Barrera, 2003). Like e-rater®, IntelliMetric™ relies on NLP. See the section about NLP above for more information. IntelliMetric™ was developed by Vantage learning and used by the College Board for placement purposes (Myers, 2003). MY Access!® is known as the instructional application of IntelliMetric™ (Vantage Learning, n.d.). More information about the structure and functions of IntelliMetric™ and MY Access!® is provided below.

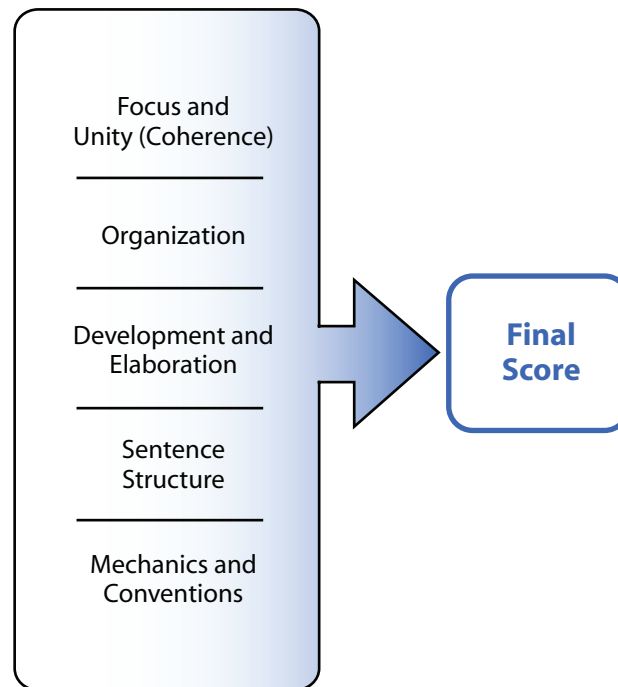
IntelliMetric™

Using a blend of artificial intelligence (AI), natural language processing (NLP), and statistical technologies, IntelliMetric™ is a type of learning engine that internalizes the “pooled wisdom” of expert human raters (Elliot, 2003, p. 71). As an advanced AI application for scoring essays, IntelliMetric™ relies on Vantage Learning’s CogniSearch™ and Quantum Reasoning™ technologies (Elliott, 2003; Shermis & Barrera, 2002; Shermis et al., 2003; Vantage Learning, 2001a, 2003c). CogniSearch™ is a system specifically developed for use with IntelliMetric™ to understand natural language to support essay scoring. For instance, it parses the text to analyze the parts of speech and their syntactical relations with one another. This process assists IntelliMetric™ to examine the essay according to the main characteristics of standard written English (Vantage Learning, 2003c). CogniSearch™ and Quantum Reasoning™ technologies together allow IntelliMetric™ to internalize each score point associated with certain characteristics in an essay response and then apply it to subsequent scoring by the system (Elliott, 2003; Shermis & Barrera, 2002; Shermis et al., 2003; Vantage Learning, 2001a). This approach is claimed to be consistent with the procedure underlying holistic scoring (Elliot, 2003). It is also claimed that the scoring system “learns” the characteristics that human raters likely to value and those they find poor (Shermis & Barrera, 2002; Shermis et al., 2003).

IntelliMetric™ needs to be trained with a set of pre-scored essays with known scores assigned by human raters. These essays are then used as a foundation to extract the scoring scale and the wisdom of the human raters (Vantage Learning, 2001a, 2002, 2003b, 2003c). The system employs multiple steps to analyze essays. First, the system internalizes the known scores in a set of training essays. In other words, the system infers the writing rubric and the essay features associated with each score. The second step includes testing the scoring model against a smaller set of

essays with known scores for validation purposes. Finally, once the model scores the essays as desired, it is applied to new essays with unknown scores (Shermis & Barrera, 2002; Vantage Learning, 2000b, 2003c).

IntelliMetric™ evaluates over 300 semantic-, syntactic-, and discourse-related features in an essay by using AI and NLP technologies (Elliott, 2003; Vantage Learning, 2001a). These text related features are identified as larger categories called Latent Semantic Dimensions (LSD) (Vantage Learning, 2003a). The LSD features are described in five broad categories. The first category, *focus and unity (focus and coherence)*, uses the features that emphasizes a single point of view, cohesiveness and consistency in purpose, and main ideas in an essay. The *organization* category analyzes transitional fluency and logic of discourse. Examples include the introduction and conclusion, coordination and subordination, logical structure, logical transitions, and the sequence of ideas in an essay. The third category, *development and elaboration*, examines the breadth of the content and the supporting ideas in an essay (e.g., vocabulary, elaboration, word choice, concepts, and support). The fourth category, *sentence structure*, focuses on sentence complexity and variety such as syntactic variety, sentence complexity, usage, readability, and subject-verb agreement. The fifth and final category of *mechanics and conventions* analyze whether the essay includes the conventions of standard American English such as grammar, spelling, capitalization, sentence completeness, and punctuation (Elliott, 2003; Vantage Learning, 2001a, 2003a). Figure 4 (next page) displays the *IntelliMetric™ Feature Model*⁶.

