# Automated Essay Scoring

Semire DIKLI
Florida State University
Tallahassee, FL, USA

## ABSTRACT

The impacts of computers on writing have been widely studied for three decades. Even basic computers functions, i.e. word processing, have been of great assistance to writers in modifying their essays. The research on Automated Essay Scoring (AES) has revealed that computers have the capacity to function as a more effective cognitive tool (Attali, 2004). AES is defined as the computer technology that evaluates and scores the written prose (Shermis & Barrera, 2002; Shermis & Burstein, 2003; Shermis, Raymat, & Barrera, 2003).

Revision and feedback are essential aspects of the writing process. Students need to receive feedback in order to increase their writing quality. However, responding to student papers can be a burden for teachers. Particularly if they have large number of students and if they assign frequent writing assignments, providing individual feedback to student essays might be quite time consuming. AES systems can be very useful because they can provide the student with a score as well as feedback within seconds (Page, 2003). Four types of AES systems, which are widely used by testing companies, universities, and public schools: Project Essay Grader (PEG), Intelligent Essay Assessor (IEA), E-rater, and IntelliMetric.

AES is a developing technology. Many AES systems are used to overcome time, cost, and generalizability issues in writing assessment. The accuracy and reliability of these systems have been proven to be high. The search for excellence in machine scoring of essays is continuing and numerous studies are being conducted to improve the effectiveness of the AES systems.

Keywords: Assessment, Writing, Feedback Mechanism, Assistive Technologies

## INTRODUCTION

The impacts of computers on writing have been widely studied for three decades. Even basic computers functions, i.e. word processing, have been of great assistance to writers in modifying their essays. The research on Automated Essay Scoring (AES) has revealed that computers have the capacity to function as a more effective cognitive tool (Attali, 2004). AES is defined as the computer technology that evaluates and scores the written prose (Shermis & Barrera, 2002; Shermis & Burstein, 2003; Shermis, Raymat, & Barrera, 2003).

Revision and feedback are essential aspects of the writing process. Students need to receive feedback from the teacher in order to increase their writing quality. However, responding to student papers can be a burden for teachers. Particularly if they have large number of students and if they assign frequent writing assignments, providing individual feedback to student essays might be quite time consuming. AES systems can be very useful because they can provide the student with a score as well as feedback within seconds. Also, the scores would be much more descriptive than the ratings provided by two human raters (Page, 2003).

Machine scoring technologies can also increase the practicality in administering large-scale assessments of writing ability (Bereiter, 2003). Employing human raters could be quite expensive in terms of time and resources. It is necessary to include more than one

**rater in large-scale writing assessments to reduce the bias the individual scorers might have. The training of multiple raters on a holistic scoring rubric is necessary but costly as well. In this case, it might be cost-effective to use an AES system (Bereiter, 2003; Chung & O'Neil, 1997; Page, 2003). Besides being a time-and money-saver, automated essay scoring systems are claimed to provide variety in feedback, not only on grammatical issues, but also on discourse related issues (Shermis & Burstein, 2003, p. xiv). Myers (2003) claims that this reduces not only the teacher's paper load, but also the issues of concern (e.g., subjectivity) with teacher assessment. Similarly, Hamp-Lyons (2001) highlights the advantages of AES technology as follows, the ability to perform repeated functions without boredom and variation, adaptability (within preprogrammed pathways), flexibility (testing can be carried out at any time, for a range of purposes, and on any number of candidates), and the ability to make decisions without being judgmental (in the sense of being biased) or confrontational (p. 121).**

**Moreover, Page (2003) states that "the automated ratings would surpass the accuracy of the usual two judges. (Accuracy is defined as agreeing with the mean of judgments)" (p. 46). Finally, providing "a third voice" (p.15) about student writing, these types of programs can be effective tools in student-teacher conferences (Myers, 2003). A number of studies are conducted to prove the accuracy and reliability of the AES systems with respect to the writing assessment and the agreement rate between human raters and AES systems are found to be high (Attali, 2004; Burstein & Chodorow, 1999; Elliot, 2000a, 2000b, 2001c, 2002, 2003b, 2003c; Landauer, Laham, & Foltz, 2003; Landauer, Laham, Rehder, & Schreiner, 1997; Nichols, 2004; Page, 2003, 2004).**

