1-1-2008

The effects of automated essay scoring as a high school classroom intervention

Kathie L Frost
University of Nevada, Las Vegas

THE EFFECTS OF AUTOMATED ESSAY SCORING AS A HIGH SCHOOL CLASSROOM INTERVENTION

by

Kathie L. Frost

Bachelor of Science
University of Arizona
Tucson, Arizona

Master of Business Administration
University of Nevada, Las Vegas

A dissertation submitted in partial fulfillment
of the requirements for the

Doctor of Philosophy Degree in Curriculum and Instruction
Department of Curriculum and Instruction
College of Education

Graduate College
University of Nevada, Las Vegas
December 2008

INFORMATION TO USERS

The quality of this reproduction is dependent upon the quality of the copy submitted. Broken or indistinct print, colored or poor quality illustrations and photographs, print bleed-through, substandard margins, and improper alignment can adversely affect reproduction.

In the unlikely event that the author did not send a complete manuscript and there are missing pages, these will be noted. Also, if unauthorized copyright material had to be removed, a note will indicate the deletion.

ProQuest LLC
789 E. Eisenhower Parkway
PO Box 1346
Ann Arbor, MI 48106-1346

UNLV
UNIVERSITY OF NEVADA LAS VEGAS

Dissertation Approval

The Graduate College
University of Nevada, Las Vegas

November 17, 2008

The Dissertation prepared by

Kathie L. Frost

Entitled

The Effects of Automated Essay Scoring as a High School Classroom Intervention

is approved in partial fulfillment of the requirements for the degree of

Doctor of Philosophy in Curriculum and Instruction

Examination Committee Chair

Dean of the Graduate College

Examination Committee Member

Examination Committee Member

Examination Committee Member

Graduate College Faculty Representative

ABSTRACT

The Effects of Automated Essay Scoring as a High School Classroom Intervention

by

Kathie L. Frost

Dr. Randall Boone, Examination Committee Chair
Professor of Education
University of Nevada Las Vegas

This quasi-experimental, mixed methods study investigated whether students writing development and proficiency, in combination with teacher-led instruction, are significantly affected by the use of an automated essay scoring (AES) system. The ninth grade standard and honors English students were divided into control and treatment groups at a large, urban high school. Student writing was examined for any changes in proficiency, measured by human- and AES-scored holistic measures. A developmental writing index was used to analyze the rate of change in pre- and post-essays. The AES system was further researched by comparing the treatment and control groups' trait score categories. Finally, treatment students were interviewed and surveyed to identify their degree of satisfaction with the AES system.

Automated essay scoring systems have moved from their original purpose of rapidly and reliably scoring high stakes testing into the classroom as an instructional tool providing holistic and trait scoring. One area of potential AES usefulness is to provide students with more writing opportunities that include feedback. While supporting research findings that student writing improves if more writing opportunities with feedback are provided, this also supports the iterative process of writing and revision.

To support teachers' optimum classroom technology integration of an AES system to supplement teacher-led instruction, an access ratio of one Internet-connected computer for each student, (i.e., 1:1) needs to be provided. System-provided or teacher-provided writing prompts (i.e., topics) can be selected to provide students with AES simulations of the summative score of high stakes testing, in concert with formative trait scoring, which gives specific recommendations to improve writing.

No gender difference was shown for the treatment participants from the AES-scored measures. The human-scored writing proficiency and development measures were inconclusive for gender and class levels due to the small sample size. By class levels, treatment honors students performed significantly better on the AES-scored proficiency measure, but the results were not supported by the human-scored measure. The other AES-scored measures analyzed by class levels, the development and trait category measures, did not show significance. However, the treatment participants expressed a high degree of satisfaction with the use of the AES system.

TABLE OF CONTENTS

APPROVAL PAGE ii

ABSTRACT iii

LIST OF TABLES viii

CHAPTER 1 INTRODUCTION 1
Purpose of the Study 1
Background 3
Holistic Scoring 4
Computer Processing Tools 7
Automated Essay Scoring's Beginnings 8
Program Essay Grade's Beginnings 8
Writers Workbench 10
Program Essay Grade in the 1990s 11
Current Automates Essay Scoring Systems 12
The Current Research 13

CHAPTER 2 REVIEW OF RELATED LITERATURE 15
Technology in the Classroom 17
Internet Access 17
Classroom Computer Use 20
Feedback 27
Computer Assisted Instruction 27
Teacher Feedback 29
Automated Essay Scoring Feedback 39
Summary 50

CHAPTER 3 METHODOLOGY 52
Research Design 53
Quasi-experimental 53
Potential Threats to Validity 55
Educational Scientific Research 56
Participants 57
Setting 57
Teachers 58
Students 59
Protocol 59
Teacher Interviews 59

Writing Prompts 59
Developmental Index 60
Treatment Group 61
Automated Essay Scoring System 61
Student Interviews 62
Control Group 63
Data Collection 63
National Writing Project Holistic Score Collection 64
Treatment Group Data Collection 65
Automated Essay Scoring System Data Collection 65
Student Survey and Interview Data Collection 65
Control Group Automated Essay Scoring System Data Collection 65
Teacher Semi-structured Interview Data Collection 66
Developmental Index Data Collection 66
Data Analysis 67
Automated Essay Scoring System Data Analysis 67
National Writing Project Holistic Scoring 68
Other Scoring 69
Conclusion 69

CHAPTER 4 RESULTS 71
Participants 74
Test Setting 75
Question One 76
Holistic Scores and Class Levels 76
Holistic Scores for Gender and Class Levels 78
Question Two 80
Words per T-unit and Class Levels 80
Words per T-unit for Gender and Class Levels 81
Question Three 83
Grammar Errors 84
Usage Errors 86
Mechanics Errors 88
Style Errors 89
Organization and Development 91
Question Four 93
Participant Experiences and Perceptions 93
Preferences for the Automated Essay Scoring System 95
Usability, Improvement from, and Effectiveness of the Automated Essay Scoring System 100
Frequency of Automated Essay Scoring System's Use 101
Summary 102

CHAPTER 5 DISCUSSION OF FINDINGS 106
Discussion of Research 106
Question One 107

Question Two 111
Question Three 112
The class level 113
Question Four 114
Participant Descriptions and Perceptions 114
School computer training and home computer access 115
Modal writing preference and self-perceived writing quality 115
The language spoken at home 116
Summary 116
Preferences for the Automated Essay Scoring System 117
Computer writing feedback and automated essay scoring system's helpfulness and holistic score 117
Teacher's importance 118
Importance of writing more 118
Preferences regarding the best thing about the automated essay scoring system 119
Preferences regarding the worst thing about the automated essay scoring system 120
Summary 122
The Automated Essay Scoring System's Usability, Improvement, and Effectiveness 123
The Automated Essay Scoring System's Frequency of Use 123
Limitations of the Study 125
Implications and Future Research 126

APPENDICES 129
A CRITERION'S SCORING CATEGORIES AND SUBCATEGORIES 129
B DEFINITIONS 132
C SEMI-STRUCTURED TEACHER INTERVIEWS 134
D TREATMENT STUDENTS' INTERVIEWS 135
E T-UNIT AND CLAUSE SCORING GUIDELINES 138
F PERMISSIONS 141

