REPUBLIC OF TURKEY
ÇAĞ UNIVERSITY
INSTITUTE OF SOCIAL SCIENCES
DEPARTMENT OF ENGLISH LANGUAGE TEACHING

AUTOMATED ESSAY SCORING SYSTEM: A RELIABILITY STUDY

THESIS BY
Ali DOĞAN

SUPERVISOR
Assoc. Prof. Dr. Şehnaz ŞAHİNKARAKAŞ

MASTER OF ARTS

MERSİN, March 2013

REPUCLIC OF TURKEY

ÇAĞ UNIVERSITY

DIRECTORSHIP OF THE INSTITUTE OF SOCIAL SCIENCES

We **certify** that thesis under the title of "**AUTOMATED ESSAY SCORING SYSTEM: A RELIABILITY STUDY**" is satisfactory for the award of the degree of **Master of Arts** in the Department of **English Language Teaching**.

.....
Supervisor- Head of Examining Committee: Assoc. Prof. Dr. Şehnaz ŞAHİNKARAKAŞ

.....
Member of Examining Committee: Assist. Prof. Dr. Erol KAHRAMAN

.....
Member of Examining Committee: Assist. Prof. Dr. Kim Raymond HUMISTON

I certify that this thesis conforms to formal standards of the Institute of Social Sciences.

.....
01 / 03 / 2013

Assoc. Prof. Dr. Haluk KORKMAZYÜREK
Director of Institute of Social Sciences

Note: The uncited usage of the reports, charts, figures and photographs in this thesis, whether original or quoted for mother sources is subject to the Law of Works of Arts and Thought. No: 5846.

II

ACKNOWLEDGEMENTS

It is a real pleasure to thank the people who helped me complete this thesis. First and foremost, I would like to express my special thanks and sincere gratitude to my advisor, Assoc. Prof. Dr. Şehnaz ŞAHİNKARAKAŞ, who always encouraged me to study, helped and guided me in completing my thesis through her invaluable suggestions, deep interest, endless assistance, constructive feedback and patience.

I would like to thank to my dearest colleagues; Kemal GÖNEN, Özlem YILDIZ, Yusuf UYAR, Duygu ÖZMEN, Neslihan GÜZEY, Selda DELİKTAŞ, Ahmet Erdost YASTIBAŞ, and Muhammed Turgay KAYIRAN for their help with the research process.

Special thanks go to my dearest son, daughter, and wife Leyla DOĞAN for their patience during this period.

Lastly, I would like to thank to my parents who always supported me spiritually throughout my life.

March 1, 2013

Ali DOĞAN

ÖZET

BİLGİSAYAR ÜZERİNDEN OTOMATİK OLARAK MAKALE ÖLÇME VE DEĞERLENDİRME: BİR GÜVENİLİRLİK ÖLÇME ÇALIŞMASI

ALİ DOĞAN

Yüksek Lisans Tezi, İngiliz Dili Eğitimi Anabilim Dalı
Tez Danışmanı: Doç. Dr. Şehnaz ŞAHİNKARAKAŞ
Mart 2013, 78 sayfa

İngilizce öğretiminde ölçme ve değerlendirme materyallerine her geçen gün yeni materyaller ekleniyor. İngilizce öğretiminde makale ölçme ve değerlendirme bilgisayar üzerinden yapılabilir mi sorusu 1966 yılında yanıtlanmaya çalışılmış ve günümüze kadar düzenli bir gelişim grafiği göstererek son zamanlarda özellikle Amerika ve Avrupa'da yaygın olarak kullanılan 'Automated Essay Scoring Systems (Bilgisayar üzerinden otomatik olarak makale ölçme ve değerlendirme sistemleri)' olarak İngilizce Öğretimi ölçme ve değerlendirme materyalleri arasında yerini almıştır.

Bu çalışmanın amacı, Zirve Universitesi YDYO'da yazma becerileri ölçme ve değerlendirme sisteminin yerine bilgisayar üzerinden otomatik olarak makale ölçme ve değerlendirme sisteminin kullanılabilirliğini araştırmaktır. Bu çalışma Zirve Üniversitesi YDYO' nda yapılmıştır. Katılımcılar ingilizce öğreniminin B1 seviyesinde Zirve Üniversitesi YDYO' da C Kurunda öğrenimlerine devam eden 50 kişilik bir öğrenci grubudur. Çalışmaya Zirve Üniversitesi YDYO C kuru final sınavı cevap kağıtlarının üç yazma becerileri okutmanı ve bir bilgisayar sistemi tarafından değerlendirilmesi ile başlanmış, elde edilen notların analizleri yapılmıştır. Çalışma üç gün sürmüştür. Bu araştırmada sayısal analiz yapılmıştır. Çalışmanın sonunda, Zirve Üniversitesi YDYO' nda kullanılmakta olan makale ölçme ve değerlendirme sisteminin daha fazla enerji, daha fazla zaman ve daha masraflı olduğu görülmüş, bilgisayar üzerinden otomatik olarak makale ölçme ve değerlendirme sisteminin Zirve Üniversitesi YDYO' na kullanılabilir bir sistem olduğu tavsiye edilmiştir.

Anahtar Kelimeler: Bilgisayar Üzerinden Otomatik Olarak Makale Ölçme
ve Değerlendirme, Bilgisayar Tabanlı Notlandırıcı,
İnsan Notlandırıcı, Yazma Becerilerinin Ölçme ve Değerlendirilmesi

ABSTRACT

AUTOMATED ESSAY SCORING SYSTEM: A RELIABILITY STUDY

Ali DOĞAN

Master of Arts, English Language Teaching
Supervisor: Assoc. Prof. Dr. Şehnaz ŞAHİNKARAKAŞ
March 2013, 78 pages

New materials have continuously been added to the assessment instruments in ELT day by day. The question of whether writing assessment in ELT can be done via e-raters was first addressed in 1996, and this system, which is commonly called "Automated Essay Scoring Systems" in especially America and Europe in recent years, has taken part in the field of assessment instruments of ELT with steady development.

The purpose of this study is to find out whether AES can supersede the writing assessment system that is used at The School of Foreign Languages at Zirve University. It is performed at The School of Foreign Languages at Zirve University. The participants of the study were a group of 50 students in level C which is the equivalent of B1. The beginning of the quantitative study includes the assessment of essays written by C level students at The School of Foreign Languages at Zirve University by three human raters and e-rater. After the study it was found that the writing assessment has been currently used at The School of Foreign Languages at Zirve University costs more energy, more time and it is more expensive. Thus, AES was suggested for use at The School of Foreign Languages at Zirve University which has proven to be more practicable.

Keywords: Automated Essay Scoring, E-rater, Human rater, Assessing Writing

ABBREVIATIONS

AES	**:**	Automated Essay Scoring
AESS	**:**	Automated Essay Scoring System
SFL	**:**	School of Foreign Languages
SPSS	**:**	Statistical Package for Social Scientists
YDYO	**:**	Yabancı Diller Yüksek Okulu
ELT	**:**	English Language Teaching
AI	**:**	Artificial Intelligence
ETS	**:**	Educational Testing Service
PEG	**:**	Project Essay Grader
IEA	**:**	Intelligent Essay AssessorTM
BETSY	**:**	Bayesian Essay Test Scoring SystemTM
LSA	**:**	Latent Semantic Analysis
NLP	**:**	Natural Language Processing
LSD	**:**	Latent Semantic Dimensions
NAEP	**:**	National Assessment of Educational Progress
CFA	**:**	Confirmatory Factor Analysis
GMAT	**:**	Graduate Management Admission Test
GRE	**:**	Graduate Record Examination
TOEFL	**:**	Test of English As A Foreign Language
TWE	**:**	Test of Written English
TOEIC	**:**	Test of English for International Communication

LIST OF TABLES

LIST OF FIGURES

TABLE OF CONTENTS

CHAPTER 3

CHAPTER 4

CHAPTER 5

CHAPTER 1

1. INTRODUCTION

1.1. Introduction

Automated Essay Scoring (AES) has become increasingly popular in academic institutions as an alternative solution to assess the writing skills in a fast, efficient, and accurate way. Automated essay scoring (AES) engines "employ computer technology to evaluate and score written prose. While some forms of writing, such as poetry, may never be covered, we estimate that approximately 90 percent of required writing in a typical college classroom can be evaluated using AES. Many universities and colleges are implementing, "one of the most innovative instructional technologies in use in college classrooms today: an online automated essay-scoring service from ETS called Criterion" (Williamson, 2003).

1.2. Background to the Study

Automated Essay Scoring 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 most widely used AES systems are Project Essay GraderTM (PEG), Intelligent Essay AssessorTM (IEA), E-Rater and CriterionSM, IntelliMetricTM and MY Access!®, and Bayesian Essay Test Scoring SystemTM (BETSY)

Although AES has become a more mainstream method of assessing essays there are critics who voice their skepticism and continuously point out the fallacies of such a grading system. Many individuals believe that such software programs are not sufficiently

sophisticated enough to imitate human intelligence thus leading to faults in the grade assessment of each paper (as cited in Williamson, 2003). Although criticisms continue to be voiced the broader assessment community suggests that automated scoring does have valid applications for the assessment of writing (as cited in Williamson, 2003).

The assessment of the writing skills of individuals in the world of academia can be a long and tedious process that may in fact, disrupt the performance of teachers and their ability to teach effectively. As a result, artificial intelligence has entered into the field of education and English as a method to resolve this sometimes repetitive and time-consuming process. Such Artificial Intelligence (A.I.) based scoring systems claims to bring relative efficiency to the automation of scoring essay by simulating human intelligence and behavior in the electronic system of a computer such software attempts to relieve the burden of grading from educators. This helps to enhance human performance in more significant areas such as the educators effectiveness in teaching the material that will be assessed on the essays. Automated Scoring as a writing assessment has been highly controversial while critics focus on the movement from indirect to direct measurement of writing many institutions are adapting their assessment methods to involve some sort of scoring system processed by artificial intelligence (Simon & Bennett, 2007).

Automated scoring technologies are finding wider acceptance among educators (Williamson, 2003). The software program Intellimetric, a scoring engine that assesses the skills of students on essays has been implemented by the Commonwealth of Pennsylvania. This engine used to score the writing of students on the state achievement examinations of Pennsylvania (as cited in Williamson, 2003). Numerous states have followed Pennsylvania as a model and have begun to implement their own Automated Essay Scoring software similar to that of Intellimetric (Williamson, 2003).

Many individuals who support such dramatic changes in assessing essays believe that not only does such automated scoring software help relieve the grading burden from the educators but also adds a level of consistency that sometimes humans cannot achieve. Many factors wear on individuals who are assessing papers including stress and exhaustion. These two factors result in a form of inconsistency from the grader. These concerns revolve around the unreliability of the evaluation of essay answers and those individuals who are assessing them (Shermis & Barrera, 2004). The automated software is not affected by such factors that human graders are prone to suffer, which increases consistency in the grading assessment of each individual student. The primary concern in the case of automation is if the task itself is computable rather than if the software is able to perform to the standards of the academic

institution implementing the system. In most cases, when it comes to the assessment of the essays and the writing skills of the students, the task itself is computable and thus the idea of implementing an automated essay scoring system is possible (Attali & Burstein,2006).

Although such automated scoring technologies are finding wider acceptance among academic institutions, the drive to find new tools and more accurate AES software mustn't diminish. Criterion has been recently updated to a 4.0 Version, which, "marks a turning point in the history and practice of writing instruction and remediation" (Williamson, 2003). The evolution of such software will only help to advance the validity of such programs and solidify its place in the world of academia as a popular method for assessing the writing skills of students.

1.3. Statement of the Problem

There are several major criticisms that question the benefits and success of Automated Scoring Systems. Primarily, critics believe that artificial intelligence is unable to provide the satisfying assessment skills that imitate human scoring on a satisfactory level. Skeptics believe that the "qualities and bases for human judgment of complex performances cannot be explained by a rubric" (Williamson, 2003). Critics state that AES fails in three main aspects; validity, accuracy, and consistency.

In terms of validity, the meaning of such a term is much more ambiguous than most terms. Contextually speaking, the definition of the term continuously evolves in order to incorporate the new tools and innovations that are seen in the sector of these automated assessment systems. As previously stated, the term "validity" loosely revolves around the idea of "how well [the program] does the job it is employed to do" (Williamson, 2003). When determining the validity of a program in assessing essay results, the study should try to assess the level of "interpretation of data arising from a specified procedure" within the parameters of the test guidelines (Williamson, 2003). Although the assessment of validity may be complex in nature in regards to automated essay scoring programs, honesty is a primary component in determining the level of validity such programs demonstrate. Simply put, validity is seen as measuring what you say you are measuring, "and that you have really thought through the importance of your measurement in considerable detail" (Williamson, 2003). Unfortunately, determining the validity of AES is an uphill battle from the beginning:

"In assessing students' skills in the critical area of effective communication, methods vary from objective tests to the often subjective evaluations of student documents…

Criticisms of such methods typically address the use of restricted norms, and failure to identify factors contributing to students' growth" (Williamson, 2003). Accuracy has been a major area of concern for instructors and educators thinking of implementing such a system but "AES scores predict as well or better than scores produced by human raters and yield a high degree of construct validity" (Williamson, 2003). Although the accuracy of such software can rival that of human raters, there are still vast improvements that can be made.

The most specific problems of The School of Foreign Languages at Zirve University can be grouped in two; in scoring writing period. The most specific problem is time because students learn their marks after Teachers and students score the writing problems. At The School of Foreign Languages at Zirve University the scoring writing procedure is a very hard, tiring, and long period. After doing the exam, essays first go to the first grader it takes some time to grade the papers because graders also have classes during the scoring period. Naturally, teachers are tired from the classes they gave during the day and there is a time limit to grade the essays. This situation demotivates the teachers and they are generally in a hurry to grade the essays. In addition, the same problems face the second grader. In the end, a teacher at The School of Foreign Languages at Zirve University has to score a minimum of forty five essays in three or four days. Moreover, the writing papers are generally graded by writing teachers and these teachers also have to grade the homework essays of their students and each writing teacher has generally a minimum of two writing classes. This is not easy for a writing teacher and it is really hard to find volunteers to teach writing at The School of Foreign Languages at Zirve University. On the other hand, students have some problems after they quit wondering about their writing marks and it demotivates students to write another essay since they haven't been informed about the essay they wrote beforehand. Another problem is the feedback problem as students can't have feedback after the exams.