Figure 4: IntelliMetric™ Feature Model

There are five key principles underlying the IntelliMetric™ system. First, IntelliMetric™ is modeled on the human brain. IntelliMetric™ “emulates the way in which the human brain acquires, stores, accesses and uses information” (Vantage Learning, 2003c, p. 5). Therefore, a *neurosynthetic* (*neuro* = brain and *synthetic* = artificially created) approach is used to duplicate the mental processes employed by the human expert raters. Second, IntelliMetric™ is considered to be a learning engine that obtains the necessary information by learning ways to examine the sample pre-scored essays by expert raters. In other words, by modeling the scoring process used by expert human raters, IntelliMetric™ learns the rubric and the essential characteristics for scoring an essay as well as the ways those characteristics are revealed in each score point. Its “error reduction function” allows IntelliMetric™ to increase its accuracy over time by detecting and “learning from” its mistakes. Third, IntelliMetric™ is systemic and based on a complex system of information processing. Another principle suggests that IntelliMetric™ is inductive. Its judgments are based on inductive reasoning and it makes inferences about how to analyze an essay based on the sample responses previously evaluated by expert human raters. Finally, IntelliMetric™ is multidimensional and non-linear. It employs multiple judgments that rely on multiple mathematical models. It is claimed that while many scoring systems are based on the *General Linear Model* (GLM), IntelliMetric™ uses a nonlinear and multidimensional approach to analyze

essays. It is also claimed that the writing process is more complex than the General Linear Model's simplistic, approach which suggests that an essay score increases as the values of text features increase and vice versa (Vantage Learning, 2003c).

One of the best attributes of IntelliMetric™ is that it is capable of evaluating essay responses in multiple languages including English, Spanish, Hebrew, Bahasa, Dutch, French, Portuguese, German, Italian, Arabic, and Japanese (Elliot, 2003). The system could be applied in "Instructional" or "Standardized Assessment" modes. The instructional mode assists students with revising and editing processes by providing holistic and diagnostic feedback on five traits (see MY Access!® section below for more information). The Standardized Assessment mode provides a holistic score and feedback on various rhetorical and analytical dimensions of an essay as well as detailed diagnostic feedback on grammar, usage, spelling and conventions, if necessary (Elliott, 2003; Vantage Learning, 2001a, 2003c).

MY Access!®

MY Access!® is a web-based writing assessment tool that relies on Vantage Learning's IntelliMetric™ automated essay scoring system. The main purpose of the program is to offer students a writing environment that provides immediate scoring and diagnostic feedback; that allows them to revise their essays accordingly; and that motivates them to continue writing on the topic to improve their writing proficiency (Vantage Learning, n.d.).

MY Access!® not only provides immediate diagnostic assessment of writing, but also constructive multilingual feedback for ELL learners in grades K–12. Currently, the system assigns essay topics and provides feedback in English, Spanish, or Chinese. However, the company plans to make this opportunity available for other languages in the future as well. Students have two options in using the MY Access!® program. One option is writing on a topic assigned in English, Spanish, or Chinese and receiving feedback in the same language. Another option is writing an essay in English and receiving feedback either in the native language or in English. Besides providing multilingual feedback, MY Access!® provides multilevel feedback – developing, proficient, and advanced – as well. The multilingual dictionary, thesaurus, and translator functions of the program allow students to receive definitions as well as synonyms of a specific word (Vantage Learning, n.d.).

MY Access!® includes several features that aim to make the writing process more feasible and effective not only for students, but also for teachers. To begin with, the program provides a web-based environment to

the user. Second, MY Access!® relies on the IntelliMetric™ scoring system and is able to provide instant feedback and scoring for an essay. This feature not only ensures consistency and accuracy among teachers as well as schools, it also gives teachers more time to focus on instruction. Also, the analytic scoring and feedback on each of the five categories provides diagnostic feedback regarding the student's writing ability. The program can provide individualized multilingual feedback (Spanish and Chinese) on different genres of writing such as informative, narrative, literary, and persuasive essays (Vantage Learning, n.d.).

MY Access!® contains over 200 operational and pilot prompts that generate instant analysis of the essay. These prompts are based on reading texts as well as literature at grade levels and they are available in following academic levels: higher education (level 4), high school (level 3), middle school (level 2), and upper elementary (level 1). Teachers can provide their own prompts, as well, bearing in mind that the system will be unable to score their students' essays because it first needs to be trained on about 300 prompts to be able to score essays automatically. MY Access!® also offers a variety of writing tools such as *writing dashboard* and *my portfolio*, which aim to facilitate the essay writing process for students. The writing dashboard feature gives students the opportunity to see their weekly progress and the my portfolio feature allows students to view a list of completed assignments, scores, reports, comments, and so on (Vantage Learning, n.d.).

Various teacher options allow teachers to have full control of the application of the program. For instance, teachers are able to create groups or customize the level as well as the type of feedback according to the proficiency level of the students. Moreover, teachers can add their own comments on student essays along with the feedback provided by the system. The *view reports* option allows teachers to generate up to ten different types of reports on their students' progress. For instance, *the student history report* provides teachers with not only an analysis of errors based on the rule categories in the system, but also with the average performance assessments of students over time. Last but not least, the MY Access!® website includes parent letters in English, Spanish, and Chinese to enable teachers to provide parents an opportunity to get involved in their children's learning process (Vantage Learning, n.d.).

Bayesian Essay Test Scoring sYstem™ (BETSY)

The final automated essay scoring system to be discussed in this article is the Bayesian Essay Test Scoring sYstem or BETSY™, which was developed by Lawrence M. Rudner⁷. BETSY™ is not of the same ilk as the commercial AES products described above and therefore should be treated more as a research tool. A more detailed discussion about BETSY™ is presented below, following a brief overview of the Bayesian approach to AES.

Bayesian approach

Another approach used in AES employs Bayesian theorem. Bayesian methods have several applications such as identifying spam and other unwanted e-mails based on their similarity with previously classified e-mail, and sorting the resumes of job applicants into various job categories according to their similarity to previously classified resumes (BETSY, n.d.). Several Microsoft products such as *Answer Wizard* of Office 95®, the *Office Assistant* of Office 97®, and numerous technical *troubleshooters* are other applications of the Bayesian approach (Rudner & Liang, 2002).

