

# **Automated Essay Scoring (AES) in an English as a Second Language (ESL) Setting**

**Doctoral Dissertation Grant (DDG)  
2004-2005 Technology**

**Semire Dikli**

**Multilingual/Multicultural Education Program  
Department of Middle and Secondary Education  
Florida State University**

**Research advisor:  
Dr. Deborah Hasson**

## **Statement of research issue or problem and relationship to TIRF's current research priority**

As a former English as a foreign language (EFL) teacher in Turkey, I observed that neither my colleagues nor the students liked the writing classes. Teachers avoided teaching writing because responding to the student papers was quite time consuming. Moreover, being non-native speakers of English language, teachers hesitated providing feedback regarding the content due to their lack of knowledge about some words or expressions in English. Also, they didn't know how and when to give feedback. Finally, both teachers and the students paid less attention to writing since it was not included in the standardized English proficiency exams. The negative attitudes toward writing urged me to find out about more about this 'scary skill', and I started thinking about the role of the computer and technology in writing assessment. Could advancements in technology assist ESL/EFL teachers in responding to student compositions? Could computers be an aid for writing teachers as in reducing the workload on grading essays? While looking for an answer, I found out that there were several Automated Essay Scoring (AES) systems on the market. These powerful programs could provide students with instant feedback. Some companies were claiming that their programs had up to 98 percent agreement with human raters. If these systems are really effective, could they assist non-native English speaking writers to improve their writing skills?

I found out that there has been little research done regarding the non-native English students and most studies included native English speaking writers. Only a few studies conducted with respect to ESL contexts by the ETS and Vantage Learning. For instance, one of the studies that ETS conducted focused on the performance of e-rater on TWE essays that were written by non-native English speakers while the other study investigated the relationship between essay length and holistic scores assigned to TOEFL-CBT essays by e-rater. Furthermore, Vantage Learning conducted two different

studies based on the essays written in languages other than English (Bahasa and Hebrew). A number of studies are conducted mainly to compare human raters with the AES systems. These studies aimed to prove the accuracy and reliability of the AES systems with respect to the writing assessment (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).

This study is expected to fill in the gap in the research not only on ESL writing and technology, but also on AES. Unlike numerous studies conducted on AES, the focus on this study is not to evaluate the validity of the AES system, but to investigate whether or not the AES system assists ESL writers to improve their writing quality. Additionally, the current study is not interested in the feedback provided in different languages but the impact of the feedback on the writing improvement on ESL students. In other words, the participants in this study are from various linguistic backgrounds and they will be writing in English.

This study is strongly related to TIRF 2004-2005 DDG research priority topic regarding technology: the demonstrable effects of the use of computer-based technology on students' learning of English as a second or foreign language. Similarly, the current study investigates the impacts of AES technology in ESL settings. As mentioned before, there is limited research done regarding the effects of AES technology in the writing improvement of non-native English speaking students. TIRF's priority topic on technology is interested in studies that explore the "differences, if any, between the effects of informed instruction with technology versus informed instruction without such technology on students' achievement in English as a second or foreign language and possibly their content learning as well, where English is a language of instruction" (<http://www.tirfonline.org/call0405.html>). This study also explores whether there is a difference between the scores of the students who are exposed to an AES technology vs. the ones who received feedback and scoring from the teacher only. Finally, like one of the themes of TIRF's current priority, the current study has implications regarding the cost effectiveness of schools' investments in language learning technology to increase the ESL students' language skills. For instance, using AES technology in writing classes can be useful for both teachers and students in that it provides students with instant and frequent feedback as well as scoring and at the same time it reduces the workload of the teacher, particularly in responding to large number of essays.

### **Theoretical Background**

This study will be conducted to find out whether the AES systems facilitate the revision process in ESL writing and improve student writing quality. The following section will briefly summarize the historical development in process writing because the focus of this study is revising and drafting of essays, which is the core idea of process writing. The next section will briefly introduce the related concepts to AES and the AES systems because such systems do incorporate the cyclic nature of writing and revision, and because the treatment in the study will be an AES system.

#### *Research in Writing*

Until the 1970s, based on the structural view of language teaching, writing was seen as a final product. That is, the writer did not revise the text and the teacher's feedback referred only to surface corrections (Riley, 1995). The increasing popularity of the communicative approach shifted the paradigm toward the process approach (Leki,

1992). As Zamel (1982) indicated, related research had earlier been concerned with the composition itself, but now the composing process has become the point of attention.