VITA 166

LIST OF TABLES

Table 1 Ethnicity of Class Levels from Which Participants Joined Study 58
Table 2 Data Measurement and Collection Timing for Each Research Question 64
Table 3 Analysis of Variance Comparisons of Outcomes 68
Table 4 Participants by Gender and Class Levels 74
Table 5 Automated Essay Scoring's Holistic Score by Class Levels 77
Table 6 National Writing Project's Holistic Score by Class Levels 78
Table 7 Automated Essay Scoring's Holistic Score by Gender and Class Levels 79
Table 8 National Writing Project's Holistic Score by Gender and Class Levels 80
Table 9 Words per T-unit by Class Levels 81
Table 10 Words per T-unit by Gender and Class Levels 83
Table 11 Automated Essay Scoring Grammar Errors by Class Levels 85
Table 12 Automated Essay Scoring's Errors by Gender and Class Levels 85
Table 13 Automated Essay Scoring Usage Errors by Class Levels 87
Table 14 Automated Essay Scoring's Usage Errors by Gender and Class Levels 87
Table 15 Automated Essay Scoring's Mechanics Errors by Class Levels 88
Table 16 Automated Essay Scoring's Mechanics Errors by Gender and Class Levels. 89
Table 17 Automated Essay Scoring's Style Errors by Class Levels 90
Table 18 Automated Essay Scoring's Style Errors by Gender and Class Levels 90
Table 19 Automated Essay Scoring's Development and Organization by Class Levels 92
Table 20 Automated Essay Scoring's Development and Organization by Gender and Class Levels 92
Table 21 Treatment Participant's Home Computer Access, Modal Writing Preference, and Perceived Writing Ability by Class Levels and Gender 94
Table 22 Treatment Participant's Percentage of Languages Spoken at Home by Class Levels and Gender 95
Table 23 Treatment Participants' Preferences for Computer Feedback, Most Help to Writing and Holistic Score Perception by Class Levels and Gender 97
Table 24 Treatment Participants Regarding the Best and Worst Things about the Automated Essay Scoring by Class Levels and Gender 99
Table 25 Treatment Participants' Usability, Writing Improvement, and Effectiveness of the Automated Essay Scoring System by Class Levels and Gender 100
Table 26 Treatment Participants' Multiple Automated Essay Scoring Submissions by Class Levels and Gender 101
Table 27 Treatment Participant's Test Submissions by Class Levels and Gender 102
Table 28 Pre-test Measurements with Significance for Class Levels or Gender and Class Levels 127

ACKNOWLEDGEMENTS

I would like to thank my dissertation chair, Dr. Randall Boone, for his advice and patience. I also want to thank the other committee members for their individualized assistance: Dr. Kent Crippen, Dr. Marilyn McKinney, Dr. Greg Levitt, and Dr. Kyle Higgins. I want to acknowledge my husband, Walt Frost, for his unending support and encouragement.

CHAPTER 1

INTRODUCTION

Purpose of the Study

Large scale testing, such as the Graduate Management Admissions Test (GMAT), created the need to objectively score large numbers of essays in a short time, thus giving rise to automated essay scoring (AES) systems. Automated essay scoring systems provide computer-based evaluation of written work (Wolcott & Legg, 1998). The expansion of high stakes testing has resulted in external, direct assessment being used to place students in certain classes or even to determine their graduation from high school. Such assessments have also challenged classroom teachers to provide more classroom assessment opportunities for students, since increased writing with feedback is known to increase the quality of student writing (Pritchard & Honeycutt, 2006). The recognition that teachers have a limited amount of grading time provided the impetus for moving AES applications beyond high stakes writing assessment and into the classroom (MacArthur, 2006). Though AES systems are used in the classroom, there is little research in that area.

This research sought to provide insights into the classroom component of the instructional use of AES systems. The focus was on the AES, *Criterion* (Educational Testing Service, 2007a), which is the application used in the school district where this study took place. Question one investigated if there is a significant difference in the

writing proficiency improvement of students who use an AES system in combination with teacher-led writing instruction compared to students who receive only teacher-led writing instruction, with assessment based on holistic scores from human raters and an AES system. Question two sought to identify if there is a significant difference in the writing development of students who use an AES system combined with teacher-led instruction compared to students who receive only teacher-led instruction, as measured by words per t-units (W/T). Question three examined if there is a significant difference between pre- and post-test AES trait error feedback categories (e.g., grammar, spelling) for those students who use an AES system combined with teacher-led instruction when compared to those students who had only teacher-led instruction. Question four explored the degree of user satisfaction for the students who used the AES system as measured by a survey and semi-structured interviews. In addition, gender was investigated for being a significant factor in the outcome of each question.

The history of writing assessment itself had an impact on the development of AES systems. During the period of 1950 to 1970, large scale writing assessments were mainly administered as objective tests (Yancy, 1999). It was during this time that AES research and development began, foreshadowing the next era of scoring (Page, 2003). From 1970 – 1986, holistically-scored essays were used for large scale writing assessment, creating an actual need for a large number of student essays to be scored in a timely manner, with validity and reliability (Yancy, 1999). It was not until the late 1990s that AES systems became commercially available (Shermis & Burstein, 2003).

Among the difficulties with tracing the historical development of AES systems are the changing ownership of the systems and the mixing and matching of various

component programs within the systems. In addition, AES applications are under continuous development, which includes adding new programs to address assessment shortcomings. Therefore, information about an AES system may be only as accurate as the date of the research's publication. Some information is limited simply due to the fact that the AES applications are proprietary commercial ventures. For example, according to Rudner and Gagne (2001), the actual variables and their statistical weights for calculating scores of Program Essay Grade (PEG) (Page & Paulus, 1968; Page, Poggio, & Keith, 1997) are unknown.

Three widely-used AES systems are backed by the development, marketing, and support staffs of large, world-wide companies (Herrington & Moran, 2001; Kelly, 2001; Kukich, 2000; Valenti, Neri, & Cucchiarelli, 2003): (a) *Criterion* Online Writing Evaluation from Educational Testing Service (ETS), (b) *WriteToLearn* from Pearson Knowledge Technologies (PKT), and (c) *MY Access!* from Vantage Learning. All of these AES systems are Web-based, so they do not require any special computer program installations for use (Kelly, 2001). This historical review provides background information and an overview of AES beginnings, leading to the focus on AES used as a classroom intervention.

Background

Holistic scoring is an accepted evaluation methodology for large-scale writing assessments (Wolcott & Legg, 1998). Human, holistic scoring is either used alone or in combination with AES systems (Williams, 2001). Technological developments in computer processing tools have moved AES systems closer to their simulation goal of

modeling human scoring of essays (Attali, 2004; Liddy, 2001; Shermis & Burstein, 2003; Shermis et al., 2006).

Holistic Scoring

Automated essay scoring systems seek to replicate the validity (i.e., measuring what is intended) and interrater reliability (i.e., consistency of agreement among readers) of holistic assessment that is achieved by humans who have been trained with scoring procedures. Wolcott and Legg (1998), in a non-vendor publication, provided the following industry-standard example of ETS's holistic scoring method by humans for a high-stakes test.

Holistic scoring describes what is found; it does not provide any remedies of how to fix any anomalies that are discovered. Thus, one of the drawbacks to holistic scoring is that students do not receive feedback on specific writing traits to enable revision. According to Wolcott and Legg (1998) the "theoretical basis" of holistic scoring is that it "encompasses all aspects of writing in its evaluation" (p. 81). This type of scoring is based on the impression of the whole paper. Scoring the whole paper means there is a balance in assessment of the rhetorical, grammatical, and mechanical parts of the paper. The sum of holistic essay scoring is said to be greater than individual scoring of the parts. This scoring uses a relative set of criteria, ranked according to other papers, thus "employing a norm-referenced approach" (Shermis & Daniels, 2003, p. 173). Holistic rubrics, set after rating numerous writing samples, describe scoring ranges. Range setting refers to identification of the papers that represent each point on the holistic scale (e.g., 6, 5, 4, 3, 2, and 1 on a 6-point scale).