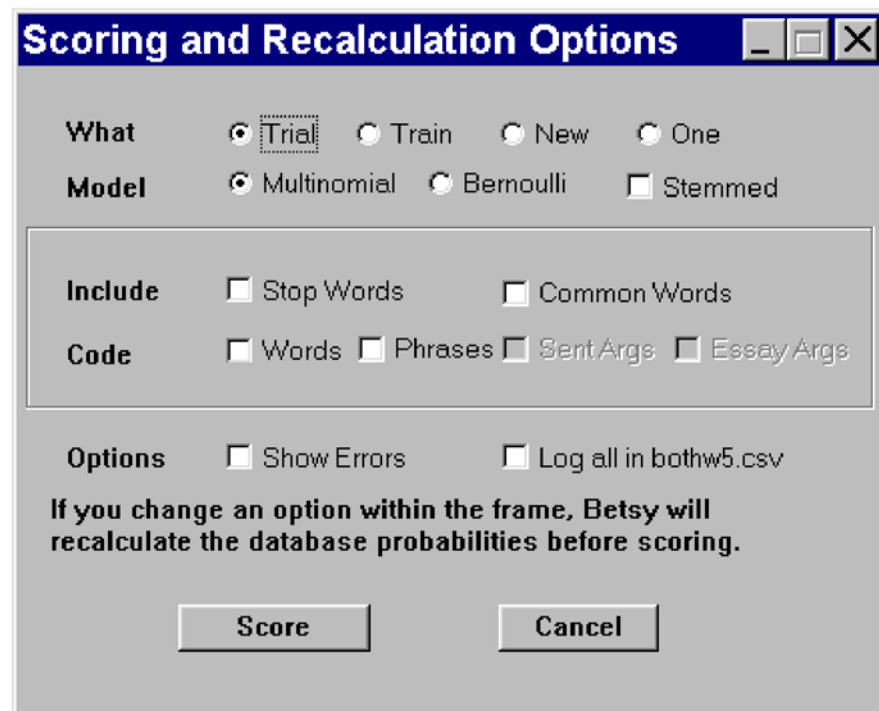
There are two Bayesian models widely used in text classification: the *Multivariate Bernoulli Model* and the *Multinomial Model*. While the former views each essay as a special case of calibrated features, the latter views each essay as a sample of calibrated features. In the Bernoulli model, the conditional probability of presence of a specific feature is estimated by the proportion of essays within each category that include the feature. In Multinomial model, on the other hand, the probability of each score for a given essay is computed as the product of the probabilities of the features included in the essay. (BETSY, n.d.; Rudner & Liang, 2002). To summarize, the Bernoulli model investigates whether a specific feature exists in an essay or not, whereas the Multinomial model checks the multiple use of a specific feature in an essay (Rudner & Liang, 2002). The Bernoulli model computes relatively slowly compared to the Multinomial model (BETSY, n.d.).

The Bayesian approach includes key concepts such as *stemming*, *stop words*, and *feature selection*. Stemming denotes the process of eliminating suffixes to get stems. For example, obtaining “educ” as a stem for educate, education, educates, educational, and educated. Stop words refer to various articles, pronouns, adjectives, and prepositions. Search engines do not list these types of words because they can cause large number of irrelevant results. One approach to feature selection is the reduction in *entropy*. By minimizing entropy, it is possible to pick the items with maximum potential information gain (Rudner & Liang, 2002).

BETSY™

The underlying idea of BETSY™ is the classification of texts based on trained materials (Valenti, Neri, & Cucchiarelli, 2003). Rudner and Liang (2002) point out that the classification of Bayesian *Computer Adaptive Testing* (CAT) is extended from two categories (master/non-master) to a three- or four-point nominal or categorical scale (e.g. extensive, essential, partial, unsatisfactory) in BETSY™. While the Bayesian CAT classification is based on optimally selected items, BETSY™ uses a large set of items. The “items” refer to a large set of essay features. These essay features include content related features such as specific words and phrases, frequency of certain content words, form related features including number of words, sentence length, number of verbs, number of commas and others, e.g., the order certain concepts are presented and the occurrence of specific noun-verb pairs. (BETSY, n.d.; Rudner & Liang, 2002).

BETSY™ needs to be trained on 1000 texts (Rudner, Garcia, & Welch, 2005) to learn how to classify new documents based on the following steps: train words, evaluate database statistics, eliminate uncommon words, determine stop words, train word pairs, evaluate database statistics, eliminate uncommon word pairs, and perhaps score the training set and trim misclassified training texts (BETSY, n.d.). After the training, BETSY™ can be applied to a set of trial texts to determine classification accuracy, several new texts, or a single text. Essay scoring typically categorizes texts into two or more groups such as *Pass/Fail* and *Advanced/Proficient/Basic/Below Basic*. Scoring is the major component of BETSY™ and several scoring and recalculation options allow users to identify what text to score and how to score it. The special options include using Microsoft Notepad® to analyze misclassifications, scoring an essay to get diagnostic feedback, and trimming misclassified training texts (Figure 5, next page) (BETSY, n.d.).

Figure 5: Scoring and recalculation window in BETSY™

It is claimed that BETSY™ includes the best features of PEG™, LSA, and e-rater® along with its own essential characteristics. In addition, the system can be applied to short essays and various content areas. Moreover, it is simple to implement and easy to explain to non-statisticians (BETSY, n.d.; Rudner & Liang, 2002; Valenti et al., 2003). The research on BETSY® is limited; however, the software can be downloaded from the official website for free for research purposes. BETSY® is in the process of being converted to VisualBasic and it will be soon become open sourced.