**Computerized scoring has many weaknesses as well. Hamp-Lyons (2001) stressed the lack of human interaction as well as the sense of the writer and/or rater as person. Similarly, Page (2003) stated that the computers could not assess an essay as human raters do because the computer would do "what it is programmed to do" and it wouldn't "appreciate" an essay (p. 51). Another criticism is the construct objections. That is, the computer counts variables that might not be "truly" important in essay grading, i.e., focusing on formal aspects rather than organizational ones (Page, 2003; Chung & O'Neil, 1997).**

## AUTOMATED ESSAY SCORING (AES) SYSTEMS

**Four types of AES systems are widely used by testing companies, universities, and public schools. The first one is Essay Grade (PEG), which is known as the first AES system built in AES history (Kukich, 2000; Rudner & Gagne, 2001; Page, 2003). The second one, Intelligent Essay Assessor (IEA), is developed by Landauer, Laham, and Foltz using Latent Semantic Analysis (LSA) features (http://lsa.colorado.edu/whatis.html). Another AES system, E-rater, has been used by the ETS (Educational Testing Service) to score essay portion of GMAT (Graduate Management Admissions Test). The final AES system is called IntelliMetric. It is developed by Vantage learning and used by the College Board for placement purposes (Myers, 2003).**

## 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 proxy measures to predict the intrinsic quality of the essays. Proxies refer to the particular writing construct such as average word length, essay length, number of semicolons or commas, and so on (Kukich, 2000; Chung & O'Neil, 1997; Rudner & Gagne, 2001).**

**One of the strengths of PEG is that the predicted scores are comparable to those of human raters. Second, the system is computationally tractable. In other words, it is able to track the writing errors made by the users. Next, its scoring methodology is straightforward. PEG contains a training stage and a scoring stage. The system is trained**

**on a sample of essays in the former stage. In the latter stage, proxy variables are determined for each essay and these variables are entered into the prediction equation. Finally, a score is assigned by computing beta weights from the training stage (Chung & O'Neil, 1997). PEG has been criticized for ignoring the semantic aspect of essays and focusing more on the surface structures (Kukich, 2000; Chung & O'Neil, 1997). Failing to detect the content related features of an essay (organization, style etc.), the system does not provide instructional feedback to the students. Also, an early version of the system was found to be weak in terms of scoring accuracy. The main concern was the vulnerability of the system to cheating. Since PEG used indirect measures of writing skill, it was possible to trick the system, i.e., writing longer essays (Kukich, 2000). PEG was modified on several aspects in 1990s. It incorporated 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). LSA approach was created by psychologist Thomas Landauer, a psychology professor at the University of Colorado at Boulder, with the assistance of Peter Foltz, a professor at the New Mexico State University and Darrell Laham, a PhD student at UC (Murray, 1998). IEA is produced by the Pearson Knowledge Analysis Technologies (PKT) (Psotka & Streeter, (n.d.); http://www.knowledge-technologies.com). A richer description of LSA and IEA is provided below.**

### Latent Semantic Analysis (LSA)

**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, (http://lsa.colorado.edu/whatis.html) paragraph, or essay.**

**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, Laham, & Foltz, 2003). The underlying idea can be summarized as "meaning of word1+ meaning of word 2 + ...........+meaning of word n = meaning of passage" (Landauer, Laham, & Foltz, 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 (http://lsa.colorado.edu/whatis.html).**

**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). The essay, then, 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. When compared to content related factors (e.g., argument, comprehensibility, style), mechanical and syntactic features are easier to separate from other factors. The reason is that content related factors are very much affected by the word choice. Previous research on automated essay scoring has concentrated on the analysis of style. Unlike other methods, the emphasis of LSA is on the conceptual content of an essay (http://lsa.colorado.edu/whatis.html; Foltz, Laham, & Landauer, 1999).**