Finally, consistency has also been a key issue for critics. Many critics believe that such programs lack consistency because of the subjective material being assessed in an objective manner. Again, research has proven that AES scores as well or better by human raters and for a longer period of time. Human raters succumb to such factors, like stress and exhaustion that hinder their performance and accuracy in assessing material as time passes. AES assesses papers in the same way as human raters and are not affected by such factors resulting in a more consistent rating program.

1.4. Significance of the Study

Although Automated Essay Scoring (AES) is becoming widely accepted as a method of assessing papers, the literature and research on the issue is not as vast other topics. AES has great benefits and we continue to expand the benefits of such innovative software while continuously improving how it assesses the writing skills of students and other individuals. The main way to accelerate this process is by continuously doing research and studies to provide constructive criticism and reveal the fallacies of such programs in order to help provide more valid and accurate ways of grading papers using artificial intelligence.

Automated scoring systems help to take the burden off educators. Such software would be a great benefit to The School of Foreign Languages at Zirve University if AESS properly assess essays. The grading burden on English instructors at the The School of Foreign Languages at Zirve University is alarmingly apparent. "Students become proficient writers through constant practice, but teachers have less time than ever to grade reading assignments," said John Oswald (Williamson, 2003). Many instructors are asked to teach twenty-five hours a week, attend weekly meetings, and grade tests and papers weekly. This hinders the performance of instructors, which as a result, reduces the effectiveness of their teaching. Such drop in performance hinders the student body from learning as much as possible. English is considered to be a difficult language to learn and even more so to speak and many instructors at The School of Foreign Languages at Zirve University believe that efficiency is key to struggle such tasks and obligations. Automated Essay Scoring (AES) will help to relieve the burden of assessing tests and weekly essays, and will enhance the performance of instructors and their effectiveness of teaching the material required of them.

There are two things that automated systems are potentially capable of doing. Primarily, these sort of software programs “can replicate scores for a particular reading of student writing and this technology is reliable, efficient, fast, and cheap. Two, automated scoring has been and will continue to be used in various large-scale assessments of student writing” (Williamson, 2003). If automated systems are able to perform these tasks with sufficiency and accuracy, it demonstrates absolute necessity of having such a automated scoring program at academic institutions.

1.5. Limitations of the Study

There are several limitations that need to be acknowledged and addressed regarding the present study, which intends to find out about the AESS. Firstly, this study has been carried out with the fifty students of intermediate level at The School of Foreign Languages at Zirve University. Second, three instructors from The School of Foreign Languages at Zirve University graded the essays as human raters. Finally, Criterion, the Online Writing Evaluation System, graded the essays as e-rater (AESS).

1.6. Research Questions

In this study the following questions will be evaluated:

1) How does Automated Essay Scoring (AES) work at the university level, particularly in a preparatory school environment?
2) Is AES valid and reliable?

CHAPTER 2

2. LITERATURE REVIEW

2.1. Types of Writing Tests

Writing is tested by two kinds of measurement, indirect and direct. In indirect testing, students are asked to respond to questions about composition often in a multiple-choice format. Indirect tests of writing are commonly referred to as objective. However, since human judgement is dominant in creating the set of questions and possible answers, Hamp-Lyons (1990) thinks indirect tests of writing are so-called objective and defines an indirect measure of writing thusly:

It does not require the test taker to write continuous prose although she or he may write some words, and there is no room for personal interpretation by the test taker since possible answers are provided and the 'correct' one already decided upon (p. 6). Direct test of writing became popular in 1970s with the emphasis on language as communication. According to Hamp-Lyons (1990), a direct test of writing has at least five characteristics:

1. Each individual actually, physically writes at least one piece of continuous text.
2. While the writer is provided with a set of instructions and material, s/he is given a considerable room within which to create a response to the prompt.
3. Each written text is read by at least one, usually more, human reader-judges who has been through some preparation or training for the evaluation process.
4. Each judgement made by readers is tied to some common standard measurement, such as a description of expected performance at certain levels or one or several rating scales.
5. Readers' responses to the writing are expressed as a number or numbers of some kind, and not written or verbal comments.

With the introduction of new approach to language as communication, direct measures of writing started to be used as the preferred means for assessing writing performance since they are closer to real discourse. Moreover, they state that the aspects of writing such as organization, coherence, and the elaboration of ideas which are not measured with indirect measures are evaluated thanks to writing samples. Jacobs et al. (1981) define the benefits of a direct test of writing as below:

1. emphasizes to learners the importance of language for communication
2. promotes a closer match between what is taught and what is tested
3. is more valid
4. is easier to prepare
5. produces more meaningful and interpretable results
6. can indicate level of proficiency and strengths and weakness in the writing skill
7. can be highly reliable if properly administered and evaluated
8. utilizes the important intuitive, albeit subjective, resources of other participants in the communication process-the readers of written discourse

2.2. Scoring Writing Methods

As direct tests of writing gained importance, the search for reliable and valid scoring was needed. Carlson and Bridgeman (1986) put forward the need to change the current scoring methods used with indirect tests and say:

With the development of competence in basic communication skills (writing, speaking, listening and reading) as a primary goal for education and with the recognition that many students pass through our educational system with inadequate English-language competence, educators are reappraising their methods and redefining their objectives (p. 126).

2.3. Holistic Scoring

Holistic scoring measures student writing for its overall quality. Trained readers use a set of instructions, called rubric, to lead their grading.

Holistic scoring is the most commonly used assessment tool in writing. As Huot (1990) mentions “many scholars see it as the major means of direct writing evaluation. Others contend that holistic scoring has proven to be the best economical, flexible and applicable of the direct writing instruments” (p. 201).

According to Gregory (1991) six reasons for the popularity of holistic scoring are:

1. Low cost, especially if compared with multiple-choice type of scoring. The biggest expense will be to raters but since most projects are brief it will not be very expensive. The efficiency of test administration: Tests can be administered in a 45-50 minute class period.

2. High reliability

3. The appeal of a holistic approach is to see things as units, as complete, and as wholes;
4. Holistic reading is thought to be face-to-face encounter because the writer's mind embodied in written expressions and reader's mind attempting to see what is being communicated.

In addition to these advantages "holistic scoring method has the advantage of being very rapid" (Hughes, 1989, p. 86). Hughes mentions that an experienced rater spends only a couple of minutes or even less to score a one-page essay. Some other researches like Carlson and Bridgeman (1986), Carlson, Bridgeman, Camp et al., (1985), Cooper (1977), Gregory (1991) and Mann (1988) agree that holistic scoring rarely takes more than two minutes per paper. However, some researchers believe that holistic scoring method is impressionistic and unreliable since ''the score must represent what a sophisticated reader interprets as a total effect'' (Lloyd-Jones, 1987, p. 164).

Perkins (1983) states in comparing scoring techniques that contradictory findings have been revealed in published research on reliability and concurrent validity issues of holistic scoring. The results of a study made by Diederich et al., 1974, show that ''out of the 300 essays graded, 101 received every grade from 1 to 9, 94% received either seven, eight or nine different grades; and no essay received less than five different grades from fifty three readers'' (p. 653).

Fortunately, training of the readers makes possible to reach high reliability. Mitchell and Anderson (1986) show the high reliability of holistic scoring they used in a study.

2.4. Raters

Wang and Brown, (2008) state in their study that there are two kinds of raters, human raters and e-raters, to grade the essays. The efficiency of automated essay scoring (AES) holds a strong appeal to institutions of higher education that are considering using standardized writing tests graded by AES for placement purposes or exit assessment purposes. However, it is not clear to what extent AES can replace human raters in judging the quality of essay writing. Research to date has mainly been conducted by testing agencies that market AES for commercial purposes. Companies such as Vantage Learning and ETS Technologies have published research results that demonstrate strong correlations and non-significant differences between AES and human scoring. However, the validity of AES tools is still a debatable issue. Some researchers criticized AES tools for their "over-reliance on surface features of responses, the insensitivity to the content of responses and to creativity,

and the vulnerability to new types of cheating and test-taking strategies" (Yang, Buckendahl, & Juszkiewicz, 2002, p. 393).

In foreign language writing assessment field, human rater has been predominantly used to evaluate writing samples both in large-scale test (such as College Entrance Exam) and classroom test. However, the human rater confronts unavoidable problems such as fatigue and inconsistency in score reliability. English writing researchers and assessors have considered using an e-rater to replace the classic method on the basis of high correlation findings and successful empirical studies in both first and foreign language writing research (August 29, 2010 by China Papers).

2.5. Human Rater

Human raters are generally the teachers, instructors, or specially trained people to grade the papers. It is very important to be trained about how to assess writing issues and especially human raters should know how to grade papers by the rubric. At The School of Foreign Languages at Zirve University human raters are not well trained but they are allowed to grade papers. In this study three instructors from The School of Foreign Languages at Zirve University took part as human raters and they graded the papers.

2.6. E-Rater

Page and Petersen (1995) explained e-raters as:

Project Essay Grade. The first automated essay scorer to be developed was Project Essay Grade (PEG). Although initial work on PEG began in the 1960s, some practical problems weren't solved until the microcomputer became popular in the late 1980s, at which time the Educational Testing Service (ETS) conducted a blind test of PEG for scoring 1,314 essays produced by students taking the Praxis test when applying for teacher certification programs. The results demonstrated that PEG was more accurate in predicting human ratings of the essays up to and including three human judges. The automated grading of essays thus was shown to be more accurate as well as more rapid and economical than the use of human judges.

2.7. IntelliMetric

A second automated essay scorer, IntelliMetric, also has proven to be highly effective (Vantage Learning, 2000). Initially made available to educational agencies in January 1998, it was the first essay-scoring tool based on artificial intelligence.

IntelliMetric relies on Vantage Learning's CogniSearch and Quantum of AES is the Intelligent Essay Assessor (IEA). Based on latent semantic analysis (LSA), the IEA is used to score the quality of conceptual content-based essays and creative narratives. Most important, LSA technology provides direct, content- based feedback to instructors or teachers (Landauer, Laham, and Foltz, 1997).

LSA provides a representation of an essay's semantic content as a vector (that is, a set of factor loadings) computed from a set of words contained in the essay. Each vector is compared with another through a cosine for comparing similarities (Landauer, Laham, and Foltz, 1997). The vector length is defined as the distance of each point from the origin.

The primary method of evaluation, 'holistic', compares an essay of unknown quality to a set of pre-scored essays (In 360,000 essays per year). The reported discrepancy rate on these massive sets of data has been less than 3 percent (Burstein et al., 2001), demonstrating that e-rater technology is a reliable measure of essay scores.

The e-rater scoring system using a six-point scale aims to implement features similar to those used in holistic scoring. To score on the higher end of the scale, an essay must remain consistent with its topic; have a strong, well-organized argument; have a strong syntactic structure; and use a diversity of words (Burstein et al., 2001). "E-rater features include discourse structure, syntactic structure, and analysis of vocabulary usage (topical analysis), . . [but does] not include direct measures of length, such as word count" (Burstein et al., 2001).

2.8. Automated Essay Scoring

Automated Essay Scoring (AES) is defined as the computer technology that evaluates and scores the written prose (Shermis, Raymat, and Barrera, 2003). AES systems are developed to assist teachers in low-stakes classroom assessment beside 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 (Burstein, 2003).

The results of a number of studies conducted to assess the accuracy and reliability of the AES systems reported high agreement rates between AES systems and human raters (Vantage Learning, 2000a, 2000b, 2001b, 2002, 2003a and 2003b). Although AES systems have been criticized for lacking human interaction, vulnerability to cheating, and their need for a large corpus of sample text to train systems, its popularity in public schools, universities, testing companies, researchers and educators is continuously growing (Rudner and Gagne, 2001).

Ramineni et al. (2012) claims that "of course, there are also challenges associated with automated essay scoring systems, such as ensuring adequate construct representation, the cost and effort of developing such systems, potential susceptibility of the systems to 'gaming' the scoring to maximize a score, and the need to validate their use for the intended purpose. However, when designed appropriately, automated scoring systems can allow a greater construct representation and authentic assessment, and may facilitate allowing some testing programs and learning environments to make greater use of constructed-response items where such items were previously too onerous to support because of the time and costs associated with human scoring" (p. 15).

Having been widely pursued, 10 different automated essay evaluation systems are now available for scoring and/or performance feedback. Automated Essay Scoring (AES) systems such as Intelligent Essay Assessor, e-rater, Project Essay Grade and IntelliMetric have become popular in recent years in both the classroom and in writing practice systems like MyAccess, WriteToLearn, and Criterion. The systems typically produce summary scores for writing assignments. In addition, the program allows for teachers to provide student-specific feedback. As a result, students can consult instructors regarding their specific feedback and make revisions on their writing based on that.

Such systems have also been used in testing and testing preparation situations such as the TOEFL, GRE, and ETS Proficiency test which offer practice exams online or in a facility. While these are graded by a human grader and AES system in real testing situations, practice tests are generally only scored through AES. For instance, the ScoreItNow site uses AES system analysis to give an idea of what one would score on the GRE writing section. In these practice situations a segmented analysis of the test writing is given, with sub-scores, annotations, and specific attention to errors and weakness. Moreover, AES systems were actually used in the real test environment for high competition exams such as the GMAT as

early as 1999. E-rater was used to score the writing portion of the exam. The frequency continued with their use in the GRE, TOEFL, and the Pearson Test of English.

AES is used in one of two ways when scoring exams. In both situations, a human score is necessary. In the first case, a human score and AES score are combined to produce an examinee's resultant score. Traditionally, two human scores are used. If there is a huge discrepancy between computer and human, within a predetermined range of uncertainty, then an additional human grading is required. In the second situation, a human score is made for the writing and the AES score is used only to check the accuracy of the human score. The human score is the only value reported. The AES is used as either confirmation or rejection of the human produced score. IEA is a particular exception to the rule that a human score operates in conjunction to an AES score. In the Pearson Test of English, IEA is the only score issued and there is no human score reported.