Summary and Discussion

There have been several studies over the past three decades that have examined ways to apply technology to writing assessment. More recently, increasingly sophisticated computer technology has enable writing performance to be assessed using AES technology (Burstein, 2003; Hamp-Lyons, 2001; Rudner & Gagne, 2001; Rudner & Liang, 2002). As Attali and Burstein (2006) maintain, AES systems do not directly evaluate the intrinsic qualities of an essay as human raters do, but they use correlations of the intrinsic qualities to predict the score of an essay. The automated essay systems described in this article employ various techniques to provide immediate feedback and scoring. While e-rater® and IntelliMetric™

use NLP techniques, IEA™ is based on LSA. Moreover, PEG™ utilizes proxy measures (*proxes*) and BETSY™ uses Bayesian methods to assess the quality of an essay. Unlike PEG™, IEA™, or BETSY™, e-rater and IntelliMetric have instructional applications (e.g., CriterionSM and MY Access![®]), as well. Finally, each AES system needs different numbers of essays to train the system. Table 1 below compares the AES systems.

Table 1: Comparison of AES Systems

AES System	Developer	Technique	Main Focus	Instructional Application	Number of Essays Required for Training
PEG™	Page (1966) ⁸	<i>Statistical</i>	Style	N/A	100–400
IEA™	Landauer, Foltz, & Laham (1997) ⁹	<i>LSA</i>	Content	N/A	100–300
E-rater®	ETS development team (Burstein, et al., 1998) ¹⁰	<i>NLP</i>	Style and content	Criterion SM	465
IntelliMetric™	Vantage Learning (Elliot, et al., 1998) ¹¹	<i>NLP</i>	Style and content	MY Access! [®]	300
BETSY™	Rudner ¹² (2002)	<i>Bayesian text classification</i>	Style and content	N/A	1000

One of the main advantages of AES system is that they can score essays instantly and provide immediate feedback. Teacher response is necessary for a student to improve his/her writing ability. However, for a teacher who teaches large classes, this can be quite time-consuming, which could possibly affect the frequency of the writing assignments given in class (Burstein et al., 2003). Since the appropriateness of feedback has been found to be highly individual specific and/or situation specific (Hyland, 1998), it will be essential to consider an effective method both for analyzing a large number of essays, but at the same time for providing individual feedback. Instructional-based AES systems (e.g., CriterionSM and MY Access![®]) make attempts to achieve this goal and aim to facilitate writing evaluation in classrooms. They are designed to supplement teachers but not to replace them (e.g., by allowing teachers to include their own scoring and feedback as well as their own prompts in the program (ETS, n.d.; Vantage Learning, n.d.).

Computers can also provide opportunities to increase practicality in the administration of large-scale writing assessment (Bereiter, 2003). Employing human raters to score essays could be quite expensive in terms

of time and resources. Large scale standardized writing tests require more than one rater in order to increase the accuracy in scoring and reduce the bias the individual scorers might have. Since hiring multiple raters and training them on a scoring rubric is necessary but costly, including an AES system in the assessment process could be a cost-effective option (Bereiter, 2003; Chung & O'Neil, 1997; Page, 2003; Sireci & Rizavi, 1999).

Most AES systems tend to focus on product rather than process in writing. While the product approach views writing assessment as a summative practice, the process approach views it as a formative practice. Drafting before submitting the final form of an essay is critical in process writing. Instructional-based AES systems (e.g., CriterionSM and MY Access![®]) make efforts to support formative assessment by allowing students to save their first and subsequent drafts on the computer and revise them based on the feedback and scoring they receive either from the computer or from the teacher. CriterionSM allows teachers to turn off the scoring feature so that students continue drafting online. While the previous version of MY Access![®] (5.0) included online portfolios only, the latest version (6.0) provides peer review opportunities and pre-writing activities to promote process writing. As Shermis and Burstein (2003) pointed out the credibility of AES systems will increase when their use moves from a summative to a more formative assessment.

AES systems are mainly developed based on English language. There are efforts to enable these programs to assess writing in various languages (Shermis & Burstein, 2003). CriterionSM, MY Access![®] and IntelliMetricTM currently include some features for English language learner (ELL) students. For instance, CriterionSM includes retired TOEFL (Test of English as a Foreign Language) prompts. CriterionSM, MY Access![®], and IntelliMetricTM contain multilingual feedback capacity (ETS, n.d.; Vantage Learning, n.d.).

As Warschauer and Ware (2006) pointed out while providing feedback in a student's native language is helpful, ELL students might need more than translation to improve their writing ability in English. The developers of AES systems need to question whether their current practices address the needs of ELL students. For instance, it will be important to investigate if the style, content, and amount of feedback provided by AES programs are appropriate for the ELL population.

One of the strongest objections to computerized scoring is that computers are not capable of assessing an essay as human raters do because computers do "what [they are] programmed to do" and do not "appreciate" an essay (Page, 2003, p. 51). Automated essay scoring systems have been criticized for eliminating the human element in writing assessment (Warschauer & Ware, 2006) and falling short of human interaction as well

as the sense of the writer and/or rater as a person (Hamp-Lyons, 2001). Landauer, Laham, and Foltz (2003) accept the fact that LSA may lack pertinent background information of the essay writers. However, Page (2003) argues against these claims by pointing out the high correlations between PEG™ and expert human raters. PEG™'s work obtaining high correlations by just looking at superficial surface features and recent reports on the high correlation of SAT essay scores with essay length indicate that modest accuracy may not be that hard to achieve.