**In the LSA based approach, the text is represented as a matrix. Each row in the matrix stands for a unique word, while each column stands for 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 (http://lsa.colorado.edu/ whatis.html). The semantics of a word is verified through all the contexts that the word occurs. The number of occurrences of each word in a text determines the semantic space. For example, 300 paragraphs and 2000 words provide a 300X 2000 matrix. Here, while each word is represented by a 300-dimentional vector, each paragraph is represented by a 2000-dimentional vector. By reducing these dimensions, LSA induces semantic similarities between words. This reduction is critical since it permits the representation of the 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)

**It is claimed that 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, i.e., grammar and punctuation, in an essay. In other words, even though the system is uses an LSA based approach to evaluate mainly the quality of the content of an essay, it includes scoring and feedback on grammar, style and mechanics as well (Landauer, Laham, & Foltz, 2000; Landauer, Laham, & Foltz, 2003; Streeter, Psotka, Laham, & MacCuish, 2004).**

**It is also claimed that IEA can successfully analyze not only the content-based essays, but also the creative narratives. The system needs to be trained on a set of domain-representative texts in order to judge the overall quality of an essay. For example, a biology book can be used to evaluate a biology essay. IEA uses three methods to analyze an essay:**

- **pre-scored essays of other students,**
- **expert model essays and knowledge source materials,**
- **internal comparison of an unscored set of essays" (Landauer et al., 2003, p. 90).**

**These methods allow IEA to compare the student essay with similar texts in terms of the content quality (Landauer et al., 2000; Landauer et al., 2003; Streeter et al, 2004). IEA, first, compares the content similarity between the student essay and other essays on the same topic that are scored by human raters and determines the closeness between them (Landauer et al., 2000; Rudner & Gagne, 2001; Streeter et al, 2004). It, then, predicts the overall score by adding "corpus-statistical writing-style" and mechanics (Hearst, 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 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).**

**Landauer, Laham, and Foltz (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 they benefited from the system greatly since it needs smaller numbers of pre-scored essays to train. Unlike other AES systems, which require 300-500 training essays per prompt, IEA only requires around 100 pre-scored essays (http://www.knowledge-technologies.com; Landauer et al., 2003).**

**Another reason is that IEA does not require a representative sample of all scores in the rubric, either. They claim that the system is so intelligent that it can determine the scale of the essay. For example, the system is able to predict what an essay with 6 point looks like in a 6 point holistic scale without seeing large numbers of essays with 6 point (http://www.knowledge-technologies.com).**

**Finally, the developers of IEA claim that the system does not evaluate the creativity and reflective thinking. It, however, assesses "expository essays on factual topics", i.e., description of a psychological theory, the 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 the audience (Landauer et al., 2003).**

### E-RATER and CRITERION

**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 & Marcu, 2000; Burstein, 2003). E-rater uses natural language processing (NLP) techniques, which identify specific lexical and syntactic cues in a text, to analyze essays (Kukich, 2000; Burstein, 2003). A detailed description of natural language processing and information regarding the structure and functions of e-rater and Criterion is provided below.**

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

**The main focus of artificial intelligence (AI) is creating intelligent machines. The applications of AI can be divided into two groups. While the first group involves new applications that cannot be done without intelligent use of computers, the second group includes applications that can replace human workers or make the humans' job easier. The examples for the first group are weather forecasting, real world simulators and computer games, robot applications to keep humans away from danger (i.e. space missions, work in nuclear polluted areas). The examples for the second group include automatic information processing like speech recognition, helpdesks, computer vision, and natural language processing (http://www.geocities.com). NLP is considered one of the most challenging areas of AI. The research in NLP contains a variety of fields including corpus-based methods, discourse methods, formal models, machine translation, natural language generation and spoken language understanding (Salem, 2000).**