Although AES has been used with more frequency in the last few years, a few problems still exist. To begin, certain groups remain unconvinced that AES is useful or appropriate for teaching writing (Anson, 2003, Herrington and Moran, 2001). In comparison, human scoring requires labor, effort, and cost. Furthermore, there can be great discrepancies between human scorers (Huot, 2002, Huot and Neal, 2006, White, 1994). Scholars believe that AES could be a solution to these problems (Ramineni et al.2012).

AES systems generally use a sample of 500 to 2000 essays to create a model for evaluation. The number of essays used for calibrating the scoring system is typically based on the number of examinees for the test. Examinees, when taking the exam, are also required to compose their essay on a computer. Handwriting is not a viable option for AES systems because the technology required to analyze handwriting is not readily available and relatively inaccurate. The inaccuracies would heavily affect the grading process of AES.

Additionally, the samples from which evaluation scale are determined must accurately reflect the population that is going to be testing and the conditions under which they will be testing. If the examinees will be sixth grade students writing in an hour time period, then the sample exams upon which the AES scoring is calibrated upon must come from the same demographic under the same time constraint. Problems arise, however, when considering the specific nature of the location from which the samples are collected. Certain communities, cities, districts may have a population that is notably better or worse than the average population. This must be considered when selecting sample exams to base the scoring rubric upon.

After the exams to serve as samples are selected, those essays are divided into two categories. The two categories they are divided into may be comprised of an equal number of essays, or one category may have significantly more. This is determined randomly. The first category is the model building set. Highly unusual essays are removed and the remaining essays are used to calibrate the e-rater models.

Highly unusual essays are determined by advisory flags. These flags can pinpoint issues such as off-topic writing, repetition of the prompt, and repetitive language. There are many such flags. They are selected in accordance with a test's specific requirements. These flags also remove essays that should not be used for calibration of AES systems and also single out essays that must be human scored.

After the model-building group has been determined, points are assigned to the essays and primary features are extracted. Features determined in the model group are compared to the human scores given for the model group. After that, weight is assigned based on the comparison to noted features of the writing. Features are, as a rule, positive performances. Negative incidents in the writing are removed, and the model set is rerun for point values (Attali, Bridgeman, and Trapani, 2010).

Model set construction, under these conditions, typically adheres to two methods: prompt-specific modeling or generic modeling. The first is centered on the selection of writing prompts examinees will choose from. For example, ten prompts result in ten e-rater models. Topic specific vocabulary is included in the features. This increases specificity and typically produces higher score than the generic model, but requires a larger sample of writings for calibration. For example, rather than 1000 essays for the entire rubric, 1000 essays are required for each prompt to give an appropriate model set from which to calibrate the electronic scoring for each prompt.

The second model, the generic model, creates a single assessment for a variety of writing prompts and considers all writing prompts together, rather than individually. Generated under the generic condition, e-rater model, results in a single rubric scoring any number of different prompts. The generic model is created by taking a best fit of the regression of the features of the essays of 10 or more somewhat related prompts. Common points of interception and a common method of weighting features is designed under this model. For generic model building only 100-200 responses are required for each prompt. Moreover, newly generated prompts that fit into the design of the prompts under which the general model are determined are integrated into the assessment system easily. The uniformity of this model is beneficial. However, the elimination of specificity-dependent

features of the writing is a downside. Quality supersedes content here and the results of performance generated under this model are generally lower.

Scoring models rely heavily upon the content and intention of the exam. Generic models are dependent on the existence of similarity across prompt pools. Understandably, model selection is dependent most on the intended use of the exam. Academic admission and employments most likely supersede training and learning environments in their concern with the specificity of an assessment system (Ramineni et al.2012).

2.9. Project Essay Grader (PEG)

A relatively young field, the history of AES goes only fourty years back. Ellis Page, regarded as pioneer of AES, designed a computer-grading program named Project Essay Grader in 1966. Utilizing the statistical capabilities of computers, researchers looked for the kind of textual features that could be extracted by computers from the texts and then applied multiple linear regression to "determine an optimal combination of weighted features that best predicted the teachers' grades" (Kukich, 2000, p. 22).

Also, Kukich (2000) added that some of the features he identified as having predictive power included 'word length, essay length in words, number of commas, number of prepositions, and number of uncommon words - the later being negatively correlated with essay scores.

Page et al. (1994) use the terms 'trins' and 'proxes' while explaining the way PEG generates a score. While trins refer to the intrinsic variables. Proxes denote the approximation of the intrinsic variables. Thus, proxes refer to actual counts in an essay.

The scoring methodology of PEG is simple. The system has a training stage and a scoring stage. PEG is trained on a sample of essays at first. In the next stage, proxy variables (proxes) are ascertained for each essay and these variables are entered into the prediction equation. Finally, a score is appointed by computing beta weights (coefficients) from the training stage 100 to 400 sample essays are needed in PEG for training purposes (Chung and O'Neil, 1997).

One of the strengths of PEG is that the predicted scores are comparable to those of human raters. Moreover, the system can computationally follow the writing errors made by the users (Chung and O'Neil, 1997). However, PEG has been criticized for ignoring the semantic aspect of essays and focusing more on the surface structures (Chung and 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 accuracy. For example, since PEG used indirect measures of writing skill, it was possible to ‘trick‘ the system by writing longer essays (Kukich, 2000, p. 13). Some changes were made in 1990 including not only several parsers and various dictionaries, but also special collections and classification schemes (Page, 2003, Shermis and Barrera, 2002).

2.10. Intelli Metric

IntelliMetric, is the first essay-scoring tool that was based on artificial intelligence (AI). Like e-rater, IntelliMetric relies on NLP. It was developed by Vantage learning and used by the College Board for placement purposes.

IntelliMetric is a type of learning engine using a blend of artificial intelligence (AI), natural language processing (NLP), and statistical technologies. “It internalizes the pool wisdom of expert human rater” (Elliot, 2003, p. 71). IntelliMetric relies on Vantage Learning’s CogniSearch and Quantum Reasoning technologies. CogniSearch was specifically developed for use with IntelliMetric to understand natural language to support essay scoring. IntelliMetric internalizes each score point associated with certain characteristics in an essay response and then applies it to subsequent scoring by the system. It is claimed that the scoring system ‘learns’ the characteristics that human raters likely to grade.

IntelliMetric is trained with a set of pre-scored essays with known scores assigned by human raters. These essays are used as a foundation to extract the scoring scale and the wisdom of the human raters. The system has several steps to analyze essays. First, the system internalizes the known scores in a set of training essays. In the next step, the system tests the scoring model against a smaller set of essays with known scores to validate the scores. Finally, once the model scores the essays as desired, it is applied to new essays with unknown scores (Vantage Learning, 2001a, 2002, 2003b, 2003c).

IntelliMetric rates over 300 semantic-, syntactic-, and discourse-related features in an essay by using AI and NLP technologies. These texts related features are identified as larger categories called Latent Semantic Dimensions (LSD). The LSD features are described in five broad categories that are focus and unity, organization, development and elaboration, sentence structure and mechanics, and conventions.

2.11. E-Rater

The electronic essay rater (e-rater) was developed by the Educational Testing Service (ETS) to assess the quality of an essay by evaluating linguistic features in the text (Burstein, 2003, Burstein and Marcu, 2000).

"E-rater uses natural language processing and information retrieval to develop modules that capture features such as syntactic variety, topic content, and organization of ideas or rhetorical structures from a set of training essays pre-scored by expert raters. These combinations are processed into the computer program to score new essays" (Wang & Brown, 2008, p. 3).

The e-rater engine was initially used to grade the Analytical Writing Assessment part of the Graduate Management Admissions Test (GMAT). Researchers at ETS 'hypothesized' groups of NLP and information retrieval extractable linguistic features that might correlate with the GMAT grading criteria. The GMAT is currently scored by two human raters on a 6-point holistic scale, with 6 being the highest score and 1 the lowest. If the difference between two raters is more than 1 point, a third rater scores for resolution. The test-taker's final score is determined through e-rater and one human-rater. Similar to the prior practice, a second human-scorer is included if there is a discrepancy between e-rater and human scorer more than one point. E-rater is able to grade both GMAT with a high degree of reliability with human raters and other types of essays (Dikli, 2006, Wang and Brown, 2008).

E-rater is a computer program that scores essays primarily on the basis of writing quality by means of feature values computed using natural language processing (NLP) techniques. The primary features of the current version of e-rater are organization, development, grammar, usage, mechanics, style, average word length, median word frequency, positive features, and two content features. Several of these primary features are based on many different sub-features that represent different aspects of writing quality that contribute to the value of the primary feature. Organization and development are the features that measure text structure based on automatically identifying sentences in an essay as they cover essay-discourse categories: introductory material (background), thesis, main ideas, supporting ideas, and conclusion. The grammar, usage, mechanics and style features together analyze over 30 error types, including errors in subject-verb agreement, preposition errors, pronoun errors, article errors, sentence fragments, missing comma, wrong word form, repetition of words, etc. These error types are summarized for each feature as proportions of

error rates relative to the essay length. Some of these error features, such as article errors, are common errors of English language learners. Lexical complexity is measured by two features. The first feature is a word frequency index representing a measure of vocabulary level, and the second feature, average word length, computed as an index of word choice. In addition, two content features measuring topic-specific vocabulary usage are also part of the e-rater feature set, but can be excluded from the scoring models for some writing prompts depending upon the nature and purpose of the assignment (Shermis, Shneyderman, and Attali, 2008). The positive feature is a measure of the correct use of collocations and prepositions and represents efforts to develop features capable of measuring positive indicators of writing rather than focusing on errors (Ramineni et al.2012).

2.12. Criterion

Criterion is a web-based essay scoring and evaluation system, which relies on other ETS technologies called e-rater and Critique writing analysis tool. As a writing analysis tool, Critique has 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 (Dikli, 2006, p. 13). Besides providing instant holistic scoring, Criterion 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 Criterion is called an advisory component. The advisory component serves as a supplement to the e-rater score, but does not give 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, Criterion provides diagnostic feedback on grammar, usage, and mechanics; style and diction; and organization and development. Criterion includes 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 supplies writing at various levels including elementary school (4th and 5th grades), middle school (6th, 7th, 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 (TOEFL). The topics are obtained from authentic retired ETS essay topics. They are taken from various ETS testing instruments such as NAEP (National Assessment of Educational Progress), English Placement Test designed for California State University, Praxis, and TOEFL. Criterion can evaluate essays on the topic for which it had been 'trained'. A minimum of 465 essays scored by expert raters train the system on a topic. However, teachers are not restricted to use the topics in the Criterion library and they can find and assign their own topics. Holistic scoring cannot be recorded for teacher-created topics, but it is possible to obtain feedback an every dimension of writing (ETS, n.d.).

The electronic portfolio and writer's handbook features are for facilitating the writing process for the students. Students can store their first and subsequent drafts online with the electronic portfolio. With writer's handbook, on the other hand, students are able to view feedback definitions, examples of correct and incorrect use, and an explanation of every error reported. Teachers have control on several features of Criterion. They can control student access to the program by activating/inactivating the website or setting start/finish dates. Teachers can also check the student access to spell check, diagnostic feedback or holistic scoring by turning on/off these features. Finally, teachers have an option to record their own feedback within the student essay (ETS, n.d).

Besides its instructional use in classrooms, Criterion can also be used for remediation and placement purposes by the schools. Some schools use Criterion for benchmark testing. Some schools use the Criterion program for exit testing. In this case, both Criterion 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.).

2.13. Validity Issues on AES

A number of research studies have been conducted by the companies that developed the AES tools. Whereas most research has conducted those to demonstrate how well AES correlated to human raters' scoring, some researchers have also investigated the threats to the validity of AES, so as to improve the performance of AES tools. Furthermore, it has been aimed to explore the effectiveness of AES for assessment in the classroom, thus expanding the potential of using AES to benefit writing instruction.

As noted by Warschauer and Ware (2006), the range of correlations between scores

produced by AES tools and those assigned by human raters, comparable to the range of correlations between two human raters' scores, has generally been supported by psychometric research. This conclusion is also supported by some of the research projects on PEG and e-rater. To examine the effectiveness of PEG for rating specific traits of writing, Page, Poggio, and Keith conducted a study in 1997. By using a sample of 495 essays written by 12th graders for the writing assessment of the National Assessment of Educational Progress, they examined how well PEG would predict the average scores of eight raters as compared to the prediction rates of two, three and four human raters. The results showed that PEG outperformed the prediction rates of two human raters on all the trait rating scores as well as on holistic scores although PEG prediction rates were lower on holistic scores and on two traits: style and mechanics compared to the four-rater prediction rates.

Another study conducted by Shermis, Koch, Page, Keith, and Harrington (2002) tested the validity of PEG by using Confirmatory Factor Analysis (CFA). In this study, PEG scores were compared to scores assigned by all possible pairs of six human raters. Five different analyses were performed to avoid overlapping. The results showed that the standardized pattern coefficient for the human pairs ranged from .81 to .89, and the median coefficient was .86. However, for PEG, the coefficients ranged from .88 to .89 with a median coefficient of .89. These findings suggested that ''the computer ratings of essays were at least as valid as pairs of human judges'' (Shermis et al., 2002, p. 15).

Research on another AES tool, e-rater, also supported the validity of AES to a great extent. A study was conducted by Burstein, Kukich, Wolff, Lu, and Chodorow (1998) to look at the validity of e-rater. In this study, 500 Graduate Management Admissions Test (GMAT) essays and 200 Test of Written English (TWE) essays were scored by applying e-rater. The correlation analyses showed that e-rater had comparable correlation rates to those between the two human raters. Whereas the two human raters correlated with each other at rates ranging from .82 to .89 across the writing prompts, e-rater correlated with Rater 1 at rates ranging from .80 to .87 and with Rater 2 at rates ranging from .79 to .87 (Burstein, et al., 1998).