For holistic scoring, two readers usually score each paper, not knowing the other score or identity of the alternate reader. Writing is evaluated against specified criteria and ranked against other papers in the same assessment. The best criteria for holistic scoring are clearly stated and easily accessible. According to ETS practices, readers are extensively trained to thoroughly understand the criteria and conduct the assessment by repeated individual ranking of papers and their public comparison of results (i.e., calibration). A description of the calibration and scoring process follows.

1. The preliminary set of range setting scoring of papers is done by the chief reader and a small group of the experienced scorers.
2. Each day begins with a table leaders meeting to simultaneously score and rank the all the range setting papers from the previous day's scoring. They publicly discuss their results and then score and rank the new range setters, using all the possible scores.
3. Then it is back to the tables, where the chief reader starts the reading session by reviewing the procedures, including ignoring unimportant qualities like hand writing and length.
4. The table leaders, experienced scorers, oversee five readers (i.e., scorers) at each table. To begin the reading session, the readers duplicate the processes of the table readers. They read and score the range setters and then publicly discuss their results until they gain a consensus.
5. Then the real exam reading begins, with the table leaders always available to help.

6. After lunch, before re-starting their scoring, the readers simultaneously score another set of range setters to preserve their attention and uniformity.

Two procedures are used to specifically prevent final scoring errors. First, intermittent scoring was independently conducted by a chief reader, a table leader and a reader. Second, each table leader circled the table and re-scored what the reader had already scored. However, discrepancies still occur. Differences in holistic scores might be caused by the paper's writing (e.g., contradictions between the content and structure). Some of the errors are accidental factors due to humans scoring (e.g., the room was too warm or the reader's attention wandered). These instances might be caught by existing procedures, but sometimes the scoring differences were not noticed. Several formulas were used for solutions. At times, the re-read score replaced the discrepant score, other times all three scores were averaged; though each scoring situation can have different standards. There are many other issues relative to interrater reliability and instrument validity for holistic scoring by humans (Blok & de Glopper, 1992; Rudner, 1992; Wolcott & Legg, 1998).

The validity and reliability of AES systems' holistic scoring has been extensively tested by vendor and non-vendor research (Burstein, Kukich, Wolff, Lu, & Chodorow, 1998; Cizek et al., 2003; Keith, 2003; Shermis & Daniels, 2003; Yang, Buckendahl, Juszkiewiez, & Bhola, 2002). The tests have insured that the AES scores and human scores match or come within one point of each other, because the AES systems have been used for scoring high stakes tests (e.g., Graduate Management Admissions Test [GMAT] and Advanced Placement Test [APT]). The process that AES systems use to arrive at a

holistic score, however, does not necessarily parallel the human scoring process (Attali, 2004).

Computer Processing Tools

Developments in artificial intelligence (AI) and natural language processing (NLP) have improved AES systems (Liddy, 2001). Artificial intelligence is the field of computer science that seeks to emulate human cognition, including machine learning, statistical inference, and adaptive computations (Dale & Douglas, 1997; Luger, 2001; Nilsson, 2005). Natural language refers to language spoken by humans (Coxhead, 2001). Natural language processing is a computer science division of AI, an attempt by computers to process human language.

Natural language processing theories began during World War II with the use of computers to break military code (Coxhead, 2001; Liddy, 2001; Reingold & Nightingale). Familiar interactions using NLP include information retrieval from Internet searches using Google and computer translation from one natural language to another (e.g., English to Spanish). Processing a stream of language must consider four basic levels: (a) "phonology -- speech sounds and how we make them," (b) "morphology -- the structure of words," (c) "syntax -- how the sequences are structured," and (d) "semantics -- meanings of the strings" (Batali, 2006). Natural language processing includes (a) algorithms (i.e., procedures for solving problems), (b) syntactic parsing (i.e., breaking into linguistic components) tools, (c) semantic analysis (i.e., representation of the meaning of linguistic components) techniques, and (d) pragmatics (i.e., actual meaning for the context) methods. Computational linguistics is a related field which can focus, for example, on syntax. Part of the growth in NLP has been the merging of previously

separate fields from electrical engineering, computer science, and linguistics: (a) speech recognition, (b) NLP, (c) computational linguistics, and (d) computational psycholinguistics (Liddy, 2001).

Dale and Douglas (1997) provided some easily understood examples of NLP "language sensitivity" or language understanding (e.g., punctuation, grammar, and syntax error) by computer applications (p. 123). For example, a period serves as the end of a sentence, which makes it a part of syntax, or indicates an abbreviation, which makes it a part of a word. Another example of linguistic sensitivity would be the identification of the misspellings between so and sew, according to the syntax of the writing. The third kind of linguistic sensitivity would process different kinds of text (e.g., the differentiation between references and citations in text). According to Dale and Douglas (1997), NLP at that time was very primitive in actual knowledge about language, rather it enabled the sophisticated processing of plain strings of text. The ongoing development of AES systems has been propelled by technological advances in (a) computer hardware, (b) the Internet, (c) AI, and (d) NLP (Attali, 2004; Liddy, 2001; Shermis & Burstein, 2003).

Automated Essay Scoring's Beginnings

Automated essay scoring tools are actually applications or systems because they are comprised of a group of computer programs. The beginnings of AES encompass PEG's beginnings, Writers Workbench, and PEG in the 1990's.

Program Essay Grade's Beginnings

The potential use of computer grading for high-stakes tests was recognized by the testing industry, resulting in the College Entrance Examination Board's (CEEB) funding of the preliminary work in essay analysis and the simulation of human essay scoring

(Page, 2003). Page (1966) attempted to convince educators that (a) there was a serious need for computers to grade essays, (2) it was feasible for computers to grade essays, and (c) instructional improvements would result from such grading (Kukich, 2000; Page & Paulus, 1968; Valenti et al., 2003; Williams, 2001). The PEG report (Page & Paulus, 1968) documented the work of Page and Paulus (1968) during their 2-year research contract with the United States Office of Education and included their preliminary work from 1965 (Kukich, 2000).

Program Essay Grade operated on main-frame computer systems, with the essays being entered via key-punched computer cards (Page & Paulus, 1968). At that time, there were no personal computers or word processing programs. The researcher's goal was to simulate human holistic raters' use of intrinsic variables, named *trins* (e.g., "aptness of word choice" or "fluency") (p. 15). Since there were no direct computer measurements for such variables, approximations or correlation variables, named *proxes*, were developed. For example, the "*trin* of fluency" would be measured by the "*prox* of actual word count" (p. 16). The most predictive *proxes* included (a)"average word length," (b) "essay length in words," (c) "number of commas," and (d) "number of prepositions" (p. 44). In order to score essays, the system first had to be calibrated (i.e., trained) by having a large number of already human-scored essays run through the system in order to set the statistical regression formula for that specific essay set. A calibration methodology is still in use with current AES systems (Burstein, Chodorow, & Leacock, 2004).

The early versions of PEG measured writing quality via surface (i.e., mechanics) features, instead of content, and correlated with human holistic scoring (Page, 1994; Page & Paulus, 1968). This correlation was as high as human raters correlated with each other

at that time. However, PEG was not well received by the education and writing communities because it used indirect measures of writing quality (Herrington & Moran, 2001; Kukich, 2000; Valenti et al., 2003). Program Essay Grade's development and the advocacy of computer essay grading continued, though more slowly, partly due to the logistical hurdles created by the required access to mainframe computers (Macdonald, Frase, Gingrich, & Keenan, 1982; Page, 2003; C. Smith & Kiefer, 1983).