**NLP is claimed to be a complex task to comprehend since it contains several levels of processing as well as subtasks. It has four categories of language tasks including speech recognition, syntactic analysis, discourse analysis and 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 the words are clustered into constituents like noun and verb phrases. Semantic analysis employs diagramming a sentence to a type of meaning representation such as a logical expression. While, discourse analysis focuses on how context impacts sentence interpretation, 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, i.e., English to German or vice versa (Brill & Mooney, 1997).**

### E-rater

**E-rater is currently used by ETS for operational scoring of the Graduate Management Admissions Test (GMAT) AWA (Analytical Writing Assessment) (Burstein, 2003; Burstein & Chodorow, 1999; Burstein & Marcu, 2000). Prior to e-rater, GMAT AWA was scored by two human raters on a 6-point holistic scale, 6 being the highest and 1 being the lowest score. If there was discrepancy between two raters by more than 1 point, a third rater was called for resolution (Burstein, 2003; Burstein & Chodorow, 1999; http://www.gmat.org). E-rater has been employed in scoring the AWA since February 1999. 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 discrepancy between e-rater and the human rater by more than 1 point, a second human rater is included (Burstein, 2003). Burstein (2003) claims that since e-rater was used to score the GMAT AWA, the discrepancy rate between e-rater and human raters has been less than 3 percent.**

**E-rater employs a corpus-based approach to model building, in which actual essay data is used to examine the sample essays. A corpus-based approach of building NLP-based tools requires researchers to usually use copyedited 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, & Swarz, 2001).**

**The features of e-rater include a syntactic module, a discourse module, and a topical analysis module. 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 identified in Quirk et al. in 1985 (as cited in Burstein, Chodorow, & Leacock, 2003). This framework includes 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 & Chodrow,1999; Burstein, Chodorow, & Leacock, 2003; Burstein & Marcu, 2000; Burstein, Kukich, Woff, Lu, & Chodorow, 1998). Finally, the topical analysis module identifies vocabulary usage and topical content (Burstein, 2003; Burstein, Chodorow, & Leacock, 2003; Burstein & Marcu, 2000). The syntactic, discourse, and topical analysis modules discussed above provided 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, Chodorow, & Leacock, 2003; Burstein & Marcu, 2000).**

**Unlike a poor essay, a good essay needs to be relevant to the topic assigned. Moreover, the variety and the type of vocabulary used in good essays are different from the ones in poor essays. The assumptions behind this module are that good essays resemble other good essays. A similar assumption is also valid for poor essays as well (Burstein & Chodorow, 1999; Burstein, Kukich, Woff, Lu, & Chodorow, 1998). A vector-spec model (Salton as cited in Burstein & Marcu, 2000) used to capture the topic or vocabulary usage (Burstein & Chodorow, 1999; Burstein, Chodorow, & Leacock, 2003; Burstein, Kukich, Woff, Lu, & Chodorow, 1998; Burstein & Marcu, 2000). The general procedure is described as follows (Burstein, 2003):**

> **...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 (p. 117).**

**In other words, 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 (http://www.ets.org/criterion/ell/faq.html).**

## Criterion

**Criterion 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; recognize discourse elements and elements of undesirable style in an essay. Besides providing instant scoring, Criterion also gives individualized diagnostic feedback based on the types of evaluations that teachers give when responding to student writing (Burstein, Chodorow, & Leacock, 2003). This web-based, real-time system allows teachers and students to see the e-rater score and relevant feedback immediately. The feedback component of Criterion is called "advisory component." The advisory component functions as a supplement to the e-rater score and it is not used to determine the score (Burstein, 2003). The feedback types that the advisory component contains are as follows:**

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

**Criterion covers a number of 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 (GRE), and non-native speakers of English (TOEFL). The topics are taken from authentic retired ETS essay topics. They are obtained from various ETS programs such as NAEP (National Assessment of Educational Progress), English Placement Test designed for California State University, Praxis, and TOEFL. Criterion is not able to assess essays on other topics. It is only capable of analyzing essays on the topics for which it has been "trained." Furthermore, 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 Criterion library, yet they can use their choice of topics. While holistic scoring can not be reported for teacher-created topics, it is possible to obtain feedback of every dimension of writing. Finally, Criterion can be used for assessment and placement purposes as well (http://www.ets.org/criterion/ ell/html).**