Investigations have revealed that writing is the process of exploring and integrating ideas. Moreover, it is not linear, but 'recursive' (Silva, 1990). As the process goes on, writers discover thoughts, express them and reformulate them (Zamel, 1983). While discussing the process/product dichotomy, researchers find it necessary to define the good and poor writer accordingly. Johns (1990) claims that the writer is the creator who goes through the process of creating and producing discourse. Similarly, Riley (1995) claims that current research in writing supports the idea that English as a Second Language (ESL) students need to pay close attention to the concepts of the reading as well as the purpose and apply these concepts in their writings to become good writers.

Automated essay scoring (AES) systems such as E-rater and IntelliMetric have been developed to assist teachers in scoring student essays and providing feedback within seconds (<http://web.lexis-nexis.com>). As Burstein, Chodorow, and Leacock (2003) emphasize, in order to improve his/her writing skills, a learner needs to get feedback from the instructor and revise his/her writing accordingly. However, for a teacher who teach 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 automated tools for writing 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).

*What is Automated Essay Scoring (AES)?*

Automated essay scoring (AES) is a system that requires computer technology to evaluate and score essays instantly. Although research on automated essay scoring began in the 1960s, for a long time, it has mainly focused on the native English speaking writers. Recently researchers have started working on models that involve languages other than English (Shermis & Burstein, 2003). For instance, Educational Testing Service (ETS) is currently working on "computer-based corpora" that characterize language variation, both for subgroups who use non-standard dialects of English and nonnative speakers of English (Burstein & Chodorow, 1999). Additionally, Vantage learning developed programs that can provide non-native English speaking students feedback in their own languages, i.e. Spanish and Chinese.

AES programs rely on various machine learning methods such as Artificial Intelligence (AI), Natural Language Processing (NLP), and Latent Semantic Analysis (LSA) to provide instant feedback and scoring. AI focuses on designing intelligent machines that have the ability to imitate the human mind. NLP is one application of AI and it has been used to summarize texts and translate them to different languages for decades. Furthermore, LSA characterizes a word used in a sentence, passage, or essay based on the semantic associations (Lauder, Foltz, & Laham, 1998).

Four types of AES systems are widely used by testing companies, universities, and public schools. The first one is known as Project Essay Grade (PEG). Given an essay, PEG is supposed to predict the score that a human rater would possibly assign to a similar type of essay. The comparisons of predictions are based on correlations between human raters as well as human raters and PEG (Page, 2003). Another AES scoring system, Intelligent Essay Assessor (IEA), is based on Latent Semantic Analysis (LSA) engine. Unlike other approaches, LSA is claimed to focus on content rather than

mechanical features of writing, i.e. grammar, spelling, and punctuation (Landauer, Foltz, & Laham, 1998). An example of an automated essay scoring system that employs text-based NLP applications is e-rater. E-rater consists of three NLP-based modules including syntactic, discourse, and topical analysis modules, to identify the scoring guide. The syntactic analyzer tracks the syntactic variety in the essay, whereas, the discourse analyzer identifies the discourse elements such as cue words, argument developments in the text. Finally, the topical analyzer captures the vocabulary use or topic identification (Burstein, 2003). E-rater is designed to analyze essay writing performance instantly according to the writing features specified on the six-point holistic scoring rubric used by the human expert raters (Burstein, Kukich, Wolff, Lu, & Chodorow, 1998). E-rater's instructional application is called Criterion and it includes the main characteristics of both approaches: automated essay scoring and diagnostic feedback (Burstein et al., 1998). Like e-rater, IntelliMetric is also use artificial intelligence and natural language processing. The instructional application of IntelliMetric is MY Access. The MY Access program includes a variety of editing and revision tools like spell checker, grammar checker, dictionary and thesaurus (Elliott, 2003a). The program assists students with revising and editing processes by providing feedback on overall performance and diagnostic feedback on five dimensions of writing including focus and unity, development and elaboration, organization and structure, sentence structure, and mechanics and conventions (Elliot, 2001a, 2003a, & 2003d).