As distinct from others, this study also tried to look at the area of discrepancy – an area where the score difference went beyond one point difference. Researchers examined the rates of discrepancies between two human raters and between each human rater and e-rater at each score level. The results showed that at the score level of 5 and 6, the rates of discrepancy between e-rater and each human rater were higher than the rates of discrepancy between the two human raters. Whereas the rate of discrepancy between two human raters was 8% at a

score level of 5 and 7% at a score level of 6, the rate of discrepancy between e-rater and Rater1 was 15% at a score level of 5 and 34% at a score level of 6. Similar discrepancy rates existed when comparing e-rater with Rater 2 (15% at score level of 5, and 31% at a score level of 6) (www.ets.org/research/dload).

Powers, Burstein, Fowles, Chodorow, and Kukich carried out a study to investigate the limits of AES tools in 2001. These researchers designed a study that specifically probed the threats to the validity of e-rater. For the purpose of this study, various writing experts and critics of AES were invited and asked to write responses to the Graduate Record Examination (GRE) writing prompts. These participants were encouraged to write in any way that they thought would "trick" the e-rater into overestimating or underestimating their essays. Furthermore, participants were asked to explain what discrepancies they would predict and what would cause those discrepancies. Once the essays were written, both human raters and e-rater scored these essays by using the holistic scoring guide designed for the GRE writing test (Wang and Brown, 2007, p. 11).

Powers et al. (2001) found that 67% of the writing samples were correctly placed in the direction of score predictions (the mean scores assigned by e-rater were higher or lower than the mean scores given by the human raters, as predicted by the participants). Seventeen percent of the essays were placed in the wrong direction (their e-rater mean scores were higher or lower than human raters' mean scores when the predictions were the opposite). The other 17% of the essays had an e-rater rating exactly the same as human rating although these essays were predicted to a higher or lower rating than e-rater rating (Powers et al., 2001). The researchers cited an example to demonstrate how e-rater could be tricked. One of the participants, a professor of computational linguistics, wrote a few paragraphs and copied them 37 times. The human raters gave his essay a score of 1, whereas e-rater gave him a score of 6 – the highest score (Powers et al., 2001).

By looking at these findings, the researchers suggested that e-rater be used together with a human rater, and that further research focus on how to "identify excessively repetitive essays, as well as those that employ questionable logic" (Powers et al., 2001, p. 14).

Research on IntelliMetric was carried out to validate IntelliMetric as an effective AES tool. Scores assigned by IntelliMetric were compared with those given by human raters to determine the agreement rates and correlational coefficient rates. In almost all of these studies, results showed high agreement rates and high correlational coefficient rates (Greer, 2002, Vantage Learning, 2001a, Vantage Learning, 2002). Most recently, Rudner, Garcia, and Welch (2006) also revealed a correlational coefficient rate as high as .83 when they

examined the relationship between IntelliMetric scoring and human raters' scoring.

Nivens-Bower (2002) conducted a comparative study at two New England community colleges. Thirty students from introductory writing classes at both colleges took the WritePlacer Plus test. IntelliMetric and then two college faculty members from each college graded their essays by using the six-point scale WritePlacer rubric for scoring essays utilized by both studies. A paired-sample t test was run to compare the group means, and the Wilcoxon signed rank test was performed to examine the range of score frequencies. The paired-sample t test revealed no significant differences in mean scores at the level of .05 and .01 (t value not reported). No significant difference in the range of score frequencies was shown by the Wilcoxon signed rank test (W value not reported). In light of these results, as well as the high correlational coefficient rates, Nivens-Bower (2002) concluded that IntelliMetric "produced results consistent with what would be expected of faculty scores" (Nivens-Bower, 2002, p. 12).

In the comparative study conducted by Vantage Learning in 2003, instructional literary analysis essays were graded by IntelliMetric (Vantage Learning, 2003). The Vantage Learning researcher collected 400 written responses from 9th and 10th grade students in English classes (the school and its location were unspecified). 350 of these samples were for training IntelliMetric and 50 for validation. Two human raters graded all the responses first. Then IntelliMetric was trained by the 350 expert-scored essays, and finally the trained scoring model of IntelliMetric evaluated the remaining 50 essays. No significant difference between the mean score assigned by the experts and the mean score assigned by IntelliMetric ($t = .265$, $p < .05$) was shown by the significance testing. The mean score averaged from human expert scoring was 2.98 with a standard deviation of 1.26 while the mean score averaged from IntelliMetric scoring was 3.18 with a standard deviation of 1.38. In addition, it revealed high agreement and high correlation coefficient rates. Based on these results, the researcher concluded that IntelliMetric performance in scoring essays in instructional environments "exceeded the performance typically found with expert scorers" (Vantage Learning, 2003, p. 6).

CHAPTER 3

3. METHODOLOGY

3.1. Introduction

The goals of this study are to find out whether AES can be used as an assessing tool at The School of Foreign Languages at Zirve University and to figure out if it is valid and reliable in assessment of writing papers. This chapter specifies the methodology of the study. First, the participants who contributed to the study are described. Then, the materials to collect data, the way the data were collected and how the scores were given are explained and presented.

3.2. Participants

After necessary permission was obtained from the school management, the study was conducted at The School of Foreign Languages at Zirve University in Gaziantep in the first term of the academic year 2011-2012 with the participation of the level C students. There were 50 students aged between eighteen and twenty-six years who participated in the study. The participants' language level was B1. The participants were randomly selected for this study.

3.3. Instruments

In this study fifty writing exam papers (essays) were scored first by three human raters, and second by e-rater. The papers were the final writing exams of the level C students at The School of Foreign Languages at Zirve University. Human raters and the e-rater used the same rubric while scoring the papers.

In the study, Criterion which is the name of the e-rater was used in this study . Criterion is the AESS widely used in the U.S.A and Europe. It is basically used as a writing assessment and writing e-portfolio material. The goal of Criterion is short a regular, valuable, efficient and rapid writing education. Criterion is a a holistic rubric and is prepared according to the 1-6 scale. Despite these features, Criterion is not an actively studied material in Turkey yet. The features of each point are tabled depending on the can-do statements belonging to each point.

Table 1: The Criterion Rubric

Students scored with 1 by Criterion rubric are thought to perform the can-do statements below and they are defined as Unsatisfactory.
• Attempts a response, but may only paraphrase the prompt or be extremely brief. • Exhibits no control over organization. • Exhibits no control over sentence structure. •Contains inaccurate word choices throughout most of the essay. •Little effort is made to persuade because there is no position taken. • Is characterized by misspellings, missing words, and incorrect word order; errors in grammar and conventions severely impede understanding throughout the essay.
Students scored with 2 by Criterion rubric are thought to perform the can-do statements below and they are defined as Insufficient.
• Provides little information and makes little attempt at development. • Is very disorganized or too brief to detect organization. • Exhibits little control over sentence structure. • Contains inaccurate word choices in much of the essay. • Is characterized by misspellings, missing words, and incorrect word order; errors in grammar and conventions are severe enough to make understanding very difficult in much of the essay. • Either a position is not clearly given or little attempt is made at persuasion.
Students scored with 3 by Criterion rubric are thought to perform the can-do statements below and they are defined as Uneven.

• Provides limited or incomplete information; may be list-like or have the quality of an outline. • Is disorganized or provides a disjointed sequence of information. • Exhibits uneven control over sentence structure. • May have some inaccurate word choices. • Contains errors in grammar and conventions that sometimes interfere with understanding. • While a position is stated, either it is unclear or undeveloped.
Students scored with 4 by Criterion rubric are thought to perform the can-do statements below and they are defined as Sufficient.
• Provides clear ideas, but sparsely developed; may have few details. • Provides a clear sequence of information; provides pieces of information that are generally related to each other. • Generally has simple sentences; may exhibit uneven control over sentence structure. • Consists mainly of simple word choices, but may contain some specific word choices. • Contains errors in grammar and conventions that generally do not interfere with understanding. • States a position and adequately attempts to persuade the reader.
Students scored with 5 by Criterion rubric are thought to perform the can-do statements below and they are defined as Skillful.
• Develops ideas with some specific, relevant details. • Is clearly organized; information is presented in an orderly way, but essay may lack transitions. • Exhibits some variety in sentence structure. • Displays some specific word choices. • May contain some errors in grammar and conventions; errors do not interfere with understanding.

Clearly states the position and persuades the reader. Students scored with 6 by Criterion rubric are thought to perform the can-do statements below and they are defined as Excellent.
• Develops ideas well and uses many specific, relevant details throughout the essay. • Is well organized with clear transitions; maintains focus. • Sustains varied sentence structure. • Exhibits many specific word choices. • Contains little or no errors in grammar and conventions; errors do not interfere with understanding. • Clearly states the position and effectively persuades the reader of validity of argument.

3.4. Research Design and Process

For this study, the first step was quite challenging because finding an e-rater program in Turkey was impossible. The reason was this field of study hasn't developed in Turkey or in the adjacent countries. Thus, several countries like France, England and Spain were applied to have permission to use the e-rater system, Criterion. However, after a long time, it was obtained from the U.S.A. A training program was provided for the human raters who took part in the study to get accustomed to the Criterion rubric. With the availability of the system, the process took only one week, and human raters and e-rater assessed the papers in three days. The rubric used for this study was Criterion's 0-6 scales holistic one. Besides, to explore the research questions of the study, quantitative data collection instruments were used.

3.5. Data Analysis

For the first research question, data was gathered from the AESS. The papers of the final writing exams of the level C students at The School of Foreign Languages at Zirve University was scored by AESS and the human raters. It took 4 days to grade the papers for the human raters, but it took just 1 hour for AESS. The duration difference is really clear between the human raters and the e-rater. Before the human raters graded the 50 essays of the students, a training session was done with the rubric and 7 essays were graded together using

the AESS rubric. This rubric is a 0 to 6 scale holistic rubric. After the training, human raters were given 4 days to grade the papers. For the second research question, the reliability and validity were measured. For the last research question, e-rater and human raters were compared to one another in regards to accuracy, consistency, and reliability.

For same reason, one of the essays couldn't be graded by the AESS. In the AESS result screen 'N/A' (not applicable) was written. There could be several reasons for this situation. The essay could be plagiarized, but in that case, the system warns against unoriginal texts. For the second possibility, the format or the style might be problematic, but again in this case, the system codes the paper as 'wrong format'. Therefore, 49 essays were assessed by the e-rater, but only one couldn't be put through that process. After having the results of the human raters and the e-rater, the data was uploaded to the SPSS (Statistical Package for the Social Sciences) in order to get the analyses reports.

3.6. Procedure

The most discouraging and time-wasting part of the study was finding the AESS Criterion. It was because the program hasn't been used by any organisation or school, even as a demo version. Firstly, the search area was Turkey, but after the talks with one of the American teachers at Bahçeşehir University, it became clear that the program has been commonly used in the U.S.A and Europe. She mentioned her past experience with the program which she had the chance to use for only two months before leaving university in France, and she added some valuable information, like the Criterion office. The answers to research questions of the study were said to be accessible with this program by this teacher. The contacts in Paris couldn't help us through they guided us to the authorized ones in America who are responsible for Europe. After some correspondences, I had the right to use AESS Criterion for limited duration at the end of a tiresome two-month period.Upon having the right of usage of AESS Criterion, I got the required permission of having research at The School of Foreign Languages at Zirve University where I currently work (see app. 2). Next phase was finding the human raters, and I succeeded in finding them as my three co-workers teaching writing helped me in this work.

For the study, 50 writing essays of C level final test were taken from Testing Office. Because of the analytic rubric for assessing writings at The School of Foreign Languages at Zirve University, holistic rubric was provided to the human raters to reassess the essays.
Due to the lack of foundations or schools using Criterion in Turkey, it was impossible to find

anyone who had applicable knowledge of the program. Through the help of Criterion manuel on the Internet and Criterion office, I learnt of how the system could be used.

3.7. Data Analysis Environment: Criterion

In this study Criterion, an automated essay scoring system, is used.

3.7.1. What is Criterion?

Criterion is an Online Writing Evaluation, a web-based, instructor-driven, comprehensive instructional system that helps students plan, write, and revise essays. Instructors can create their own topics or select from the Criterion topic library at any level from fourth grade through upper-level college. It is important to choose the appropriate level for the students while using Criterion, otherwise the acquired result will be invalid and the whole process will be in vain. For that reason, setting the level of students in the association using the system same makes the process work faultlessly.

The system provides annotated diagnostic feedback and a holistic score to both student and instructor within 20 seconds so that students can quickly review, revise and resubmit essays. Instructors can also provide their own feedback within the system, as well as view summary reports of Criterion feedback.

Students can access a password-protected website, plan their essays, and submit them on topics assigned by their instructor. Users receive online diagnostic feedback to analyze elements of grammar, usage, mechanics, style, organization and development, and an overall holistic score from e-rater, a proprietary automated scoring engine developed by Educational Testing Service. Potential errors from these categories are identified within the essay itself, and targeted feedback is offered for each identified error.

The Criterion Online Writing Evaluation captures and displays actual errors and features within the student's essay, including fragments, run-on sentences, agreement errors, misspellings, sentences containing errors of commonly confused words, and sentences missing initial capitalization and final punctuation. Heuristic-based diagnostic feedback helps students focus on their errors and features as they revise their essays, while a work-in-progress revision capability allows them to make revisions as they review each category of feedback. Students can ask instructors for advice about an error or feature by writing questions and comments on any given assignment. Instructors can also insert their

own comments about the essay both within the essay and in a message board. The Criterion system also provides students with online access to a level-specific Writer's Handbook with reference materials on how to improve specific aspects of their writing. All of this feedback is designed so that the student can utilize it along with the revision capability.

All of the student's writing, scores, annotated feedback, and instructor comments are saved to a secure virtual portfolio that both the instructor and student can access.

3.7.2. How to use Criterion

The automated essay system, Criterion, is used as following the the steps below.

3.7.3. Registering The System

Accession to the Criterion system is through the Criterion website. It is a paid program, and everything is executed online. In the website of Criterion, any necessary information is available for the students, teachers and administrators. If this is seen inadequate, then online, mail, telephone and visual backings are offered.