## 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) (Elliot, 2003d; Shermis & Barrera, 2002; Shermis, Raymat, & Barrera, 2003). Like e-rater, IntelliMetric relies on NLP, which determines "the meaning of a text by parsing the text in known ways according to known rules conforming to the rules of English language" (Elliott, 2003a, p. 7). MY Access is known as the instructional application of IntelliMetric (http://www.vantagelearning.com). More information about the structure and the functions of the 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, 2003d, p. 71). As an advanced artificial intelligence application for scoring essays, IntelliMetric relies on Vantage Learning's CogniSearch and Quantum Reasoning technologies (Elliot, 2003d; Shermis & Barrera, 2002; Shermis, Raymat, & Barrera, 2003; Vantage learning, 2001a, 2003a). 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 (Elliott, 2003a). CogniSearch and Quantum Reasoning technologies together allow IntelliMetric to internalize each score point that is associated with certain characteristics in an essay response and then apply to subsequent scoring by the system (Elliot, 2001a, 2003d; Shermis & Barrera, 2002; Shermis, Raymat, & Barrera, 2003). This approach is claimed to be consistent with the procedure underlying holistic scoring (Elliot, 2003d). 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, Raymat, & Barrera, 2003).**

**IntelliMetric needs to be "trained" with a set of essays that have been scored beforehand including "known scores" determined by human expert raters (Elliott, 2001a; 2003a). The system employs a multi-stage method in analyzing essay responses (Shermis & Barrera, 2002).**

**In the first step, IntelliMetric, internalizes the known score points of a set of responses. Subsequently, the model is tested against a smaller set of response with known scores that aides in validation and generalizability of the model. Once these are confirmed, the model is used to score new responses whose scores are unknown. Responses are targeted if they are evaluated to be atypical with regards to the standards previously set by the essay scoring or by standard American English (p. 15).**

**IntelliMetric evaluates over 300 semantic, syntactic and discourse related features in an essay by using AI and NLP technologies (see AI and NLP section above for more information) (Elliot, 2001a, 2003d). These text-related features are identified as larger categories called Latent Semantic Dimensions (LSD) (Elliott, 2003a). The LSD features are described in five broad categories. The first category, focus and unity, uses the features that emphasizes a single point of view, cohesiveness and consistency in purpose and main ideas in an essay. The development and elaboration category examines the breadth of the content and the supporting ideas, i.e. vocabulary, elaboration, word choice, concepts, and support, in an essay. The third category, organization and structure, analyzes transitional fluency and logic of discourse. The examples contain introduction and conclusion, coordination and subordination, logical structure, logical transitions, and sequence of ideas. The category of sentence structure focuses on sentence complexity and variety such as syntactic variety, sentence complexity, usage, readability, and subject-verb agreement. Finally, the category of mechanics and conventions analyze whether the essay includes the conventions of standard American English, i.e. grammar, spelling, capitalization, sentence completeness, and punctuation (Elliot, 2001a, 2003a, & 2003d).**

**There are five key principles underlying the IntelliMetric system. First of all, IntelliMetric is modeled on the human brain. IntelliMetric "emulates the way in which the human brain acquires, stores, accesses and uses information" (Elliott, 2003a, 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 a learning engine, which obtains the information necessary by learning the 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 seeing its mistakes. Third, IntelliMetric is systemic and it is 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. Unlike other automated essay scoring systems, Intellimetric employs multiple judgments that rely on multiple mathematical models. It is claimed that while many scoring systems are based on the General Linear Model, IntelliMetric uses a nonlinear and multidimensional approach to analyze essays. It is claimed that 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 (Elliott, 2003a) .**