Another criticism is the construct objections. Construct objections question the extent to which computers measure variables that are critical in scoring essays (Page, 2003). Both PEG™ and IEA™ have been criticized for their focus on essay constructs. The main focus of PEG™ is the surface features (e.g., word order and essay length) in writing rather than the meaning and content (Chung & O'Neil, 1997; Kukich, 2000). While IEA™ is superior to other AES systems in terms of assessing the content of an essay (Landauer et al., 2003; Rudner & Gagne, 2001), it fails to provide information regarding word order (Chung & O'Neil, 1997; Landauer et al., 2003). Vantage Learning, PKT, and ETS are currently working on adding new and improved features in an effort to increase what are already remarkably high accuracy rates. See Table 1 on page 23 for more information regarding the main focus of other AES systems.

An important issue with machine scoring is whether the computer can be fooled by writers or not (Page, 2003; Powers, Burstein, Chodorow, Fowles, & Kukich, 2001; Sireci & Rizavi, 1999). The developers of AES systems try to employ algorithms to defend against writers who try to cheat the computer. For instance, PEG uses an algorithm to alert the odd elements in an essay. When the computer flags an essay it is aside for human evaluation (Page, 2003). Similarly, CriterionSM and MY Access![®] programs flag anomalous essays for human scoring (ETS, n.d., Vantage Learning, n.d.). An earlier version of PEG™ was found to be vulnerable to cheating. Since the system mainly employed word count, word length, essay length, number of semicolons or commas, and so on (Chung & O'Neil, 1997; Kukich, 2000; Rudner & Gagne, 2001), it was possible to "trick" the computer by, for example, writing longer essays to receive higher scores (Kukich, 2000). A study was funded by the GRE (Graduate Record Examinations) Board to determine whether e-rater could be tricked into assigning a lower or higher value to an essay than it deserved (Powers et al., 2001). The results of the study revealed that e-rater might reward a poor essay. The findings suggested that e-rater was not ready to use by itself and it should be paired with human raters, particularly for high-stakes assessment purposes (Powers et al., 2001). One can argue whether the scores provided by one or two human raters are the right criteria. The AES systems could be more accurate considering that they are based on hundreds of reads of

the same essay and they are pooled across multiple raters. In other words, AES systems are likely to eliminate the between rater variance.

One of the main characteristics of AES systems is that they need to be trained on a large set of pre-scored essay samples in order to be able to evaluate the student essays effectively (Burstein, 2003; Chung & O'Neil, 1999; Elliott, 2003; Landauer et al., 2003; Rudner & Liang, 2002). The systems can only score the pre-scored prompts from their own libraries. Although teachers have the opportunity to assign their own prompts, the computer is not capable of scoring those prompts since it is not trained to assess the essays with unfamiliar prompts. Thus, the essays written on new prompts need to be scored either by teacher or an expert human rater. On the other hand, the AES systems are only as good as what they learn from the calibration sample. The calibration samples can be optimized by exposing the system to a large number of training essays on a particular prompt. For example, an AES system is able to score new NAEP essays if it is exposed to a large number of previously scored NAEP essays.

The AES systems described in this paper are claimed to be accurate and valid. For example, in their 2002 study, Rudner & Liang reported that the Bayesian approach presented accurate results in text categorization as high as .80 (Rudner & Liang, 2002). The correlations and agreement rates between e-rater®, IEA™, IntelliMetric™, IEA™, or PEG™ and expert human raters have been found to be high, as well (Attali, 2004; Burstein & Chodorow, 1999; Landauer et al., 1997; Nichols, 2004; Page, 2003; Landauer et al., 2003; Vantage Learning, 2000a, 2000b, 2001b, 2002, 2003a, 2003b). While reviewing information regarding agreement, it is critical to understand the difference between exact agreement and adjacent agreement. Exact agreement requires two or more raters to assign same exact score on an essay (e.g., two raters assign 5 on a 1–6 scoring scale). On the other hand, adjacent agreement requires two or more raters to assign a score within one scale point of each other (e.g. one rater assigns 5 and another rater assigns 6 respectively on a 1–6 point scoring scale) (Cizek & Page, 2003; Elliott, 2003). It is clear that exact agreement is harder to achieve and that adjacent agreement results in higher agreement rates (Cizek & Page, 2003). Table 2 (next page) compares the agreement rates across three different constructed-response scoring modes (expert scoring, standard human scoring, and IntelliMetric™ scoring) that were used to assess the writing responses of eighth-grade students from a statewide testing program (Vantage Learning, 2002, p. 4). It shows how the adjacent agreement rates between humans and IntelliMetric™ and humans can be higher than the exact agreement rates. The study was conducted by Vantage Learning using the IntelliMetric™ program. First, two expert raters scored each essay response. Then, two traditional human scorers independently scored those responses, and finally, IntelliMetric™ scored each response.

In this study, while “expert” referred to an individual who had a degree in English as well as at least five years of experience in analyzing writings in large-scale writing assessment programs, “traditional human scorer” was defined as an individual who usually attended a one-day training session on writing assessment in a large, statewide scoring session in writing (Vantage Learning, 2002).

Table 2: Comparison of Expert Scoring, Human Scoring, and IntelliMetric Scoring

	Human 1 to Human 2	Human 1 to IntelliMetric	Human 2 to IntelliMetric	Human 1 to Experts	Human 2 to Experts	IntelliMetric to Experts
Exact	.52	.53	.56	.58	.54	.73
Adjacent	.94	.96	.95	.96	.97	.99
Discrepant	.6	.4	.5	.4	.3	.1

Increasing the reliability of AES systems has always been of great interest to AES researchers. The most common way to enhance the reliability of an AES system is to calibrate the system with a large number of sample essays to make sure that it is well-trained. Another way could be using the accuracy as a function of alternative calibration pools. Employing different training sets will ensure the inclusion of more than one calibration pool, which might help better assess the reliability of AES systems.