3.7.4. Accounts

Criterion has three varieties of account. The first one;

3.7.5. Admin Account

Admin account is supplied to the one(s) who will manage the system on behalf of the purchaser after the buying process is completed. By using ''add new student, edit school information, delete school, email all administrators in school, email all instructors in school, email all students in school, advanced import, export report data, archive portfolios, edit my user information, school administrator options, instructor options, and search'', those using the admin account have the options of adding-deleting classes- instructors- students, seeing all, correcting, deleting, adding, commenting, reporting, examining reports etc... That is, they have the right to do anything on behalf of the purchaser.

3.7.6. Instructor Account

Instructor account is given by the administrator account to the ones teaching. After logging into the account, they have the right to use the following options, ‘‘add new class, edit class information, delete class, batch print reports, email all students in class, import student information, export report data, archive portfolios, edit my user information, add me to an existing class, delete me from a class, and search’’.

3.7.7. Student Accounts

Student account is given by the administrator account to the ones who will be having education. After logging into the account, students are faced with different options; the first one is to view your portfolio for the class, and the other is assignments given by the instructor.

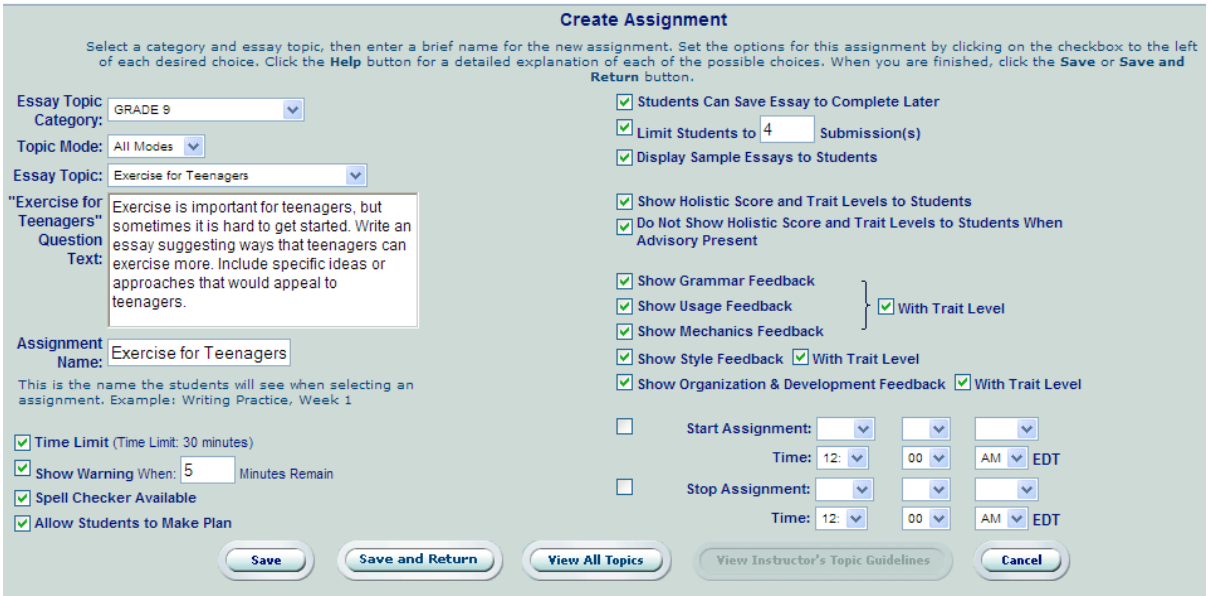
3.7.8. E-portfolio

In addition to being an online writing assessment tool, Criterion is also a writing e-portfolio.

3.7.9. Giving assignment

There are some steps to be taken for giving assignments to the students. The first one is a very crucial point for Using Topic Library AESS. The topic library in Criterion is quite extensive. In Using Topic Library there are many topics appropriate for all levels and essay types, their suitability to the writing criteria was tested, and they were applied to certain groups in different times. What the instructor needs to do after which type of essay is going to be used is choosing the suitable topic supplied by Using Topic library. For the topics in this section, there is also an explanation part, and all topics are easy for the students to understand clearly. As seen in Figure 1, Creating Assignment screen is prepared as a check list.

Figure 1: Creating assignment screen



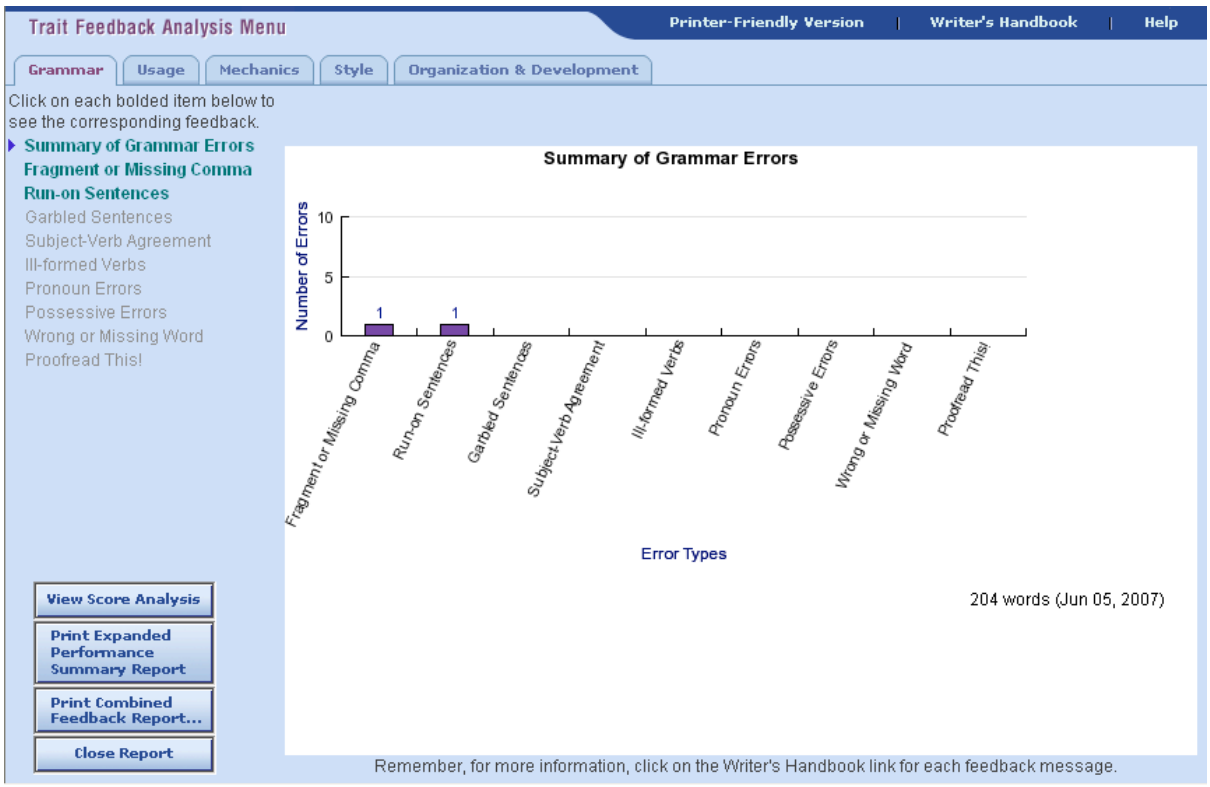
Thanks to this feature, one doesn't need to have a deep computer literacy. Creating Assignment is very essential for the system. Principally, the assignment in the Essay Topic Category should be chosen suitable for the level of students. On the contrary case, all assessments and studies will be invalid. In Topic mode, all mode or suitable topic modes in the menu can be chosen. When this is completed, next step is writing clear explanations about what you want the students to write consisting of 3 or 4 sentences in Question text section. Then comes the time limit part where we set the required time for the task, and we set a reminder about the minutes remaining. Next step is 'make a plan' option valuable for the students (See in figure 1). Students can choose this part to make a brainstorm and outline plan template.

The first type of plans is an outline (see in appendix 4) which is a good way to help organize ideas about a topic. A student can list main ideas first and then, under each main idea, can list some examples, reasons, or details that help support the main idea. The second type is the list(see in app 5). Making a list can help a student get started. Using the list plan, the student thinks about the topic and then lists any words, ideas, or examples that could be used in the essay. Up to 20 rows can be added to the list about a topic. Next essay planning type is the Idea tree (See in app 6). If the main ideas about the topic are already known, it

may be helpful for the student to put them in an idea tree. This plan can help the student see where examples, details, and reasons are needed to support the main ideas. The template contains five main idea columns, with space for four support elements for each main idea. The screen can be scrolled horizontally and vertically. Sometimes just writing down the first ideas that come to mind after reading the topic is a good way to help a student think of even more ideas. If an instructor wants students to use a template which is not one of the eight in the system, students can create a plan using a printed template provided by the instructor, and enter the completed plan into the 'Free Writing' template (as shown in app 7). Putting ideas and examples into an idea web can help a student see how they are related to one another. The student can try starting with some main ideas and then fill in the examples, details, or other ideas that will help support the main ideas (see in app 8). Using The Compare & Contrast (Attribute Tables) type of plan can help a student organize ideas about ways in which the two things are alike and ways in which they are different (see in app 9). If the essay is about why something happened or about what might happen in the future, using Cause & Effect (Fishbone Mapping) type of plan can help explain the causes behind a particular effect or result. This type of plan can also be used to show the opposite: the different effects that might be caused by a particular event (see in app. 10-11). If the assignment requires you to explain opinions or views about a topic, The Persuasive (Argument Diagram) type of plan helps to list and organize the main ideas (arguments). Space is also provided to list examples and/or reasons that support each argument.

Besides, there is an option to let the students complete the task at the same time or complete half of the task and do the rest of it after a while. Another functional option is that it is adjustable how many times a student can submit. The opportunity to have feedback after submitting and rewriting it with the help of feedback is very helpful for motivation. Moreover, instructor can share one of the samples of the assignment if he wants, and it is also important to have an example for the students. This system gives two different kinds of feedback. The first one is 1-6 scale holistic rubric (see in appendix 1) and the second one is Trait Feedback Analysis consisting of Usage Feedback, Mechanics Feedback, Style Feedback, Organization Feedback, Development Feedback and Grammar Feedback both of which are quite purposive for enhancing the writing process of students.

Figure 2: Trait feedback analysis menu



The Criterion system generates specific feedback regarding student submissions. The Trait Feedback Analysis shows the errors the student has made, explains why they are errors, and offers suggestions as to how to correct them. Diagnostics are generated for the following traits:

Table 2: The list of main and sub-traits used to grade essays in Criterion.

Main Trait	*Sub-traits*
Grammar	*Fragment or Missing Comma* *Run-on Sentences* *Garbled Sentences* *Subject-Verb Agreement* *I'll-formed Verbs* *Pronoun Errors* *Possessive Errors* *Wrong or Missing* *Word Proofread This!*
Style	*Repetition of Words* *Inappropriate Words or Phrases* *Sentences Beginning with Coordinating Conjunctions* *Too Many Short Sentences* *Too Many Long Sentences* *Passive Voice*
Organization and Development	*Introductory Material* *Thesis Statement* *Topic Relationship & Technical Quality* *Main Ideas* *Supporting Ideas* *Conclusion* *Transitional Words and Phrases Other*

Mechanics	*Spelling* *Capitalize Proper Nouns* *Missing Initial Capital Letter in a Sentence* *Missing Question Mark* *Missing Final Punctuation* *Missing Apostrophe* *Missing Comma* *Hyphen Error* *Fused Words* *Compound Words* *Duplicates*
Usage	*Wrong Article* *Missing or Extra Article* *Confused Words* *Wrong Form of Word* *Faulty Comparisons* *Preposition Error* *Nonstandard Verb or Word Form* *Negation Error*

3.8. CREATING REPORTS

The reports in Criterion system are divided into two. The first one is the instructor report in which instructors can create a number of reports containing information about Classes, Students, etc. from the Main Navigation screen by using the drop-down menu in the Select a Report field. This feature has great importance for the students' writing processes. There are eight different detailed report types such as classes report, classes access information report, errors report, holistic score summary, school roster report, student access information report, students report, class roster report. Thanks to these reports, students can have a detailed understanding of which parts should be improved and where they are incapable. The reports and their functions are as follows:

1. Classes Report: Appears on Instructor's Home Page as the default report listing the instructor's classes. Displays the names of the classes, the last instructor log-in date and time, the number of students registered for each class, and the total number of submissions per class.

2. Classes Access Information Report: Displays the names of the classes which the Instructor has created, as well as the access IDs and passwords assigned to the classes.

3. Errors Report: Shows data for administrator assignments giving the percentage of essays that have the same types of errors on any or all assignments. Instructors can use this information to focus lessons on areas that need further study. A similar report provides the same information for all assignments within a specific class.

4. Holistic Score Summary: Provides graphs showing the percent of essays at each score for all classes as well as the number of essays with each score. A button also allows instructors to see essays which received an advisory, but no score. The color coding and descriptions in the graphs are based on standards defined at the Administrator level. Instructors can use this information to help target lessons to students who are at different proficiency levels.

5. School Roster Report: Summarizes each student's work for all of an instructor's classes by showing the assignment name, holistic score, and number of errors/comments for each trait feedback category.

6. Student Access Information Report: Lists the names of all students in the instructor's classes who have access to the Criterion system, as well as their IDs, passwords, and last log-in date. Instructors can quickly monitor how frequently students are accessing the system and whether they are keeping up with assignments.

7. Students Report: Display students' names, the title of their most recent or any specific assignment, the date/time each student last accessed the assignment, the number of submissions each student made for the assignment, the holistic score achieved by each student, and notification of the existence of new comments.

8. Class Roster Report: Summarizes each student's work for a particular class by showing the assignment name, holistic score, and number of errors/comments for each trait feedback category.

3.9. Reports for Student Use

Criterion system has the second report type as student reports. Students can create three types reports containing information about their essays, errors, progress such as Submitted Essays Report. The following reports are available to students:

1. **Submitted Essays Report:** Documents the assignments given to the student; the date the student submitted each assignment; the number of times the student submitted an assignment; comments made by the instructor to the student; and the holistic score the student received on the most recent submission of the assignment.