**IntelliMetric could be applied in "Instructional" or "Standardized Assessment" modes. The instructional mode assists students with revising and editing processes by providing feedback on overall performance and diagnostic feedback on rhetorical dimensions such as organization and analytical dimensions such as sentence structure in an essay (Elliot, 2001a, 2003a, & 2003d). Additionally, IntelliMetric includes a variety of editing and revision tools like spell checker, grammar checker, dictionary and thesaurus (Elliott, 2003a). IntelliMetric provides students with detailed diagnostic feedback on grammar, spelling, and conventions as well (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 (Elliot, 2001a, 2003a, & 2003d).**

**It is claimed that IntelliMetric provides scores as accurate as human experts do (Elliott, 2001a). It is also claimed that the agreement rate between human raters and IntelliMetric is as high as 97 percent- 99 percent of the time. The developers of IntelliMetric state that they are aware of the fact that there is no scoring method –no matter whether it is human or computerized- that is 100 percent reliable. IntelliMetric may not "catch" all of the inauthentic responses in an essay, yet it effectively (around 95 percent) "catches" these types of responses (Elliott, 2001a).**

**One of the best attributes of IntelliMetric is that it is capable of evaluating essay responses in multiple languages. The system has already been used to analyze essays in English, Spanish, Hebrew, and Bahasa. Currently, it is available for text evaluation in a variety of languages including Dutch, French, Portuguese, German, Italian, Arabic, and Japanese (Elliot, 2003d).**

### 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 go on writing on the topic to improve their writing proficiency (http://www.vantagelearning.com).**

**MY Access provides not only immediate diagnostic assessment of writing, but it also provides constructive multilingual feedback for ESL 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 to 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 (http://www.vantagelearning.com).**

**MY Access includes several features that can make the writing process more feasible and effective not only for students, but also for teachers. For instance, the program can provide with individualized multilingual feedback (i.e., Spanish and Chinese) on different genres of writing such as informative, narrative, literary, and persuasive essays. 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. However, the system cannot score the essays written on these prompts since it needs to be trained on about 300 prompts to be able to score those essays automatically.**

**MY Access also offers a variety of writing tools that stimulate essay writing for students. For example, "writing dashboard" gives students the opportunity to see their weekly progress. In addition, the model essays scored by IntelliMetric allow students to view essays at each score point. Another example is the "my portfolio" feature, in which students can view a list of completed assignments, scores, reports, comments, etc.**

**The final feature, teacher options, allows teachers to have the 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 essay along with the feedback provided by the system. Last but not least, the website includes parent letters in English, Spanish, and Chinese for teacher use so that they can involve parents in their children's learning process (http://www.vantagelearning.com).**

## SUMMARY AND DISCUSSION

**There have been several studies that searched for ways to apply technology to writing assessment. One way is to use AES systems to assess the writing performance (Hamp-Lyons, 2001). A learner needs to get feedback from the instructor and revise his/her writing accordingly (Burstein, Chodorow, & Leacock, 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. However, for a teacher who teaches large classes, this is quite a time consuming process, which might also affect the frequency of the writing assignments given in class. The reason for developing AES systems is not only to provide students with opportunities to practice writing, but also to provide them with quick and accurate feedback regarding grammatical errors, style, content, and organization (Burstein et al., 2003).**

**AES systems can be a great assistance to teachers in responding to large number of essays and assign frequent writing assignments without worrying about scoring the first and subsequent drafts.**

**The AES 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 to assess the quality of essays. Unlike PEG or**

**IEA, e-rater and IntelliMetric systems have instructional applications (Criterion and My Access) as well.**

**Both Criterion and MY Access contain some functions for not only native English speaking students, but also for non-native English speaking students. For instance, Criterion includes TOEFL (Test of English as a Foreign Language) topics and some features of MY Access can provide multilingual feedback (i.e., Spanish and Chinese). Finally, except for IEA, the remaining three AES systems are unable to detect plagiarism.**