Writers Workbench

The next significant application was Writer's Workbench (WWB), a series of programs developed on a main-frame UNIX computer by Bell Laboratories for use by their professional writing staff (Reid & Findlay, 1986; C. Smith & Kiefer, 1983). The WWB system was adapted for use at Colorado State University, and foreshadowed the use of AES systems in the classroom (C. Smith & Kiefer, 1983). It was not an AES system for use with high-stakes testing, but rather had the goal of helping university students improve their quality of writing. Its use was expanded in 1981-1982 to over 3000 students in classes such as basic writing, college composition, and advanced writing.

Research sought to determine WWB's stylistic measurements usefulness to university students for identifying writing quality and thus encouraging effective revision on drafts (Dale & Douglas, 1997). The study quantitatively compared WWB to human holistic scoring of essays. The human scorers were formally trained with procedures adapted from ETS's methodology for scoring the APT, thus insuring maximum reliability. The 44 placement essays were from the 1982 Colorado State University Composition Placement Examination given to every entering freshman. The selected

essays' three holistic scores from three raters did not vary more than one point. The samples were also representative of the whole, 1 to 9 point, holistic scale. The results of the study were limited to the writing topic used for the essays.

In WWB, simple statistical correlations were run between the holistic score and 27 style measurements (e.g., sentence length, readability, and spelling). Length showed the highest correlation, followed by spelling. The Kincaid readability score (Kincaid, Fishburne, Rogers, & Chissom, 1975) showed the third highest correlation. This readability formula combines and weighs the average sentence length and the average syllables per word. The fourth highest correlation was average word length. The scoring results were particular to the essay topic sample, so they were not generalizable. The revision factors found to directly affect the re-scoring of an essay were identified as those affecting writing fluency: (a) sentence length, (b) word length, and (c) readability. The computer algorithm used by WWB was very simple, not using parsers to break the text into grammatical parts nor using any form of NLP. Rather, WWB used statistical information and heuristics (a trial and error learning process). The style measurement criteria also indicate another issue with AES that still exists today. There is no single, recognized standard defining an ideal written essay that could be used to calibrate human scoring or to select master text (Valenti et al., 2003).

Program Essay Grade in the 1990s

With the advent of PCs in the 1980's, development began again on PEG (Page & Petersen, 1995; Page et al., 1997; Shermis, Koch, Page, Keith, & Harrington, 1999, 2002). Throughout the 1990's, the development and widespread testing of PEG continued to be based on *trins* and *proxes*, requiring a number of human-graded essays to set the

statistical coefficients for a grading set (Chung & O'Neil, 1997). So the human holistically-scored essays, as described in the background section, continued to be used to train or calibrate PEG. Additional parsers (enabling the separation of text into grammatical units) were added and trait ratings were included for (a) content, (b) organization, (c) style, (d) mechanics, and (e) creativity (Page, 1994; Valenti et al., 2003).

As PEG moved into the commercial realm, little has been documented on the actual *proxes* that are used in the rating calculations (Shermis, Mzumara, Olson, & Harrington, 2001). The *proxes* could change for each essay topic because the system needed to be re-calibrated for each set of essays. Program Essay Grade's correlation rate with human raters reached as high as .87 (Kukich, 2000). A Web browser interface was added, but the quality of the scoring continued to be based on statistical calculations with observable *proxes,* but without the use of any NLP.

Current Automates Essay Scoring Systems

The major commercial AES systems consist of a holistic scoring engine and a formative scoring engine, using a variety of NLP, statistical, and AI approaches. Holistic scoring engines became commercially available between 1998 and 2000 (Shermis & Burstein, 2003). These systems use norm-based scoring, with a set of human-scored essays or expert texts (which were scored by several people following the human, holistic scoring model) used to calibrate scoring formulas for system-specific variables. A second set of un-scored essays are then used to test the calibrations. As a result of contracts to score the GMAT, the holistic scoring engines of *Criterion* (i.e., *e-rater*) and *My Access!* (i.e., *IntelliMetric*) have been psychometrically evaluated as having high computer-to-human holistic scoring correlations (Rudner, Garcia, & Welch, 2006). *Intelligent Essay*

Assessor's (IEA) machine scoring engine (i.e., *Latent Semantic Analysis*) reports similar high results for correlations between system and human holistic scores (Kukich, 2000). Information is limited about the Web-based user interfaces and the accuracy or usefulness of diagnostic feedback of any of these AES systems (Burstein et al., 1998; Shermis & Burstein, 2003).

A modular system, *e-rater* uses "syntactic variety and discourse structure (like PEG) and content analysis (like IEA)" (Rudner & Gagne, 2001). The prototype of *e-rater* was tested in 1998, with the pilot version of the automated GMAT Analytic Writing Assessment (AWA) providing the test data. From 1999 until 2006, *e-rater* was used as one grader and a human as the other grader (i.e., instead of two human graders) for the GMAT (Williams, 2001). To create the holistic score, *e-rater* extracts linguistic features and develops a statistical model to correlate the features to writing quality (e.g., syntax or topical content) and to assign a ranking, which is the holistic score (Attali, 2004; Burstein, Chodorow, & Leacock, 2003; Burstein & Higgins, 2005). The use of *e-rater* for high stakes testing is supported by many studies showing the high correlation (i.e., 97% average) between the human holistic scores and the computer-based holistic scores, meaning the scores matched or were one point away from each other (Burstein et al., 1998).

The Current Research

Automated essay scoring systems' development has benefited from the development of technology such as AI and NLP (Shermis & Burstein, 2003). The increase in the use of high stakes testing has also increased the need for students to

practice their writing. Automated essay scoring systems provide a way for students to have more writing practice with potentially skilled diagnostic feedback.

Sufficient research has not yet been conducted to ascertain if AES use is really helpful to students. This research sought to examine if the use of an AES system significantly correlated to participants' improvements in writing proficiency or writing development rate of change. Another question sought to identify whether an AES system's use significantly correlated to changes in the categories or quantities of trait errors that students have in their post-essay products. Finally, students' perceptions about the usefulness of an AES system as a writing intervention were analyzed.

CHAPTER 2

REVIEW OF RELATED LITERATURE

Technology and Writing

High stakes testing (e.g., Nevada's state-mandated writing proficiency assessment for high school graduation) has increased the need for writing practice in the classroom (Nevada Department of Education, 2006-2007). According to educational research, in order to improve their writing, students must write more and receive feedback on their writing (Burstein, Chodorow et al., 2003; Nippold, Ward-Lonergan, & Fanning, 2005; Page, 2003; Pritchard & Honeycutt, 2006). The purpose of writing feedback is to guide the student to revision, which results in an improvement in the quality of writing.

Much of the current AES research has been conducted by institutions developing commercial testing or teaching materials (Warschauer & Ware, 2006). Some of the institutional research has appeared in peer reviewed journals and has been presented at national educational conferences (e.g., National Council on Measurement in Education [NCME] and American Educational Research Association [AERA]). This review of the literature synthesizes current knowledge related to students' use of AES in the classroom.