2. **Errors Report:** Counts words and errors/comments by category for the most recent assignment, a selected assignment, or all the assignments. If there are no errors/comments for a particular category, a message indicating so will be displayed. The report shows data for the most recent attempt for the assignment.

3. **Progress Report:** Shows the student's progress over time at both the Holistic Score and Trait Level.

3.10. Using the Text Editor Option

The Text Editor option allows teachers and students to receive full trait feedback on writing that does not respond to a specific topic or prompt. Students can use text editor to provide themselves with full feedback for any submitted writing sample on any topic, though without a score. Instructors can create more open-ended assignments which do not require students to write on a specific topic or prompt. These open-ended assignments will receive full feedback but no score. In the create assignment screen, instructor should select the desired grade/level against which the student submission will be evaluated by using the drop-down menu in the essay topic category field. In the second, updated create assignment screen that appears, instructor can enter any text he wishes in the enter essay prompt field. He should use this option to create generic, open- ended assignments like "Write a

descriptive essay about anything you choose" or "Write a persuasive essay on a favorite issue."

The second problem that was faced in the process of study was finding the Criterion rubric. However, after a period of 15 days, Criterion rubric was available (see app. 1). Criterion 1-6 scale holistic rubric was used. Next step was having a training with the human raters and 10 essays were assessed with this rubric. It was realized that more essays should have been used for assessing. The essays of the students were typed again as soft and true copies. This was because the essays for AESS system need to be in soft copy. The process of assessing in AESS Criterion took about one hour, so the system was found to be more efficient in terms of speed, energy, and time than human raters. The results were analyzed in SPSS and inter-rater reliability was assessed in Pearsons R. At the end of the study done at The School of Foreign Languages at Zirve University, AESS Criterion was seen as more practical, faster and energy saving than classical writing assessment . As long as it was used in the correct way, AESS Criterion was more reliable than any other human rater.

3.11. Conclusion

The participants, materials, research design and procedure and data analysis process were overviewed in detail in this chapter. In addition to these, the challenges faced during the process were mentioned to help further studies and prevent them from getting into trouble with the possible hardships of this field.

CHAPTER 4

4. FINDINGS

4.1. Introduction

This chapter will focus on the analysis of the data gathered by the human raters grades and e-raters grade in an order determined by the research questions. The analysis of each data collection tool will be mentioned in different sections supported by the excerpts taken from the data collection tools. In the end the results of the findings will be discussed.

4.2. The Usability of Automated Essay Scoring (AES) at the University Level, Particularly in a Preparatory School Environment

This study was conducted to clarify whether AESS is practicable at The School of Foreign Languages at Zirve University. The Process of AESS can be analyzed in three topics; substructure, portfolio and assessment.

4.2.1. Substructure

A certain substructure is needed to use Criterion, one of the common programs of AESS. Therefore, institutions lacking the required substructure cannot use Criterion, which seems as a weak point for the system. To use the system, teachers and students primarily need to have personal computers. As Criterion is a process-based system, the users mentioned above can study in detail whereas this is not an indispensable requirement. Should the institution supply students or teachers with a computer laboratory that might be satisfactory? Nevertheless, it is not possible to benefit from Criterion effectively if the timetable of labs is problematic. To give an example, there are four different steps to complete writing an essay. Having these steps at certain intervals may increase the motivation and success rate of students. A two-hour period might have a negative impact on the students if the process is speeded up. Giving computers to students and teachers is not enough; the Internet with an average speed needs to be provided. All sections of logging in and out are processed online in the system. Criterion is an online system requiring no installments. It can be used in any computers having access to the Internet. Another topic is purchasing Criterion system institutionally. It is a system which is commonly used in America and Europe and rewarded many times of all AESS. Criterion is a paid system which is not currently used in any of the institutions in Turkey. This has also affected this study negatively. However, as far as it is understood from the researches done in the study, the

system will come into use in a few years. The first step for this is having institutions or organizations willing to use the system, and the second is establishing companies which can provide purchasing and education services. According to this study, the first step mentioned above is ready. There are many schools willing to use this system. Regarding the second one, some companies have interest on this issue. The system will possibly be for sale in our country in a few years.

The last and the most important element of usability of Criterion is user seminars. Criterion is very essential for both students and teachers. Ones having an average knowledge of the system may have seminars for Criterion usage. Others lacking the knowledge of using computers need to get it before the seminars. After getting ready about the computer literacy, 'How to use Criterion' seminars should be given to the students and teachers as their accounts differ. These seminars have to be both practical and informative. Finally, teachers should know how to use students' accounts right along with their accounts.

4.2.2.Portfolio

Next step after completing substructure is writing essays. Right after the teachers provide the homework with proper topics, students should follow these steps in order of brainstorming, outlining, writing the first draft, and submitting the essay. The time between giving and submitting homework is called portfolio. All steps from the first point of learning the topic to the submitting point are recorded for teachers to assess. These steps are as in the following:

4.2.3.Brainstorming

The assignments given by the teachers are available for a certain period of time and date decided by the teacher himself. After getting the assignment, students need to research the topic in detail to get enough information. Then comes the brainstorming point in which students focus themselves on a certain way. To give an example, ''Write an essay explaining the reasons why do people go abroad?'' is given to the students. Students should research the topic and organize 'the reasons why people go abroad' by adding their background and knowledge.

The following are examples;

Reasons why people go abroad

* People go abroad for education
* People go abroad for job opportunities

* People go abroad for health problems
* People go abroad for travel & adventure
* People go abroad for holiday
* People go abroad because of war
* etc...

4.2.4.Outlining

Students enlarge the ideas they find in brainstorming by adding supporting details and choosing a suitable outline template in Criterion system.

4.2.5. Writing The First Draft and Submitting the Final Draft

Students write 1st draft upon completing the outline. While writing the first draft, students have the chance to correct their spelling and grammar mistakes with the help of spell checker and grammar checker. Then students send their first drafts to the teachers, who give feedback. Having received the feedbacks, students make the required corrections and send in the final draft. For the last point, they upload their final draft to the system. With the completion of upload, students can see their student report, and so do the teachers.

This study was conducted step by step as mentioned above at The School of Foreign Languages at Zirve University. The hardest part of the study was to find the Criterion system and learn how it works.

Besides its instructional use in classrooms, Criterion can also be used for remediation and placement purposes by the schools. Some schools use Criterion for benchmark testing. Some schools use the Criterion program for exit testing. In this case, both Criterion 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.). Consequently, AESS is practicable in any university foreign language schools in terms of usage and implementation.

4.3. The Validity and Reliability of AES

Validity refers to how well an instrument measures what it is supposed to measure. One key to rubric validity is carefully selecting criteria that match the concepts and skills taught. Criterion rubric is designed by ETS introducing themselves as:

‘‘We advance quality and equity in education for people worldwide by creating assessments based on rigorous research.’’ ETS develops, administers and scores more than 50 million tests annually; including the TOEFL® and TOEIC® tests, the GRE® General and Subject Tests and The Praxis Series™ assessments, in more than 180 countries, at more than 9,000 locations worldwide. So the validity and reliability of Criterion is checked every time by linguistics and academics at assessment center of ETS. It is clear in the results of the study that Criterion is reliable. Reliability is the extent to which an instrument yields consistent results when used repeatedly under the same conditions. When two different graders use the reliable rubric on the same performance, they will give similar scores. This is called inter-rater reliability. It is found in the study that the points human raters gave and the ones of e-raters are consistent as understood by comparing the tables of both raters.

Table 3: *The Analysis of the Three Human Raters and E-rater*

Correlations

		rater1	rater2	rater3	erater
rater1	Pearson Correlation	1	,583*	,652*	,716*
	Sig. (2-tailed)		,000	,000	,000
	N	49	49	49	49
rater2	Pearson Correlation	,583*	1	,641*	,712*
	Sig. (2-tailed)	,000		,000	,000
	N	49	49	49	49
rater3	Pearson Correlation	,652*	,641*	1	,890*
	Sig. (2-tailed)	,000	,000		,000
	N	49	49	49	49
erater	Pearson Correlation	,716*	,712*	,890*	1
	Sig. (2-tailed)	,000	,000	,000	
	N	49	49	49	49

*. Correlation is significant at the 0.01 level (2-tailed).

In the table, the analysis of the three human raters and e-rater is presented. As it can be inferred from the detailed numbers in the table, the correlation between the human raters 1, 2 and e-rater is weak whereas this correlation is significant between the e-rater and human rater 3, and it has an impact on the reliability. 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 (Dikli, 2006, p. 27).

In spite of the training given to the three human raters before the study, the reliability acquired was low. Having graded 8 essays according to the 0-6 scale holistic rubric together, three human raters started grading 50 essays provided by the study. As the human raters at The School of Foreign Languages at Zirve University were accustomed to the analytic rubrics used by School of Foreign Languages, the training of holistic rubrics to assess/ grade the writing papers failed to increase the reliability value of the study. During the process, a great need for well-trained raters was realized. To hinder this problem in further studies, intense training programs will be needed. The more training a rater receives about the rubric, the more reliable the study becomes, as noted in the previous table.

Table 4: ***Correlation Among the Human Raters***

Correlations

		rater1	rater2	rater3
rater1	Pearson Correlation	1	,583*	,652*
	Sig. (2-tailed)		,000	,000
	N	49	49	49
rater2	Pearson Correlation	,583*	1	,641*
	Sig. (2-tailed)	,000		,000
	N	49	49	49
rater3	Pearson Correlation	,652*	,641*	1
	Sig. (2-tailed)	,000	,000	
	N	49	49	49

*. Correlation is significant at the 0.01 level (2-tailed).

In table 4 the correlation among the human raters can be seen. Like the previous correlation, this one is also weak. At The School of Foreign Languages at Zirve University, essays are graded by three human graders, and the mean of these grades makes up the final score of the writing part. As a result of this table, it is possible to claim that there are reliability problems at The School of Foreign Languages at Zirve University in assessment and evaluation of the essays. Whether analytic rubrics or holistic ones are used during the

essay assessment process, it is clear that teachers should have an intense period of training to be fully qualified to assess any type of essays and any types of rubrics.

Table 5: ***Reliability Score Between the Mean Number of Human Raters and E-rater***

Correlations

		e-rater	rater total
e-rater	Pearson Correlation	1	,890*
	Sig. (2-tailed)		,000
	N	49	49
rater total	Pearson Correlation	,890*	1
	Sig. (2-tailed)	,000	
	N	49	49

*. Correlation is significant at the 0.01 level (2-tailed).

Unlike the other two tables, it is remarkable that the reliability score between the mean number of human raters and e-rater is high in table three. Despite the difference in the previous tables, in the last one the correlation numbers which are showing the reliability value are same. Therefore, it can be concluded that the correlation between the average scoring of three human raters and e-rater's result are significant.

CHAPTER 5

5. DISCUSSIONS AND CONCLUSION

5.1. Introduction

This chapter will focus on the discussion. Also, the limitation of the study and implications for the future studies will be mentioned.

5.2. The Usability of Automated Essay Scoring (AES) at the University Level, Particularly in a Preparatory School Environment

There are some requirements to use AES at universities to assess writing. These are seminars about how to use AES, access the Internet, personal computers, and a computer laboratory at schools. Carefully planned lab times are arranged if the system is used in computer laboratories at schools because timing may be demotivating for the students if they are forced to do writing activities in a limited time. The second issue about AES is that institutions or universities have to buy the system if they want to use it affectively. In this study, Criterion was used as an AES tool to assess the students' writing. It is commonly used in America and Europe. It enables students to brainstorm about a given topic, share their ideas, outline their ideas, write and get feedback from their teachers. Upon finishing writing their final draft, they upload and get their results in detail in a report prepared by the system. The final process takes a few minutes, so it is not time-consuming, but time-saving. It is beneficial for both students and teachers. For teachers, it allows teachers to follow and monitor their students' writing process, and give individual and detailed feedback by using the report given by the system. For students, they can get immediate feedback about their writing in detail, so they can see their improvements in their writing by matching teacher and system feedback and also improve their writing by analyzing teachers' feedback for their first draft and system feedback for their final draft. Also, the system minimizes the time that is required for assessing students' writing. Three human raters in the study assessed and evaluated fifty pieces of writings in twenty days, but Criterion did the same job in one hour. To conclude, if the requirements of the system are met, how the system works is understood, AES can be used effectively at university level.

5.3. The Validity and Reliability of AES

The second research question concentrates on the validity and reliability of AES. Though the correlation between human raters is weak, the correlation of the mean number of human raters and e-rater is the same and comparable, which means AES measures what it is supposed to measure (Burstein, Kukich, Wolff, Lu, & Chodorow, 1998). Also, it is known that AES tools like Criterion and IntelliMetric are developed with the help of experts on the field (Burstein et al., 1998; Vantage Learning, 2003). The experts are given sample writing tests/papers and assess them. Then the systems analyze their grading to structure the way they assess writings. The systems develop their own database. After the database is completed and developed, the experts and systems assess writings different from the ones used to form the database. Their results are matched, so the validity of the systems is measured.

Though the correlation between the mean number of human raters and e-rater is the same, the correlation between human raters and the correlation between e-rater and each human rater varies as it is supported in Burstein et al. Criterion has a high correlation with the third human rater, but a weak one with the first and second human raters while human raters have a weak correlation with each other. The weak correlation between human raters may result from the different types of rubric used in the institution and in the system, fatigue, loss of attention, negative or positive mood, which results in great discrepancies between human raters (Huot, 2002; Huot & Neal, 2006; White, 1994). Therefore, the results are not consistent with each other. In spite of the differences in terms of correlations between each human rater and e-rater, and between each human rater, there is a high agreement between the correlations of the mean number of the raters. This indicates that AES (Criterion) is an accurate and reliable assessment tool that can be used at institutions (Vantage Learning, 2000a, 2000b, 2001, 2002, 2003a and 2003b).