**There are some similarities among the four AES systems as well. First of all, they all need to be trained on large numbers of essay samples in order to be able to evaluate the student essays effectively. Next, almost all systems provide holistic scoring along with feedback on various domains of writing. Furthermore, all four systems are claimed to be very accurate and valid. The inter-rater reliability between each system and expert human raters are found to be high (Attali, 2004; Burstein & Chodorow, 1999; Elliot, 2000a, 2000b, 2001c, 2002, 2003b, 2003c; Landauer, Laham, & Foltz, 2003; Landauer, Laham, Rehder, & Schreiner, 1997; Nichols, 2004; Page, 2003, 2004).**

**AES is a developing technology. Many AES systems are used to overcome time, cost, and generalizability issues in writing assessment. The search for excellence in machine scoring of essays is continuing and numerous studies are being conducted to increase the accuracy and effectiveness of the AES systems.**

## BIODATA and CONTACT ADDRESSES of AUTHOR

**Semire Dikli is a PhD candidate in the department of Middle and Secondary Education at the Florida State University. She is also teaching an undergraduate level course: Second Language Testing and Evaluation at the Florida State University. Her research interest is writing and technology.**

**Currently, she is writing her dissertation prospectus on Automated Essay Scoring (AES) in an English as a second language (ESL) setting. Ms. Dikli has attended several local, national and international conferences. A native of Turkey where she was a teacher of English as a Foreign Language (EFL), she is eager to return home armed with ways to dramatically improve the teaching and learning of writing and technology.**

**Semire DIKLI**
**Florida State University**
**Tel: (850) 576 38 66**
**E-mail: ssd0960@garnet.acns.fsu.edu**
**Adress: 345 Pennell Circle# 3**
**Tallahassee, FL 32310 USA**

## REFERENCES

***A description of a new AI system with superior learning capabilities*, Retrieved on June 06, 2004 at http://www.geocities.com/ainew.geo/index.html.**

**Attali, Y. (April, 2004). *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. (June, 2004). *Automated essay scoring with e-rater V.2.0*. Paper presented at the Conference of International Association for Educational Assessment (IAEA), Philadelphia, PA.**

Bereiter, C. (2003). Automated essay scoring: a cross disciplinary approach. In Mark D. Shermis and Jill C. Burstein (Eds.), *Foreword* (pp. vii- ix), Lawrence Erlbaum Associates: Mahwah, NJ.

Brill, E. & Mooney, R. (1997). An overview of emprical 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 Mark D. Shermis and Jill C. Burstein (Eds.). *Automated essay scoring: a cross disciplinary approach.* Mahwah, NJ: Lawrence Erlbaum Associates.

Burstein, J., & Chodorow, M. (June, 1999). Automated essay scoring for nonnative English speakers. *Proceedings of the ACL99 Workshop on Computer-Mediated Language Assessment and Evaluation of Natural Language Processin,* College Park, MD.

Burstein, J., Kukich, K., Wolff, S., Lu, C., & Chodorow, M. (April, 1998). Computer analysis of essays. *Proceedings of the NCME Symposium on Automated Scoring*, Montreal, Canada.

Burstein, J., Chodorow, M., & Leacock, C. (August, 2003). *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. & Marcu, D. (2000). *Benefits of modularity in an Automated Essay Scoring System* (ERIC reproduction service no TM 032 010).

Burstein, J. (2003). The e-rater scoring engine: Automated essay scoring with natural language processing. In M. D. Shermis & J. Burstein (Eds.). *Automated essay scoring: A cross-disciplinary perspective.* Mahwah, NJ: Lawrence Erlbaum Associates.

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.

Chodorow, M. & Burstein, J. (2004). Beyond essay length: evaluating e-rater's performance on TOEFL essays (Research report no 73). Princeton, NJ: Educational Testing Service (ETS).

Chung, K. W. K. & O'Neil, H. F. (1997). *Methodological approaches to online scoring of essays* (ERIC reproduction service no ED 418 101).

Educational Testing Service (ETS). (n.d.). *E-rater*. Retrieved on May 06, 2004 at www.ets.org/e-rater

Educational Testing Service (ETS). (n.d.). *Criterion*, Retrieved on May 06, 2004 at http://www.ets.org/criterion/ell/faq.html.