The improved writing outcomes expected from students were in the areas of writing proficiency and development. Writing proficiency, reflected in a holistic, overall evaluative perspective of an essay, is greater than the sum of the evaluations of any specific writing traits (Wolcott & Legg, 1998). Writing development, part of language development, is defined as characteristics of individual development located at some point along a continuum (Wolfe-Quintero, Inagaki, & Kim, 1998). There is limited research on student writing outcomes with the use of an AES system. Therefore, this literature review focuses on two lines of research: technology use in the classroom and various forms of feedback. Manual index searches were conducted in the *College Composition and Communication Journal*, *Computers and Composition*, the *English Journal*, and the *Journal of Technology and Teacher Education*. Other references were found in the discovered articles and the *Handbook of Writing Research* (Fitzgerald, Graham, & Fitzgerald, 2006). Studies were gathered using the descriptors *word processing* and *computer writing* in the electronic databases of EBESCO, ERIC First, Academic Search Elite, Education Full-Text, and ProQuest Dissertations. References were also gathered using the descriptors *computer and writing*, *computer and composition*, *writing assessment*, *writing feedback*, and *writing revision*. The research populations covered a wide age range of grade and age levels, from preschool through professional writer, though most were college students.

Automated essay scoring systems are based on the theoretical model of cognitive processing of the human brain, with no accommodation for social learning (Ware & Warschauer, 2006). The feedback research examines teachers and computer feedback. Though many teacher factors (e.g., professional development, teacher technology skills,

instruction, pedagogy, and curriculum integration) affecting technology use in the classroom may appear in the research, they were not investigated in this study. This study is an investigation into improvement in students' writing outcomes with the use of an AES system as a supplement to teachers' instruction.

Technology in the Classroom

The potential student outcomes from the use of an AES system were dependent on the access to Internet connectivity and the use of educational technology in the classroom. The current AES systems are Web-based, requiring an Internet connection for use. Students must create their essays with a text editor, either internal or external (i.e., word processors) to an AES system. To receive evaluations, students must deliberately submit an essay to the AES system that has been set-up by a teacher.

Internet Access

A variety of studies show that Internet access is improving in schools. Gender is no longer thought to be a computer access issue, according to the 2003 U. S. Census Survey (Day, Janus, & Davis, 2003). For K – 12 school children, 83.2% of males and 83.6% of females used computers at school. Internet usage at school was similarly balanced, with 42.2% males and 44.3% females using the Internet.

The National Center for Education Statistics (NCES) periodically administers a short Fast Response Survey System (FRSS) survey (Parsad & Jones, 2005) to public school teachers that includes items about their computer and Internet usage. The FRSS data for 2003 were collected from October of 2003 through February of 2004. Surveys were mailed to a selective and representative sample of 1,207 public schools in the 50 states and the District of Columbia, resulting in an un-weighted response rate of 91%. In

the fall of 2003, nearly 100% of public schools had Internet access, as compared to the 1994 results of 35%. There were no discernible differences in Internet access by school characteristics, as grade level, school size, and percentage of minority enrollment. The comparisons were tested for statistical significance and documenting data were available.

The Internet access for instructional rooms (e.g., classrooms, computer labs, and libraries) grew from 3% in 1994 to 93% in 2003. Across school characteristics, access measured from 90% to 97%. However, that means that 3% to 10% of instructional rooms still did not have Internet access in 2003. According to the data provided, those Internet-deficit classrooms were most likely to be found at urban schools or schools that had 75% or more of students eligible for free or reduced-price lunches (i.e., indicating a high number of students at the poverty level). The ratio of students to Internet-connected computers averaged 4.4:1, but the actual level of connectivity related to school characteristics. Again, schools having 75% or more of the student population eligible for free or reduced-price lunches had a 5.1:1 ratio, while a lower 4.3:1 ratio was found at schools that had the fewest numbers of students at the lowest of poverty levels.

An unpublished study by Boone and Frost (2005) revealed that difficulties still exist in finding dependable classroom computer access. A Delphi study was conducted at a large urban school district with English teachers whose classes were AES users at five middle schools and five high schools. The purpose of the study was to develop a consensus on the best instructional strategies for using an AES and to identify which AES features were most beneficial to the students and/or teachers. Phase 1 asked participants two questions: List five specific teaching strategies that were useful when students used the AES system, and list five software features beneficial to students and/or teachers. The

responses were aggregated and sent out as Phase 2 for teachers to (a) rate the items on a five-point Likert scale, (b) select the three most important, and (c) briefly explain why those three were most important (Boone & Frost, 2005; Likert, Roslow, & Murphy, 1934). The sample population of 65 teachers had a 30% response rate for Phase 1 and 35% for Phase 2. The domain analyses of the Phase 1 responses revealed a third category: Technical difficulties with software, access, and hardware. The teachers noted in Phase 2 that the students had difficulties with logging into the AES software and the AES system losing student work. The teachers also had difficulties getting class time in computer labs. When time was found in the computer labs, it was difficult to find enough computers in working condition.

Originally, computer labs were thought to be the solution to making computers available to all students in a classroom at one time (McCanne, 2004; Stuebing, Celsi, & Consineau, 1998). More recently though, Hokanson and Hooper (2004) defined computer labs as "ghettoized," keeping computers separated from classroom learning (p. 249). To resolve the access issue, schools shifted toward groups of computers in individual classrooms (McCanne, 2004). The availability of only one computer in a classroom has been deemed as tokenism, that is not really making technology available for integration into the classroom curriculum (Hokanson & Hooper, 2004).

According to the 1999 FRSS report (Smerdon et al., 2000), computer use by teachers' and students' was related to the number of Internet-connected computers available in the classroom, not how many were available throughout the school. The 2003 FRSS report (Parsad & Jones, 2005) showed schools beginning to make laptops available to students. The optimum student to computer ratio (i.e., 1:1) for the use of an AES

system in the classroom can be provide by two methods: Computer carts (i.e., providing temporary 1:1 access) or a laptop initiative (i.e., providing continuous 1:1 access) that supplies computers to individual students (Grant, Wang, & Potter, 2005; Russell, Bebell, & Higgins, 2004). Access is a primary research consideration, since access is the issue that makes it difficult to conduct truly experimental studies about educational technology. Quasi-experimental studies are conducted because participants' computer access must be guaranteed first, rather than randomly selecting participants from a general pool of students.

Classroom Computer Use

For decades research has been underway on whether the classroom use of any kind of educational technology impacts student outcomes. Extensive research was conducted on the use of word processing in the classroom. A meta-analysis has investigated the research on student outcomes from the classroom use of categories (i.e., reading or math) of educational technology over a span of 7-years. More recently, a broad study was conducted about the classroom use of educational technology affecting student outcomes.

The research on word processing is of interest to this study since participants used a text editor to enter their essays into an AES system. A meta-analysis of word processing studies investigated document length and quality outcomes (Bangert-Drowns, 1993). Three criteria were used to select studies from the years 1984 to 1990: (a) the difference in the research methodology was only the modality of writing, one group used word processing and the other used hand writing, (b) the studies were retrievable from universities' and college libraries, and (c) treatment outcomes were quantitatively

measured. Essay length was measured by word count, and writing quality was measured by a holistic score.

Twenty-one characteristics of each study were coded for comparison. Those characteristics were divided into four categories: (a) eight variables were instructional treatment, (b) six were methodological features, (c) five were setting features, and (d) two were publication features. The participant descriptions were only the school grade and the researchers' unexplained determination of writing ability (e.g., low, average, or high). Computers were defined as either mainframe terminals or microcomputers, with all studies taking place in computer labs, except for one study taking place in a classroom. The functionality of the word processing applications was not provided, but must be considered primitive compared to today's word processors (Russell, 1999).

Out of 32 studies in the classic Bangert-Downs (1993) meta-analysis, only 4 studies showed positive correlations between length and holistic writing quality. Thus, the increased length did not necessarily mean increased quality. Of the 28 studies measuring writing quality, 66% reported an improvement with the use of word processing. Of those studies, 20 had enough information to calculate an effect size, identifying a significant, though small, .21 standard deviations (SD). Student computer skills were not considered in this meta-analysis. The Bangert-Downs (1993) research has been cited for the finding that writing on a computer increases writing quality (MacArthur, 2006; Russell, 1999). Overall, word processing has the reputation of being beneficial to "struggling" writers or those with learning disabilities (MacArthur, 2006, p. 253).