5.4. Conclusion

When writing tests are assessed and evaluated by Criterion (an AES tool), the results are reliable because there is consistency between the average score of Criterion and human raters. Also, this consistency supports the validity of the system. Therefore, Criterion is a valid and reliable tool to assess writing, and this makes it a tool that can be used at institutions. What is important is to meet the requirements of the system, to teach students

and teachers about how to use the system, and to organize the settings for the use of the system. If these are done, institutions can use the system effectively.

5.5. Limitations of the Study

The present study has the following limitations.

1. AESS is a paid system. For the system, either the students or the institution willing to use, should make the payment.
2. The institution willing to use the system should either provide a computer laboratory in which each student can have access to the Internet or every student should have his own computer.
3. For the AESS to be used effectively, both the teachers and students who are going to use it need to have an intense training about the program.
4. In order to use the system, there should be companies selling these kinds of systems because internationally AESS is seen as business.
5. To use the system it is possible to get in touch with international companies; however, continual support for training and the system might be very difficult. The reason behind this difficulty is that the trainers coming from other countries have to work in Turkey.
6. The lack of studies and researches about AESS in Turkey forms one of the serious problems.
7. In any institution in which AESS will be used, the administration has to be innovative, ready to follow educational technologies and spend money on necessary basis. The foreseeing managers who can adopt the innovations of this century play a vitally important role in the usage and extending of systems like AESS.
8. The institutions willing to use the system have to be patient inasmuch as both the teachers and students need time to use the system.

5.6. Implications for Future Research

AESS is not only for assessment but also used for e-portfolio. The study was only performed as part of the essay grading, and e-portfolio wasn't studied. A study about the usage of AESS for e-portfolio can be done for further studies. All the participants (students, teachers, graders, and administrations) attending the studies of AESS can contribute to the study to be faster and more reliable if they have training.

6. REFERENCES

Anson, C.M. (2003). Responding to and assessing student writing: The uses and limits of technology. In P. Takayoshi & B. Huot (Eds.), Teaching writing with computers: An introduction (pp. 234–245). New York: Houghton Miðin Company.

Attali, Y., Bridgeman, B., & Trapani, C. (2010). Performance of a generic approach in automated essay scoring. Journal of Technology.

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* (pp. 113–122). Mahwah, NJ: Lawrence Erlbaum Associates, Inc.

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

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.

Carlson, J. G., Bridgeman, B., Camp, R. and Waanders, J. 1985).

Carlson, S. B., Bridgeman, B. (1986). Testing ESL student writers. In K. L. Greenberg, H. S. Wiener, & R. A Donovan (Eds.), Writing assessment: Issues and strategies (pp. 126-152). NY: Longman.

China Papers. (n.d.). Retrieved September 16, 2012 from, http://www.chinapapers.com.

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

Cooper, C. R., & Odell, L. (Eds.). (1977). Evaluating writing: Describing, measuring, judging. Urbana, IL: NCTE.

Diederich, P. B., French, J. W., & Carlton, S. T. (1961). Factors in judgments of writing ability (ETS research bulletin RB-61-15). Princeton, NJ: Educational Testing Service.

Dikli, S. (2006). An Overview of Automated Scoring of Essays. *Journal of Technology, Learning, and Assessment, 5(1).* Retrieved July 15, 2012, from http://www.jtla.org.

Educational Testing Service (ETS). (n.d.). Retrieved on August 17, 2012, from http://www.ets.org.

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.

Greer, G. (2002). Are computer scores on essays the same as essay scores from human experts? In Vantage Learning, *Establishing WritePlacer Validity: A summary of studies* (pp. 10–11). (RB-781). Yardley, PA: Author.

Gregory, K. (1991). More than a decade's highlight? The holistic scoring consensus and the need for change. (ERIC Document Reproduction Service No. ED 328 594).

Hamp-Lyons, L. (1990). Basic concepts. In L. Hamp-Lyons (Ed.), Assessing second language writing in academic contexts (pp. 5-15). Norwood, NJ: Ablex.

Herrington, A., & Moran, C. (2001). What happens when machines read our students' writing? *College English, 63(4),* 480–499.

Hughes, A. (1989). Testing for language teachers. NY: Cambridge University Press.

Huot, B. (1990). Reliability, validity and holistic scoring: What we know and what we need to know. College Composition and Communication, 41(2), 201-213.

Huot, B. (2002). *(Re)Articulating writing assessment for teaching and learning.* Logan, UT: Utah State University Press.

Huot, B., & Neal, M. (2006). Writing assessment: A techno-history. In Charles MacArthur, Steve Graham, & Jill Fitzgerald (Eds.), *Handbook of writing research* (pp. 417–432). New York: Guildford Publications. in Education: Principles, Policy and Practice, 15, 91–105.

Jacobs, H. L., Zinkgraf, S. A., Wormuth, D. R., Hartfiel, V. F., & Hughey, J. B. (1981). Testing ESL composition: A practical approach. Rowley, MA: Newbury House.

Klobucar, A., Deane, P., Elliot, N., Ramineni, C., Deess, P., & Rudniy, A. (2012). Automated essay scoring and the search for valid writing assessment. In: C. Bazerman, C. Dean, J. Early, K. Lunsford, S. Null, P. Rogers, & A. Stansell (Eds.), International advances in writing research: Cultures, places, measures (p. 103-119). Fort Collins, Colorado/Anderson, SC: WAC Clearinghouse/Parlor Press http://wac.colostate.edu/books/wrab2011/chapter6.pdf.

Kukich, K. (2000). Beyond Automated Essay Scoring. *IEEE Intelligent Systems, 15(5),* 22–27.

Learning & Assessment, 10. Retrieved from http://ejournals.bc.edu/ojs/index.php/jtla/article/view/1603/1455

Lloyd-Jones, R. (1987). Test of writing ability. In G. Tate (Ed.), Teaching composition (pp. 155-176). Texas Christian University Press. Mance. San Francisco, CA: Jossey-Bass.

Mann, R. (1988). Measuring writing competency. (ERIC Document Reproduction Service No. ED 295 695).

Mitchell, K. & Anderson, J. (1986). Reliability of holistic scoring for the MCAT essay. Educational and Psychological Measurement, 46, 771-775.

Nivens-Bower, C. (2002). Faculty-WritePlacer *Plus* score comparisons. In Vantage Learning, *Establishing WritePlacer Validity: A summary of studies* (p. 12). (RB-781).

Yardley, PA: Author. Norusis, M. J. (2004). SPSS 12.0 guide to data analysis. Upper Saddle River, NJ: Prentice Hall.

Admission test scores to writing performance of native and nonnative speakers of English TOEFL Research Report #19). Princeton, NJ: Educa- tional Testing Service.

Page, E. B. (1994). Computer Grading of Student Prose, Using Modern Concepts and Software, *Journal of Experimental Education, 62,* 127–142.

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. & Petersen, N. S. (1995). The computer moves into essay grading: Updating the ancient test. *Phi Delta Kappan, 76,* 561–565.

Page, E.B., Poggio, J.P., & Keith, T.Z. (1997, March). *Computer analysis of student essays: Finding trait diðerences in student profile.* Paper presented at the Annual Meeting of the American Educational Research Association, Chicago, IL. (ERIC Document Reproduction Service No. ED411316).

Perkins, K. (1983). On the use of composition scoring techniques, objective measure, and objective tests to evaluate ESL writing ability. TESOL Quarterly, 17(4), 651-671.

Powers, D.E., Burstein, J.C., Fowles, M.E., & Kukich, K. (2001, March). *Stumping e-rater: Challenging the validity of automated essay scoring.* (GRE Board Research Report No. 98-08bP). Princeton, NJ: Educational Testing Service. Retrieved November 15, 2012, from http://www.ets.org/Media/Research/pdf/RR-01-03-Powers.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.M., Garcia, V., & Welch, C. (2006). An evaluation of the IntelliMetricTM essay scoring system. *Journal of Technology, Learning, and Assessment, 4(4).* Retrieved August 6, 2012, from http://escholarship.bc.edu/jtla/vol4/4/ 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).

Shermis, M. D., Shneyderman, A., & Attali, Y. (2008). How important is content in the ratings of essay assessments? Assessment Shermis, M.D., Koch, C.M., Page, E. B., Keith, T.Z., & Harrington, S. (2002). Trait ratings for automated essay grading. *Educational and Psychological Measurement, 62(1),* 5–18.

Shermis, M.D., Koch, C.M., Page, E. B., Keith, T.Z., & Harrington, S. (2002). Trait ratings for automated essay grading. *Educational and Psychological Measurement, 62(1),* 5–18.

Technology and Teacher Education, 8 (4). Retrieved from http://www.citejournal.org/vol8/iss4/languagearts/article1.cfm.

Vantage Learning (2001a). Applying IntelliMetricTM to the scoring of entry- level college student essays. (RB-539). Retrieved September 1, 2012, from http://www.vantagelearning.com/content_pages/research.html.

Vantage Learning (2002). *A study of expert scoring, standard human scoring and IntelliMetricTM scoring accuracy for statewide eighth grade writing responses.* (RB-726). Retrieved September 10, 2012, from http://www.vantagelearning.com/content_pages/research.html.

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. (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. (2003). *A comparison of IntelliMetricTM and expert scoring for the evaluation of literature essays.* (RB-793). Retrieved December 2, 2012, from http://www.vantagelearning. com/content_pages/research.html.

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.

Wang, J., & Brown, M. S. (2008). Automated essay scoring versus human scoring: A correlational study.

Wang, Jinhao, & Brown, Michelle Stallone. (2007). Automated essay scoring versus human scoring: A comparative study. *The Journal of Technology.*

Warschauer, M. & Ware, P. (2006). Automated writing evaluation: Defining the classroom research agenda. Language Teaching Research, 10, 2, 1–24. Retrieved June 18, 2012, from http://www.gse.uci.edu/faculty/markw/awe.pdf.

White, E. M. (1994). Teaching and assessing writing: Recent advances in understanding, evaluating, and improving student performance.

Yang, Y., Buckendahl, C.W., Juszkiewicz, P.J., & Bhola, D.S. (2002). A review of strategies for validating computer automated scoring. *Applied Measurement in Education, 15(4),* 391–412.

7. APPENDIXES

7.1. Appendix 1: Criterion Rubric

Score of 6: Excellent

1.Develops ideas well and uses many specific, relevant details throughout the essay.
2. Is well organized with clear transitions; maintains focus.
3. Sustains varied sentence structure.
4. Exhibits many specific word choices.
5. Contains little or no errors in grammar and conventions; errors do not interfere with understanding.
6. Clearly states the position and effectively persuades the reader of validity of argument.

Score of 5: Skillful

1. Develops ideas with some specific, relevant details.
2. Is clearly organized; information is presented in an orderly way, but essay may lack transitions.
3. Exhibits some variety in sentence structure.
4. Displays some specific word choices.
5. May contain some errors in grammar and conventions; errors do not interfere with understanding.
6. Clearly states the position and persuades the reader.

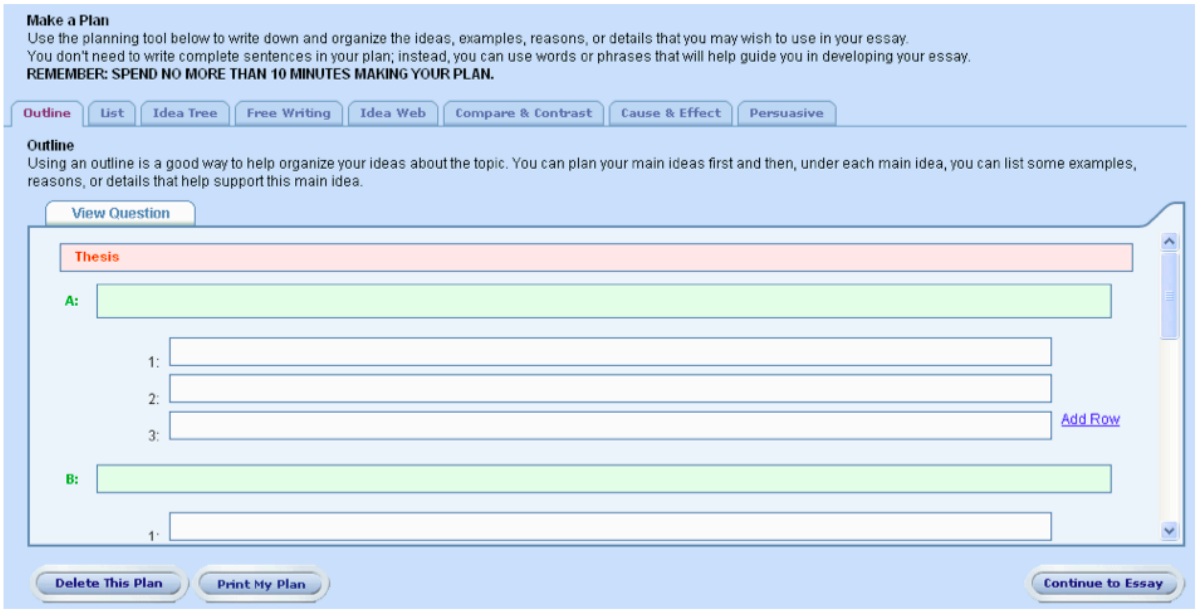
Score of 3: Uneven

1. Provides limited or incomplete information; may be list-like or have the quality of an outline.
2. Is disorganized or provides a disjointed sequence of information.
3. Exhibits uneven control over sentence structure.
4. May have some inaccurate word choices.
5. Contains errors in grammar and conventions that sometimes interfere with understanding.
6. While a position is stated, either it is unclear or undeveloped.