### **Research Methodology**

The AES program that will be used in this study is MY Access, which is supported by the IntelliMetric scoring system. MY Access is selected because of its availability. The producing company of the program, Vantage Learning, will provide the software and the research setting. Nevertheless, the study will be conducted independently. The research questions are the following:

1. Is there a difference in the essay scores between English as Second Language (ESL) students who are exposed to the automated essay scoring (AES) and teacher feedback system compared to the ones who receive teacher feedback and scoring only?
2. Is there any significant improvement in students' writing scores by condition in terms of five dimensions of writing (focus and unity, development and elaboration, organization and structure, sentence structure, mechanics and conventions) on each of the prompts assigned?

In this study, the data will be collected from adult ESL students who are attending to the Florida Community College at Jacksonville (FCCJ) during the fall semester of 2005. A total of 200 students from high intermediate or advanced classes will participate in this study. A true experimental design will be employed to collect the data. Within a class, half of the students will be randomly assigned to the control groups while the rest will be assigned to the experimental groups. The experimental groups will be exposed to the MY Access program and receive teacher feedback (not scoring), whereas the control groups will receive teacher feedback and scoring only. Students will be writing 2-3 drafts on up to eight essay prompts (topics) from the MY Access library in a 12-week period. Only the data from students who will have written an essay with at least two drafts on a minimum of five topics will be included in the study.

All students will be typing the essays on the computer in class in a lab session. While the students in the control groups will type their essays on a word document, the students in the experimental groups will use the MY Access website to type the essays. The students will be assigned eight essay prompts at the beginning of every week (or every other week). They will revise the essays in class on another day based on the feedback and scoring they have received from the teacher or from the program. Student access to the MY Access program out of class will be controlled by inactivating the program at the end of each class. Students in both conditions will have access to the same type of tools during the study. That is, all the tools but the spell checker will be turned off both on the MY Access program and the Microsoft Word program. Finally, both groups will be allowed to use a bilingual dictionary.

The data will be collected through various sources including pre-tests and post-tests, student essays on eight prompts (both first and subsequent drafts), analytic scoring and feedback that were assigned to the essays by the MY Access program as well as the teachers, teacher meetings/interviews, and surveys. Students will be assigned a pre-test (a writing prompt) at the beginning of the semester and a post-test (a similar writing prompt) at the end of the semester. The pre-tests will be used to sense what students' writing is like and the post-test will be used to determine whether there was a significant improvement in the writing scores of the students both in the experimental and control groups after the treatment. The pre- and post-tests will be scored by two raters to obtain inter-rater reliability based on a 6-point analytic scoring scale generated by the program. Teacher meetings and/or interviews will be conducted to ensure that all teachers implement the same curriculum, use the same scoring scale, and assign the same topics according to the pre-determined timeline. All teachers will be provided training about how to use the program and the scoring scale by the company prior to any intervention. A computer literacy survey, a demographic survey and an opinion survey will be utilized to investigate students' computer skills and to gather demographic information about both their backgrounds and their perceptions regarding the feedback they received.

The first question will be answered based on student scores on pre-tests and post-tests.. A t-test will be used to compare the results. The second question will be addressed through MANCOVA and post hoc analyses. The data will be collected from the first and second drafts on each of the eight topics that have been scored by the MY Access program or by the teacher during the semester. Descriptive statistics (mean, standard deviation, range) will be used to summarize the basic features of the data.

A pilot study will be conducted with only one class in the summer semester of 2005 and the main study will be conducted with a larger sample in fall 2005.

#### **Statement of implications of research for theory, policy, and/or practice**

There has been a paradigm shift in writing from the product approach toward the process approach. Process approach requires great amount of time from the teachers because of its iterative nature. Teachers need to read and respond to many drafts of student writing. AES systems can be a great assistance to the teachers in evaluating student essays, thus reducing the amount of effort toward correction, and providing a greater amount of time for developing content skills, i.e., organization, focus and unity.

It is possible to speculate on the extent that the AES systems would assist ESL writers. AES systems provide students the opportunity to submit as many drafts as they want and get instant scoring and feedback. The feedback on content skills, i.e.

organization, focus and meaning, they receive might help good writers pay attention to those aspects of writing. On the other hand, it might also be possible that poor writers would get lost in the information pool generated by the computer since the feedback provided by the program can sometimes be quite extensive.

It is hoped that this study will fill in a gap in the AES research since there has been almost no research done regarding the use of AES systems in ESL contexts. I believe the results of this study will contribute to the research in ESL writing by providing both ESL teachers with further insights regarding the use of AES programs in writing classes. The results will also be a great assistance to the English teachers in the world. The non-native speaking English teachers will feel more confident in teaching writing to their students.