It should be noted here that a report commissioned by the Carnegie Corporation of New York, *Writing Next: Effective Strategies to Improve Writing of Adolescents in Middle and High Schools* (Graham & Perin, 2006), recommended that writing be taught with word processors. It was one of eleven key elements of writing instruction that were identified through a meta-analysis. The research encompassed the areas of writing-to-learn and learning-to-write. Only studies that reliably measured quality were included in the learning-to-write category, and this was the category where word processing research was located.

Following the practice of earlier meta-analyses on writing, studies had to be experimental or quasi-experimental. In addition to including 7 studies from grades K to 12 in the Bangert-Drowns (1993) meta-analysis, 11 other studies about word processing were collected. An effect size of 0.20 is small, 0.50 is medium, and 0.80 is large (Lipsey & Wilson, 2000). Word processing was identified as a process supporting student writing that had a medium, positive effect size (Graham & Perin, 2006). The effect size for low-achievers was even higher than a medium effect size. Therefore, the researchers suggested, not only did word processing have a significant positive effect on student writing quality; it seemed more effective in increasing the writing quality of lower-achieving writers.

In meta-analysis on student outcomes with the use of educational technology, Waxman, Lin and Michko (2003) selected 42 studies, including about 7,000 students. Selected studies met the following criteria: (a) Focus on teaching and learning with technology in K to 12, (b) classes had face-to-face meetings over 50% of the time, (c) quantitative, experimental, and quasi-experimental research that had been published in

refereed journals; (d) compared a technology group to a non-technology group or compared a group based on a pre- and post-test, and (e) included enough statistical data to create effect sizes.

Student outcomes from teaching and learning with technology were compared to student outcomes from traditional instruction. Separate results were provided for cognitive, affective, and behavioral outcomes. Cognitive outcomes were from researcher-based test, authentic assessment, and standardized tests. No test descriptions were provided. The 29 study-weighted comparisons of effect sizes for cognitive outcomes were small, positive, and significant. Affective (e.g., attitude) outcomes were positive and non-significant, while behavioral (e.g., time on task) outcomes were negative and non-significant. Overall, across all outcomes, technology showed a positive mean effect size that was small but significant. An ANOVA showed that the generalizations hold true across all the different research studies.

The researchers noted that though the overall student outcomes were significantly positive, an important limitation of the study was that meta-analysis findings were constrained by the quality of the primary study data. Out of a possible 200 teaching and learning with technology research studies, only 47 had enough statistical data to calculate effect sizes. Only 25% of the selected studies used randomized, experimental design. Waxman et al. (2003) also pointed out that many studies lacked the details for the 57 variables that were coded. About 25% of the studies lacked the details of the software being used. The selected research was published in the five years (i.e., 1997 – 2003) prior to the Waxman (2003) meta-analysis, meaning that the studies used hardware and software that is now over a decade old. While these same shortcomings were noted in

many of the studies presented in this literature synthesis, their purpose is to guide future research.

An Institute of Education Sciences report (Dynarski et al., 2007) to Congress was the result of research on student outcomes with the use of 16 different educational technology products. The measurements used were "student test scores, classroom activities, and roles of students and teachers" (p. xiv). There were four different groups (i.e., first grade, fourth grade, sixth grade, and mostly ninth grade) of participants, and all teachers in each group were randomly selected to be control or treatment groups. The groups are described further with the test results.

The software selection for the study was based on the product information that vendors voluntarily provided to the research committee. The research committee selected products that had, at the minimum, some research that indicated a positive effect from their use. Sixteen products were chosen out of the 160 that were submitted. The vendors helped in the selection of schools, which had higher minority populations and lower socioeconomic status (i.e., the target population) than average. The schools were also selected on the basis of not using software similar to what was being tested, in order to guarantee a difference between the treatment and control groups. The schools chose which of the selected software products would most likely fit their needs. The vendors provided software training to all the participant teachers.

The researchers (a) administered tests toward the beginning and end of the school year, (b) conducted three classroom observations, (c) collected data from teacher questionnaires, (d) assembled student records, and (e) gathered product records on both treatment and control groups. The outcome analyses were based on student test scores,

classroom activities, and roles of teachers and students. Three implementation findings focused on the classroom use of educational technology. First, the research found that across all four groups, the teachers believed the training prepared them to use the technology products, but their confidence decreased somewhat with the use of the software. Second, the technical difficulties were minor, meaning that they were easily resolved and most teachers indicated they would use the products again. Third, the use of educational technology was found to impact the treatment classroom behavior of both the students and teachers. The students were more likely to be working on their own, and the teachers were more likely to facilitate than to lecture.

The effectiveness of the educational technology was determined by analyzing the pre- and post-test scores of the treatment and control groups and correlating the results to school and classroom characteristics that were tracked.

1. The first grade group encompassed 13 districts, 42 schools, 158 teachers, and 2,619 students. This treatment group used five reading software products. The differences in the reading test scores from the treatment and control groups were not statistically significant. However, the large differences between schools' reading software test scores did correlate with the student-teacher ratio.
2. The fourth grade group included 11 districts, 43 schools, 118 teachers, and 2,265 students. This treatment group used one of four reading software products. Again, the differences between the treatment and control groups' reading test scores were not statistically significant. Differences in effect sizes did correlate with the amount of product use, but this was not a causal finding.

3. The sixth grade group involved 10 districts, 28 schools, 81 teachers, and 3,136 students. This treatment group used one of three math and pre-algebra software products. The differences in the math test scores between the treatment and control groups were not statistically significant. The differences between schools' test scores were not affected by any of the school or classroom characteristics measured by the study.
4. The final group, mostly ninth graders, contained 23 districts, 10 schools, 69 teachers, and 1,404 students. This treatment group used one of the three algebra software products. The math test scores' differences between the treatment and control groups were not statistically significant. The differences between schools' test scores were not affected by any of the school or classroom characteristics measured by the study.

The overall finding was that the test scores of the randomly assigned treatment groups, using a variety of reading and math educational software, were not significantly different from the control groups' test scores. This first report only evaluated software product categories (i.e., reading or mathematics) to determine the effectiveness of educational technology, while the follow-up report looked at the individual products.

Two of the meta-analyses reviewed here, Bangert-Drowns (1993) and Graham and Perin (2006) found a significant positive effect on writing quality from the use of word processing. The Bangert-Downs (1993) meta-analysis included 32 studies, but only 20, with 1,328 participants ranging from elementary school through college, had enough information to create an effect size. The Graham and Perin (2006) meta-analysis included 18 studies on word processing, but 7 of them were from the Bangert-Drowns (1993)

study. The Carnegie report (Graham & Perin, 2006) did recommend that writing be taught with word processors.

The meta-analysis by Waxman et al. (2003) analyzed 42 studies with 7,000 participants and found the research to show that the use of educational technology in the classroom had a small, significant positive effect on student outcomes. However, the most resent research by Dynarski et al. (2007), with 9,424 participants, indicated that the positive effects from the use of educational technology are not statistically significant. One of the groups in the Dynarski et al. (2007) study did indicate that frequency of classroom use of educational technology can affect outcomes, and the frequency is impacted by the ease of classroom computer access for teachers.