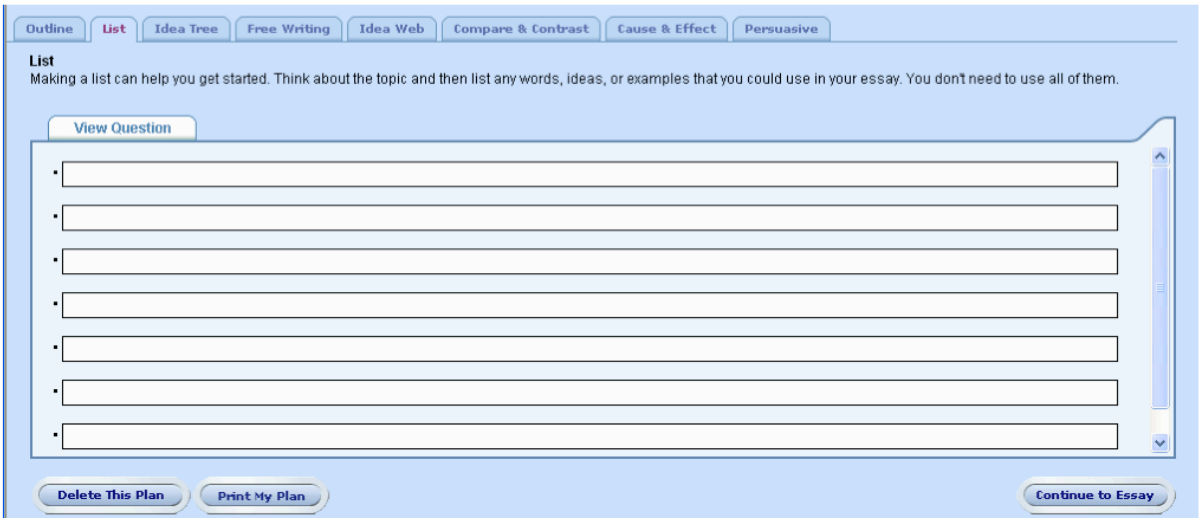
Score of 1: Unsatisfactory

1. Attempts a response, but may only paraphrase the prompt or be extremely brief.
2. Exhibits no control over organization.
3. Exhibits no control over sentence structure.
4. Contains inaccurate word choices throughout most of the essay.
5. Is characterized by misspellings, missing words, and incorrect word order; errors in grammar and conventions severely impede understanding throughout the essay.
6. Little effort is made to persuade, either because there is no position taken or no support is given.

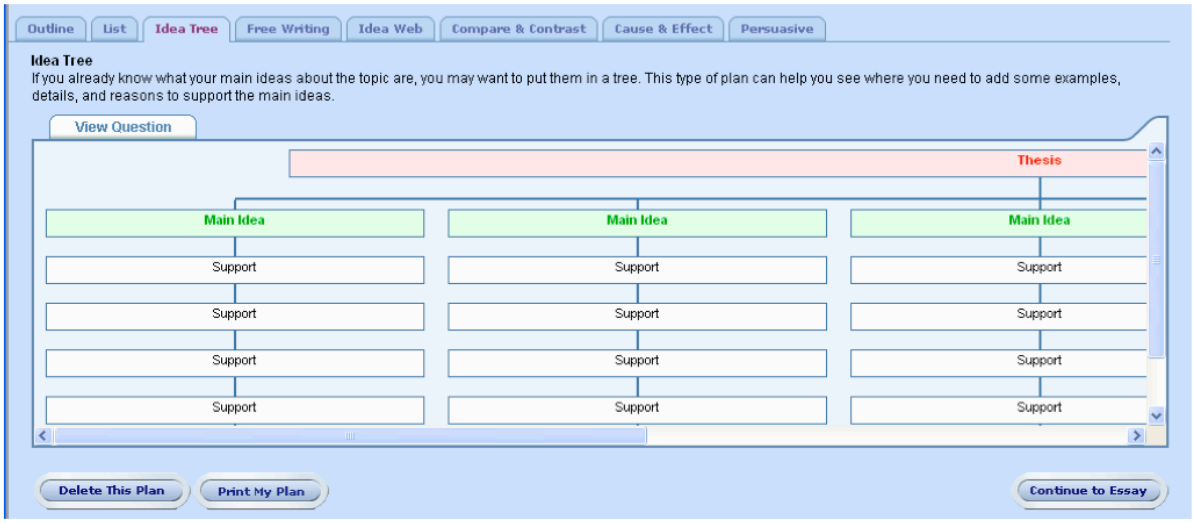
7.2. Appendix 2: Outline



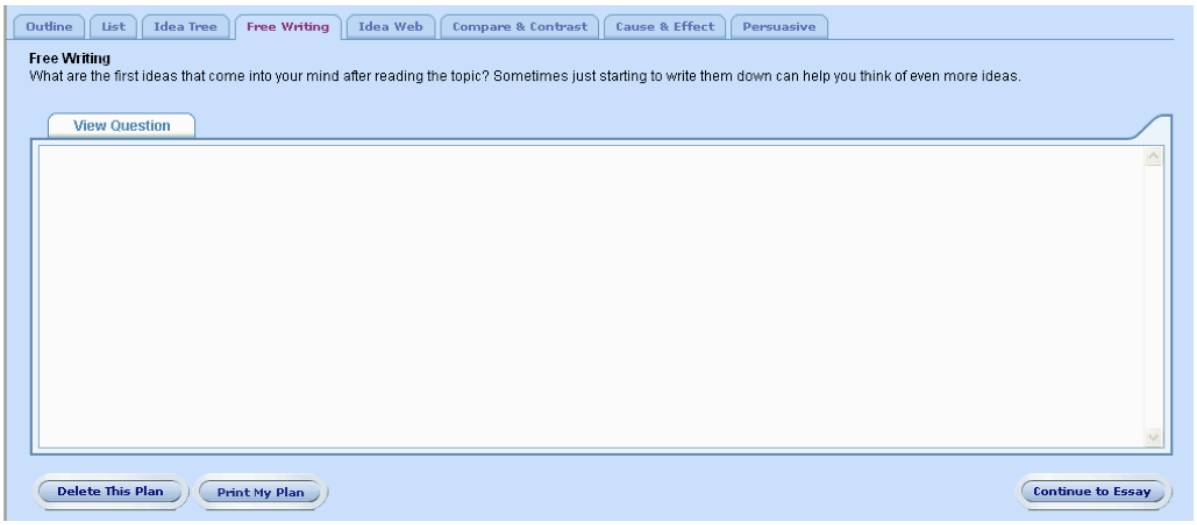
7.3. Appendix 3: List



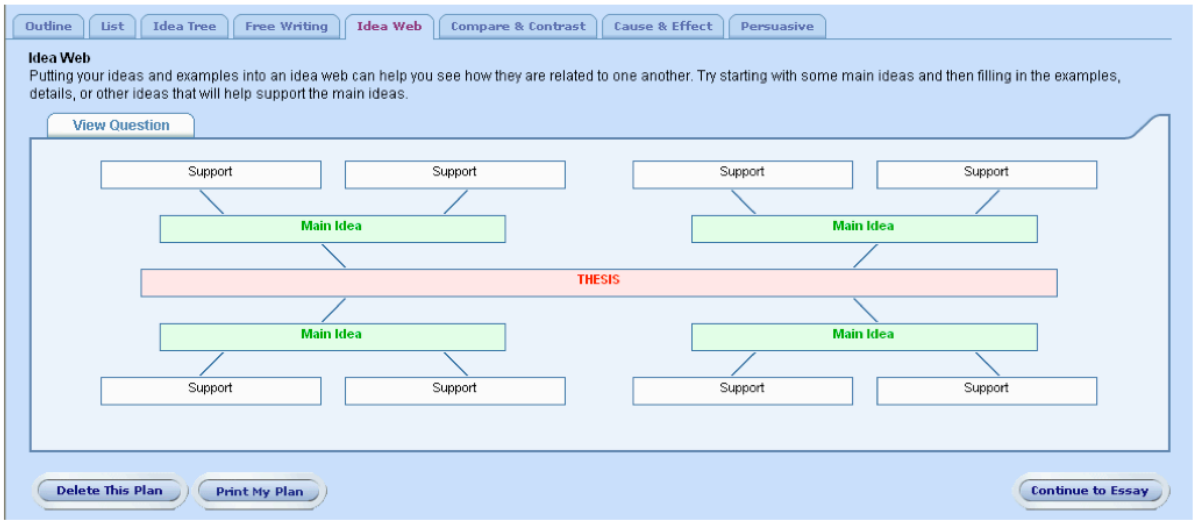
7.4. Appendix 4: Idea Tree



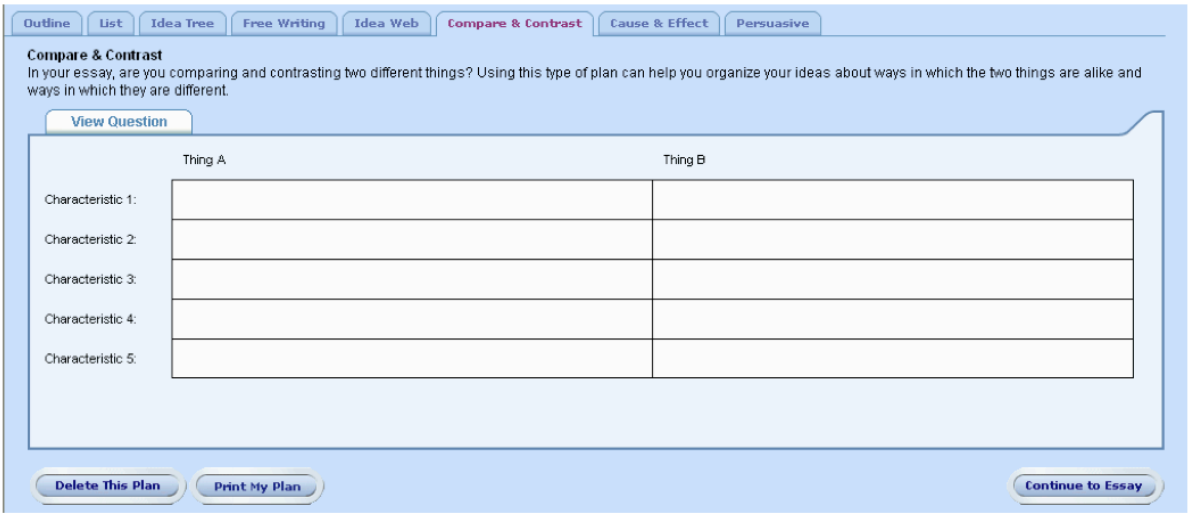
7.5. Appendix 5: Free Writing



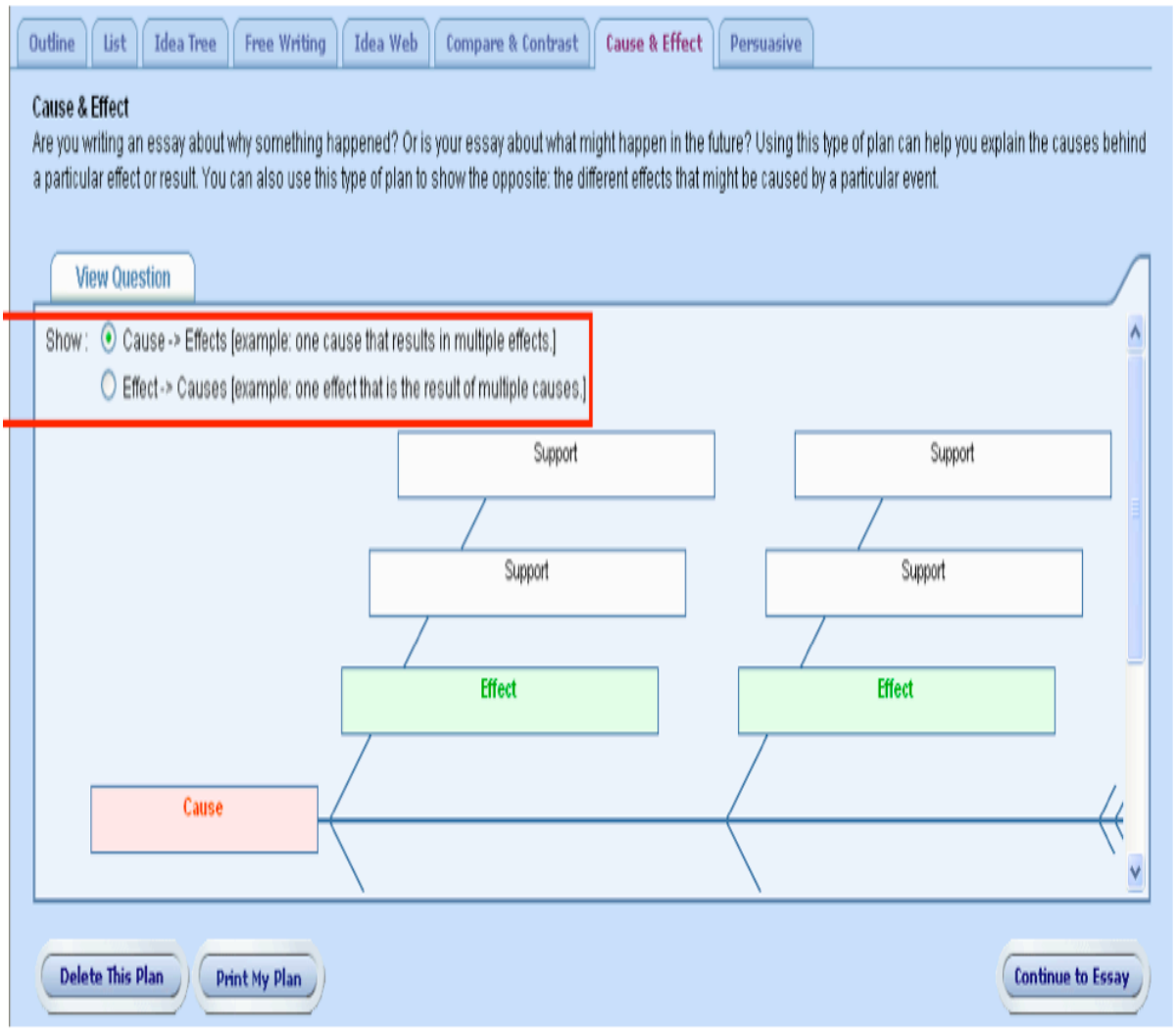
7.6. Appendix 6: Idea Web



7.7. Appendix 7: Compare & Contrast



7.8. Appendix 8: Cause



7.9. Appendix 9: Effect