MY Access!® and CriterionSM are student based tools that have emerged from a computer technology that was originally created to help testing organizations score large numbers of essays. Currently, these systems are being used in writing classes at various schools and universities as writing tools. While AES systems assist teachers in writing classes, they are not free of charge. In the future, it would be interesting to see these systems as a public utility rather than a proprietary vendor-created-and-owned system. For instance, federal government could use NAEP essays and collect writing samples so that new essay prompts would be available for teachers to use in their writing classes. This would allow more teachers and students to benefit from the AES systems in writing classrooms.

The demand for incorporating AES systems in writing assessment is increasing. Although some teachers and educators may fear that AES technology will eventually substitute humans, the producers of classroom-based AES systems (e.g., MY Access!® and CriterionSM) claim that the main role of these systems is not to replace teachers in writing classes but to assist them. An effective way of using AES technology to score essays is to

incorporate the AES system into the writing evaluation process as a second or third rater. As Monaghan and Bridgeman (2005) suggested, using an AES system as a check point to compare the scores assigned by human readers can be an effective way of incorporating the AES technology in writing assessment. In other words, the AES systems can be used both to verify human scoring and to represent a collection of human judges in large-scale writing assessments.

Today AES systems are widely being used as instructional tools in classrooms (e.g., MY Access!® and CriterionSM) and as a co-rater in scoring large-scale standardized writing assessments (e.g., ETS has used e-rater along with a human rater to score GMAT essays since 1999) without excluding the human element. Although AES is a developing technology (Shermis & Burstein, 2003) the search for better machine scoring is ongoing as investigators continue to move forward in their drive to increase the accuracy and effectiveness of AES systems.

Endnotes

1. See <http://www-formal.stanford.edu/jmc/whatisai/whatisai.html> for more information.
2. See <http://www.ed.gov/programs/naep/index.html> for more information.
3. See www.ets.org/redirect/tests.html for more information.
4. See www.ets.org/praxis for more information.
5. See www.ets.org/redirect/tests.html for more information.
6. The Editors of the JTLE have altered Vantage Learning's original model in an effort to present the information more clearly (see Figure 4). Please refer to Vantage Learning (2003a, p. 73) to view the model's original configuration.
7. Lawrence Rudner is currently the chief statistician with the Graduate Management Admission Council (GMAC).
8. See <http://www.pearsonkt.com/papers/IEEEdebate2000.pdf> for more information.
9. See <http://pareonline.net/getvn.asp?v=7&n=26> for more information.
10. See http://www.edres.org/betsy/three_prominent.htm for more information.
11. See <http://www.vantage.com/pdfs/intellimetric.pdf> for more information.
12. See <http://www.edres.org/betsy/history.htm> for more information.

References

- Attali, Y. (2004, April). *Exploring the feedback and revision features of Criterion*. Paper presented at the National Council on Measurement in Education (NCME), San Diego, CA.
- Attali, Y. & Burstein, J. (2006). Automated Essay Scoring with e-rater V.2. *Journal of Technology, Learning, and Assessment (JTLA)*, 4(3).
- Bereiter, C. (2003). Foreword. In Mark D. Shermis and Jill C. Burstein (Eds.), *Automated essay scoring: a cross disciplinary approach* (pp. vii–ix). Mahwah, NJ: Lawrence Erlbaum Associates.
- BETSY. (n.d.). Retrieved September 06, 2005, from <http://edres.org/betsy>
- Brill, E. & Mooney, R. (1997). An overview of empirical natural language processing. *AI Magazine*, 18(4), 13–24.
- Burstein, J. (2003). The e-rater scoring engine: Automated Essay Scoring with natural language processing. In M. D. Shermis and J. C. Burstein (Eds.), *Automated Essay Scoring: A cross disciplinary approach* (pp. 113–121). Mahwah, NJ: Lawrence Erlbaum Associates.
- Burstein, J. & Chodorow, M. (1999, June). *Automated Essay Scoring for nonnative English speakers*. Proceedings of the ACL99 Workshop on Computer-Mediated Language Assessment and Evaluation of Natural Language Processing, College Park, MD.
- Burstein, J., Chodorow, M., & Leacock, C. (2003, August). *Criterion: Online essay evaluation: an application for automated evaluation of student essays*. Proceedings of the 15th Annual Conference on Innovative Applications of Artificial Intelligence, Acapulco, Mexico.
- Burstein, J., Kukich, K., Wolff, S., Lu, C., & Chodorow, M. (1998, April). *Computer analysis of essays*. Proceedings of the NCME Symposium on Automated Scoring, Montreal, Canada.
- Burstein, J., Leacock, C., & Swartz, R. (2001). *Automated evaluation of essays and short answers*. Proceedings of the 5th International Computer Assisted Assessment Conference (CAA 01), Loughborough University.
- Burstein, J. & Marcu, D. (2000). *Benefits of modularity in an Automated Essay Scoring System* (ERIC reproduction service no TM 032 010).
- Chung, K. W. K. & O'Neil, H. F. (1997). *Methodological approaches to online scoring of essays* (ERIC reproduction service no ED 418 101).