Feedback

Contemporary recursive processes of writing are identified as planning, drafting, and revision (Graham & Perin, 2006; Pritchard & Honeycutt, 2006). In order for revision to occur, students must receive feedback on their writing. There are three important research strands regarding feedback's effects on students' writing. First, computer assisted instruction (CAI) research has studied student outcomes and use strategies. Second, the types and possible effects of teachers' writing feedback have been the subject of several studies, though more are needed (Graham & Perin, 2006). Finally, the functionality and feedback of the AES system were reviewed.

Computer Assisted Instruction

One of the best features of CAI is that it provides immediate feedback, as do AES systems (Educational Testing Service, 2007a; Waxman et al., 2003). Christmann, Badgett, and Lucking (1997) conducted a meta-analysis on research that compared the

academic achievement of secondary students using CAI over a 12-year period, 1984 to 1995, across a variety of subjects areas. Computer-assisted instruction was defined as "programmed learning using microcomputers," while traditional instruction was "non-computer-based methods of instruction" (p. 283). The criteria for selecting studies specified a minimum of 20 secondary school students, but ranged from 28 to 425 students in the experimental and control groups, with a mean of 133 students. The research selected was correlational, quasi-experimental, or experimental, with academic achievement as the dependent variable and CAI as the intervention variable. From the total population of more than 1000 research studies, only 26, encompassing a total of 3,694 students, met the criteria.

The research question examined the academic achievement differences between students who only received traditional instruction and those who received both traditional and CAI instruction during consecutive years. The 39 effect sizes from the 26 studies ranged from -0.455 to 0.844. The positive overall mean effect size of 0.187 was lower than the 0.250 Cohen (1977) recommended to be a small effect. On the average, 57.2% of students who received both traditional and CAI instruction achieved higher academic scores than those students who only received traditional instruction. A typical student who used CAI moved from the 50.0 percentile to the 57.2 percentile. Because the results from the use of CAI indicated an academic achievement "improvement of 7.20 percentile ranks," the researcher concluded that CAI with traditional instruction was more effective than traditional instruction alone for students in grades six though twelve (p. 286). The study limitations included a lack of (a) participant descriptions (beyond being secondary students), (b) academic measurement descriptions, and (c) software descriptions. In

addition, software systems in this meta-analysis could not have been as sophisticated as technology that is available in today's AES systems (Waxman et al., 2003).

Computer assisted instruction may also provide students with opportunities for effective individualized response strategies. Brooks and Crippen (2001) analyzed 14,000 Web transactions from a site that simulated an Advanced Placement (AP) chemistry test (Crippen, 2000). The site was designed for linear use and based upon "repetitive testing and feedback" (p. 6). Tutoring was provided in the form of extended text explanations. From a database of 200 questions, each quiz was randomly generated with 8 items and corresponding tutoring. Students' items, answers, and tutoring use were automatically tracked.

The detailed analysis found that 24 of the 300 students participating in the study devised a back-and-forth methodology to answer questions one-at-a-time instead of eight-at-a-time, as designed and expected (Brooks & Crippen, 2001). Learning was measured by the average score per item. The learning rate of those 24 students was calculated as statistically significant, measuring at twice the rate of other students using the system. The researchers attributed this learning difference to the one-at-a-time item strategy reducing cognitive load (Brooks & Crippen, 2001). Students can replicate this optimum strategy of one-at-a-time error correction with an AES system (Educational Testing Service, 2007a). However, AES systems do not track how many or which errors are corrected.

Teacher Feedback

Teacher comments were analyzed from several different viewpoints. Straub (2000) categorized his own feedback to his college English class students in comparison

to teacher strategies for integrating assessment theory in the classroom. Sommers (1982) studied students' responses to teacher feedback, while Yagelski (1995) investigated students' responses to peer and teacher feedback. Smith (1997) analyzed the genre of end comments, and Matsumura, Pathey-Chavez, Valdes, and Garnier (2002) examined student writing in relation to teacher feedback.

Straub's (2000) research was selected because it clearly stated an example of response theory. The research provided background information on a teacher's point of view on feedback in the writing process. The study is a classroom-based, teacher-researcher examination of response to student writing. The researcher's goal was to provide suggestions to other teachers on how they might examine their own response practices in order to integrate assessment theory into their classrooms. The researcher-teacher examined his classroom responses from the perspective of seven response principles that were presented as teacher strategies, as follow:

1. Turn your comments into a conversation (p. 6).
2. Do not take control of the student's text (p. 8).
3. Give priority to global concerns of content, context, organization and purpose before getting (overly) involved with style and correctness (p. 10).
4. Limit the scope of your comments and the number of comments you present (p. 14).
5. Select your focus of comments according to the stage of drafting and relative maturity of the text (p. 14).
6. Gear your comments to the individual student (p. 15).

7. Make frequent use of praise (p. 17).

The study's participant was one student from the teacher-researcher's English course at Lehigh University. Two of the student's essays and the researcher's responses to them provided the data for the study. Generalizations could not be made about the use of this response theory because each teacher individualized its use.

The Sommers (1982) research question was, "...do teachers comment and students revise as the theory predicts they should" (p. 149)? The population description was defined only as 35 teachers at New York University and University of Oklahoma. The researcher studied teachers' commenting styles on first and second drafts, with all teachers commenting on the same sets of three student essays. This implies that a set is a first draft and the corresponding second draft. The teachers' essay comments were triangulated by having Writers Work Bench (WWB) score one of the papers and by conducting interviews with a representative sample of teachers and their students. WWB was a prototype for the current AES systems.

Much descriptive information was lacking in this research article. Since the teachers were from colleges, the participants must be college students, but no further information was provided (e.g., level of writing skill). No definitions were given to as to what constituted a first or second draft, nor were essay topics provided. A "representative number" of teachers and students were interviewed, but the actual number was not given (Sommers, 1982, p. 149). The interview questions were not provided, and the responses were generalized. The WWB's assessment responses were defined as "a sharp contrast" to the "arbitrary and idiosyncratic" comments from the teachers (p. 149). Computer comments were further described as "calm, reasonable language," while teacher

comments appeared "hostile and mean-spirited" (p. 149). No criteria were provided as a basis for identifying these characterizations.

The first finding was that the teacher's messages appropriated the student's text so the student was no longer focused on their purpose for writing, but rather the focus was on the teacher's purpose in commenting. This was identified as happening most often when teachers gave surface error corrections on the first drafts, as was found in the sample essays. The researcher also identified that the teachers' messages provided conflicting information because there was no way to determine which comments were primary and which were secondary. The comments gave editing and development recommendations on the same draft, confusing the revision process with editing and proofreading.

The second finding was that the same teacher comments were given to all of the texts, thus lacking specificity. According to the student interviews, they had difficulty understanding what the teachers' comments meant for them to do in their writing. This research lacked empirical methodology but did identify the frequency of teachers identifying "usage, diction, and style errors" on first drafts (Sommers, 1982, p. 150). The same feedback sequence Sommers (1982) recommended was later described by Straub (2000); different comments should be given on different drafts, first focusing on content and logic. It should be noted that Sommers' 1982 research compared teacher comments to WWB's computer evaluation comments, with the inference that the computer was more accurate than the teachers.

The Yagelski (1995) research studied the relationship between a senior high school classroom context and the revisions by the student writers. The quantitative data

of this study was triangulated with qualitative data such as field notes and interview transcripts. Students' writing was collected and coded for frequency and the type of revisions: (a) surface, (b) stylistic, (c) structural, or (d) content. The study was conducted at a senior-level, advanced composition course at a high school located in a suburb of a large midwestern city, over the period of a semester. The essays of 21 students were selected from the assignment genres of (a) description, (c) persuasion, and (c) cause-and-effect – a total of 55 essays having a total of 154 drafts, with an average of three drafts per essay. The first draft was the version submitted for peer review. The draft submitted for teacher review was labeled the second draft. There actually could have been more renditions of the essays than indicated by the version labels.