Elliot, S. (2000a). *A study of expert scoring and IntelliMetric scoring accuracy for imensional scoring of Grade 11 student writing responses* (RB- 397). Newtown, PA: Vantage Learning.

Elliot, S. (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.

Elliot, S. (2001a). *About IntelliMetric* (PB-540). Newtown, PA: Vantage Learning.

Elliot, S. (2001c). *Applying IntelliMetric Technology to the scoring of 3rd and 8th grade standardized writing assessments* (RB-524). Newtown, PA: Vantage Learning.

Elliot, S. (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.

Elliot, S. (2003a). *A true score study of 11th grade student writing responses using IntelliMetric Version 9.0* (RB-786). Newtown, PA: Vantage Learning.

Elliot, S. (2003b). *Assessing the accuracy of IntelliMetric for scoring a district- wide writing assessment* (RB-806). Newtown, PA: Vantage Learning.

Elliot, S. (2003c). *How does IntelliMetric score essay responses?* (RB-929). Newtown, PA: Vantage Learning.

Elliot, S. (2003d). 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.

Foltz, P. W., Laham, D. & Landauer, T. K. (1999). Automated Essay Scoring:Applications to Educational Technology. *Proceedings of EdMedia '99*. Retrieved on 5/15/04 from http://www-psych.nmsu.edu/ ~pfoltz/reprints/Edmedia99.html.

Hyland, F. (1998). The impact of teacher written feedback on individual writers. *Journal* of Second Language Writing, 7 (3), 255-286.

Kukich, K. (September/October, 2000). Beyond Automated Essay Scoring. In Marti A. Hearst (Ed)*, The debate on automated essay grading*. IEEE Intelligent systems, 27-31. Retrieved on November 12, 2004, http://que.info-science.uiowa.edu/~light/research/mypapers/autoGradingIEEE.pdf.

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). Mawhwah, NJ: Erlbaum.

Landauer, T. K., Laham, D., & Foltz, P. W. (September/ October, 2000). The Intelligent Essay Assessor. In Marti A. Hearst (Ed)*, The debate on automated essay grading*. IEEE Intelligent systems, 27- 31. Retrieved on 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 Mark D. Shermis and Jill C. Burstein (Eds.), *Automated Essay Scoring and Annotation of Essays with the Intelligent Essay Assessor.* Mahwah, NJ:Lawrence Erlbaum Associates.

Latent Semantic Analysis (LSA). (n.d.). Retrieved on May 8, 2004 from http://lsa.colorado.edu/whatis.html.

Lemaire, B. & Dessus, P. (2001). A system to assess the semantic content of student essays. *J. Educational Computing Research*, Vol. 24 (3), 305-306.

Murray, B. (1998). *The latest techno tool: essay grading computers*. American Psychological Association (APA), 8 (29). Retrieved 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.* Mahwah, NJ: Lawrence Erlbaum Associates.

**Nichols, P. D. (April, 2004). 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.* Mahwah, NJ: Lawrence Erlbaum Associates.**

**Pearson Knowledge Technologies (PKT) official website, http://www.knowledge-technologies.com.**

**Psotka, J, & Streeter, L.(n.d.) *Automatically critiquing writing for army educational settings.* Retrieved on December 02, 2004 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).**

**Salem, A. B. M. (2000). *The potential role of artificial intelligence technology in Education* (ERIC document reproduction service no ED 477 318).**

**Shermis, M. D. & Burstein, J. (2003). *Automated Essay Scoring: A Cross Disciplinary Perspective*. Mahwah, NJ: Lawrence Erlbaum Associates.**

**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., Raymat, M. V., & Barrera, F. (2003). *Assessing writing through the curriculum with automated essay scoring* (ERIC document reproduction service no ED 477 929).**

**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 at http://www.k-a-t.com/papers/essayscoring.pdf.**

**Vantage Learning. *(n.d.). My Access*. Retrieved on May 06, 2004 at http://www.vantagelearning.com.**