- Cizek, G. J. & Page, B. A. (2003). The concept of reliability in the context of automated essay scoring. In M. D. Shermis & J. Burstein (Eds.), *Automated essay scoring: A cross-disciplinary perspective* (pp. 125–145). Mahwah, NJ: Lawrence Erlbaum Associates.
- Elliot, S. (2003). IntelliMetric: from here to validity. In Mark D. Shermis and Jill C. Burstein (Eds.), *Automated essay scoring: a cross disciplinary approach*. Mahwah, NJ: Lawrence Erlbaum Associates.
- Educational Testing Service (ETS). (n.d.). Retrieved on May 06, 2004, from <http://www.ets.org>
- Foltz, P. W. (1996). Latent Semantic Analysis for text-based research. *Behavior Research Methods, Instruments and Computers*, 28(2), 197–202. Retrieved June 6, 2004 from <http://www-psych.nmsu.edu/~pfoltz/>
- Foltz, P. W., Laham, D. & Landauer, T. K. (1999). *Automated Essay Scoring: Applications to educational technology*. Proceedings of EdMedia '99. Retrieved May 15, 2004 from <http://www-psych.nmsu.edu/~pfoltz/reprints/Edmedia99.html>
- Hamp-Lyons, L. (2001). Fourth generation writing assessment. In T. Silva and P. K. Matsuda (Eds.), *On Second Language Writing* (pp. 117–125). Mahwah, NJ: Lawrence Erlbaum Associates, Inc.
- Hyland, F. (1998). The impact of teacher written feedback on individual writers. *Journal of Second Language Writing*, 7(3), 255–286.
- Kukich, K. (2000, September/October). Beyond Automated Essay Scoring. In M. A. Hearst (Ed.), *The debate on automated essay grading*. IEEE Intelligent systems, 27–31. Retrieved November 12, 2004, from <http://que.info-science.uiowa.edu/~light/research/mypapers/autoGradingIEEE.pdf>
- Landauer, T. K., Foltz, P. W., & Laham, D. (2004). *What is LSA?* Retrieved Nov 15, 2004, from <http://lsa.colorado.edu/whatis.html>
- Landauer, T. K., Laham, D., & Foltz, P. W. (2000, September/ October). The Intelligent Essay Assessor. In M. A. Hearst (Ed.), *The debate on automated essay grading*. IEEE Intelligent systems, 27–31. Retrieved November 12, 2004, from <http://que.info-science.uiowa.edu/~light/research/mypapers/autoGradingIEEE.pdf>
- Landauer, T. K., Laham, D., & Foltz, P. W. (2003). Automated Essay Scoring: A cross disciplinary perspective. In M. D. Shermis and J. C. Burstein (Eds.), *Automated Essay Scoring and annotation of essays with the Intelligent Essay Assessor* (pp. 87–112). Mahwah, NJ: Lawrence Erlbaum Associates.

- Landauer, T. K., Laham, D., Rehder, B. & Schreiner, M. E. (1997). *How well can passage meaning be derived without using word order? A comparison of Latent Semantic Analysis and humans*. Proceedings of the 19th Annual Conference of the Cognitive Science Society, (pp. 412–417). Mahwah, NJ: Erlbaum.
- Lemaire, B. & Dessus, P. (2001). A system to assess the semantic content of student essays. *Educational Computing Research*, 24(3), 305–306.
- Monaghan, W. & Bridgeman, B. (2005, April) E-rater as a quality control on human scorers. *ETS RD Connections*. Retrieved April 03, 2006, from http://www.ets.org/Media/Research/pdf/RD_Connections2.pdf
- Murray, B. (1998). The latest techno tool: Essay grading computers. *American Psychological Association (APA)*, 8(29). Retrieved April 16, 2005, from <http://www.apa.org/monitor/aug98/grade.html>
- Myers, M. (2003). What can computers and AES contribute to a K–12 writing program? In M. D. Shermis & J. Burstein (Eds.), *Automated essay scoring: A cross-disciplinary perspective* (pp. 3–20). Mahwah, NJ: Lawrence Erlbaum Associates.
- Nichols, P. D. (2004, April). *Evidence for the interpretation and use of scores from an Automated Essay Scorer*. Paper presented at the Annual Meeting of the American Educational Research Association (AERA), San Diego, CA.
- Page, E. B. (2003). Project Essay Grade: PEG. In M. D. Shermis & J. Burstein (Eds.), *Automated essay scoring: A cross-disciplinary perspective* (pp. 43–54). Mahwah, NJ: Lawrence Erlbaum Associates.
- Page, E. B. (1994). Computer Grading of Student Prose, Using Modern Concepts and Software, *Journal of Experimental Education*, 62, 127–142.
- Pearson Knowledge Technologies (PKT). (n.d.). Retrieved November 8, 2004, from http://www.knowledge-technologies.com/papers/IEA_FAQ.html
- Powers, D. E., Burstein, J. C., Chodorow, M., Fowles, M. E., & Kukich, K. (2001). *Stumping e-rater: Challenging the validity of automated scoring*, (GRE No. 98–08). Princeton, NJ: Educational Testing Service.
- Pro's and con's of Essay Assessing Software*. (n.d.). Retrieved June 17, 2005, from <http://people.emich.edu/lhubbard/lori444/prosandcons.htm>
- Pspotka, J. & Streeter, L.(n.d.) *Automatically critiquing writing for army educational settings*. Retrieved December 02, 2004, from http://www.hqda.army.mil/ari/pdf/critiquing_writing.pdf

- Rudner, L. & Gagne, P. (2001). *An overview of three approaches to scoring written essays by computer* (ERIC Digest number ED 458 290).
- Rudner, L., Garcia, V., & Welch, C. (2005). An Evaluation of Intellimetric™ Essay Scoring System Using Responses to GMAT® AWA Prompts (GMAC Research report number RR-05-08). Retrieved April 8, 2006, from <http://www.gmac.com/gmac/researchandtrends/>
- Rudner, L. M. & Liang, T. (2002). Automated essay scoring using Bayes' theorem. *The Journal of Technology, Learning, and Assessment*, 1(2). Retrieved September 18, 2005, from <http://www.jtla.org>
- Salem, A. B. M. (2000). *The potential role of artificial intelligence technology in Education* (ERIC document reproduction service no ED 477 318).
- Shermis, M. & Barrera, F. (2002). *Exit assessments: Evaluating writing ability through Automated Essay Scoring* (ERIC document reproduction service no ED 464 950).
- Shermis, M. D. & Burstein, J. (2003). *Automated Essay Scoring: A cross disciplinary perspective*. Mahwah, NJ: Lawrence Erlbaum Associates.
- Shermis, M. D., Raymat, M. V., & Barrera, F. (2003). *Assessing writing through the curriculum with Automated Essay Scoring* (ERIC document reproduction service no ED 477 929).
- Sireci, S. G. & Rizavi, S. (1999). *Comparing computerized and human scoring of students' essays* (ERIC document reproduction service no ED 463 324).
- Streeter, L., Psotka, J., Laham, D., & MacCuish, D. (2004). The credible grading machine: Essay scoring in the DOD [Department of Defense]. Retrieved on January 10, 2005, from <http://www.k-a-t.com/papers/essayscoring.pdf>
- Valenti, S., Neri, F., & Cucchiarelli, A. (2003). An overview of current research on automated essay grading. *Journal of Information Technology Education*, (2), 319–330.
- Vantage Learning. (n.d.). Retrieved May 12, 2004 from, <http://www.vantagelearning.com>
- Vantage Learning. (1999). *Construct validity of IntelliMetric with International Assessment* (RB-323). Newtown, PA: Vantage Learning.
- Vantage Learning. (2000a). *A study of expert scoring and IntelliMetric scoring accuracy for dimensional scoring of Grade 11 student writing responses* (RB-397). Newtown, PA: Vantage Learning.

- Vantage Learning. (2000b). *A true score study of IntelliMetric accuracy for holistic and dimensional scoring of college entry-level writing program* (RB-407). Newtown, PA: Vantage Learning.
- Vantage Learning. (2001a). *About IntelliMetric* (PB-540). Newtown, PA: Vantage Learning.
- Vantage Learning. (2001b). *Applying IntelliMetric Technology to the scoring of 3rd and 8th grade standardized writing assessments* (RB-524). Newtown, PA: Vantage Learning.
- Vantage Learning. (2002). *A study of expert scoring, standard human scoring and IntelliMetric scoring accuracy for statewide eighth grade writing responses* (RB-726). Newtown, PA: Vantage Learning.
- Vantage Learning. (2003a). *Assessing the accuracy of IntelliMetric for scoring a district-wide writing assessment* (RB-806). Newtown, PA: Vantage Learning.
- Vantage Learning. (2003b). *How does IntelliMetric score essay responses?* (RB-929). Newtown, PA: Vantage Learning.
- Vantage Learning. (2003c). *A true score study of 11th grade student writing responses using IntelliMetric Version 9.0* (RB-786). Newtown, PA: Vantage Learning.
- Warschauer, M. & Ware, P. (2006). Automated writing evaluation: Defining the classroom research agenda. *Language Teaching Research*, 10, 2, 1–24. Retrieved June 05, 2006, from <http://www.gse.uci.edu/faculty/markw/awe.pdf>

Acknowledgements

I would like to thank Dr. Deborah Hasson for her valuable comments on the description of AES systems in the earlier drafts of this article. I am grateful to my fiancé Daniel Malaxa for helping me better understand the computer jargon the developers of AES systems used. I am also grateful to TESOL International Research Foundation (TIRF) to provide me with funding to do research in AES. Finally, I acknowledge the helpful suggestions and corrections provided by the JTLE reviewers and editors.

Author Biography

Semire Dikli is an advanced PhD candidate in the department of Middle and Secondary Education at the Florida State University. She taught English as a foreign language (EFL) both at the middle school and university levels in her home country, Turkey, for four years. She has also taught an undergraduate level course: Second Language Testing and Evaluation at the Florida State University for three years. Her research interest is Second language writing and technology. Currently, she is writing the last two chapters of her dissertation [An exploratory study of Automated Essay Scoring (AES) in an English as a Second Language (ESL) setting]. She received a doctoral dissertation grant from TESOL International Research Foundation (TIRF) in 2005 to complete her study. She can be contacted at ssd0960@garnet.acns.fsu.edu.



The Journal of Technology, Learning, and Assessment

Editorial Board

Michael Russell, Editor
Boston College

Allan Collins
Northwestern University

Cathleen Norris
University of North Texas

Edys S. Quellmalz
SRI International

Elliot Soloway
University of Michigan

George Madaus
Boston College

Gerald A. Tindal
University of Oregon

James Pellegrino
University of Illinois at Chicago

Katerine Bielaczyc
Museum of Science, Boston

Larry Cuban
Stanford University

Lawrence M. Rudner
Graduate Management
Admission Council

Marshall S. Smith
Stanford University

Paul Holland
Educational Testing Service

Randy Elliot Bennett
Educational Testing Service

Robert Dolan
Center for Applied
Special Technology

Robert J. Mislevy
University of Maryland

Ronald H. Stevens
UCLA

Seymour A. Papert
MIT

Terry P. Vendlinski
UCLA

Walt Haney
Boston College

Walter F. Heinecke
University of Virginia

www.jtla.org