The essay coders were trained and had an interrater reliability of 92% on all codes. The coding of second and third drafts of the essays identified three statistical findings. First, the essay's genre had no significant influence on revision. Second, students made more surface (i.e., 31%) and stylistic (i.e., 50.7%) changes than structural (i.e., 4.2%) or content (i.e., 14.1%) changes. Finally, students made more changes to their second drafts (i.e., 37.7 changes per draft after teacher comments) than to their first drafts (i.e., 30.9 changes per draft after peer feedback). Even when a version received the teacher's feedback, 75% of students' changes were surface and stylistic, which the researcher noted supported the Sommers (1982) findings.

Another Sommers' (1982) research finding, the same teacher comments were given to all texts, was addressed in S. Smith's (1997) study of end comment genres. She proposed to (a) identify primary genres within teachers' repertoires, (b) determine features of these genres, and (c) define the patterns of genre usage. The first sample

analysis used comments from 10 Penn State teaching assistants on 208 papers written by first-year composition and rhetoric classes in 1993. No further information was provided about the teaching assistants. The randomly selected sample was representative of all possible scholastic grades from A through F. The second sample used data gathered by Connors and Lunsford (1988) for a large scale study of student errors. From their appeal to 1500 teachers, they randomly selected 300 papers from a national collection of 21,500 papers from 300 teachers (i.e., 20% response rate). S. Smith (1997) then discarded those papers that did not contain end comments and randomly selected papers for each grade category, as in the first sample, resulting in 105 end comments.

A detailed description was given about S. Smith's (1997) methodology for collecting the primary genres that made-up a teacher's end comments. A primary genre was described as a single sentence, a phrase, or a fragment. The 16 primary genres were categorized into 3 groups: (a) judging genres, (b) reader response genres, and (c) coaching genres. More detailed descriptions of the primary genres included their positive or negative tone and an explanation of how the genres are grouped. For example, end comments typically began with a positive evaluation, followed by a negative evaluation and coaching, and ended with either coaching or a positive evaluation. This study did not include any investigation of when end comments were used, nor did it reflect on the existence of additional comments in the participant essays.

The teachers in the S. Smith (1997) study developed standardized patterns (i.e., conventions) of end comments that were not as individualized as theoretical expectations might suggest. "More than four out of five teacher evaluations of the entire paper are positive, despite the even distribution of grades across the sample" (p. 253). This seemly

supports the Straub (2000) feedback strategy of using a lot of positives, but S. Smith (1997) hypothesized it created insincere feedback that could reduce the effectiveness of teachers' comments – but that hypothesis was not measured.

The judging genre comments formed the largest part of the primary end comments repertoire of a teacher. The grammatical subject patterns of the judging genres used an impersonal term, "the paper," in 46% of the evaluative statements (S. Smith, 1997, p. 256). The persuasiveness genre (i.e., a judging genre) about the writers' argument typically appeared on A and B papers, and two-thirds were positive. The evaluations genre (i.e., a judging genre) about a topic tended to appear on papers graded C or below, and three-fourths were positive. Judging genres also followed tone conventions, with 5 that were usually positive, 2 that were usually negative, and 4 that were not associated with negative or positive. When used, 86% percent of judging genres were positive and most frequently written as fragments about the entire paper, for example, "good paper" (S. Smith, 1997, p. 255).

While the other two genres made up only 5 of the 16 total primary genres, they also revealed patterns of construction and usage. There were two reader response genres that allowed the teacher to respond like an active reader. The identification genre was a response to the personal experience rather than the writing. The reading experience genre was often used as evidence to support an evaluation. It was usually written as an "I" statement, providing the teacher's point of view and a more personalized response. The coaching genres were composed of three different types of comments. First, suggestion genres for the paper currently being evaluated pertained to content 84% of the time and expression (e.g., clarity) 16% of the time. Second, coaching genres for future papers

focused 35% on content, 47% on expression, and 18% requested the student put more effort in the future paper. The final coaching genre offered assistance to the students.

A teacher usually made an end comment by selecting four or five primary comments from the repertoire, resulting in a secondary genre. Eighty-eight percent of the end comments began with a positive evaluation, transitioned to negative evaluation and coaching, and concluded with coaching and positive evaluation. Nearly all the conventions in the primary and secondary end comment genres followed the key patterns across the national and Penn State samples. Suggestions were made to improve the effectiveness of teacher comments. Overall, teachers' written responses were not as individualized as expected.

Clare, Valdez, and Patthey-Chavez (2000) studied teachers' written feedback in relation to the quality of students' work in five urban middle schools, as part of a University of California at Los Angeles (UCLA) study funded by the U. S. Department of Education. The data were collected as part of a larger study on evaluating large-scale school reform affects on student learning. Over a period of 2 years, 64 essays, including rough and final drafts from 4 "typical" language arts assignments, were studied (p. 4).

The seventh-grade participants were mainly minority students who were English language learners (ELL), as were 44% of their schools' populations. The schools' enrollments were specifically defined by ethnicity: (a) Asian, (b) African American, (c) Latino, (d) White, or (e) other. The largest percentages of the students were Latino. The schools' free or reduced lunch participation ranged from 56.6% to 86.9%. The 11 middle school teachers' experience varied from 2 to 28 years. Teachers submitted an information sheet on each project, along with four samples of student's work. The information

provided included (a) the categorical identity of the feedback provider (e.g., peer, teachers, peers and teachers, or none), (b) the writing genre, and (c) the mean number of words in the students' essays (i.e., 270). Two samples were to be "medium" quality and two of "high" quality (Clare et al., 2000, p. 5). The teachers' criteria for the quality ratings were not provided.

A researcher categorized each essay's feedback as either content feedback, which "encouraged students to add or delete content and/or restructure content" or surface feedback, defined as "word choice, spelling, grammar, and punctuation," (Clare et al., 2000, p. 3; Olson & Raffeld, 1987). The random re-categorization of 20% of the feedback showed an interrater reliability of 80% (Clare et al., 2000). The amount of feedback was identified with a ratio calculated by dividing the number comments and edits by the number of words in an essay. Bilingual raters used three standards-based, 4-point scales "measuring organization, content, and writing mechanics, use of language, grammar, and spelling (MUGS)" (p. 8). The scales were developed by the University of California, Los Angeles in partnership with the Los Angeles Unified School District and United-Teachers, Los Angeles. Each dimension was rated on a 4-point scale, from 1 (i.e., poor) to 4 (i.e., excellent). Using the same interrater reliability methodology described previously, 81% agreement was found.

The relationship between the type of teacher feedback and student writing quality was analyzed using correlation coefficients. T-tests for paired samples investigated the quality changes between earlier and final drafts. Regression analyses identified the influence of teacher feedback on the quality of the final drafts. A more qualitative

analysis tracked teacher recommendations from draft to draft to determine if students implemented the recommendations.

The analysis of the nature and amount of teacher feedback to middle school students revealed that (a) 8% of middle school students did not receive any feedback on their drafts, (b) 58% received surface-level feedback, and (c) 38% received content-level feedback. Essays receiving surface-level feedback increased in length by an average of 16.86 words, but essays receiving content-level feedback increased by an average of 48.1 words – more than twice as much of an increase. In spite of this fact, the quality from students' first drafts to final drafts remained constant, with no effect from either type of feedback. Thus, higher quality first drafts became higher quality final drafts, while lower quality first drafts became lower quality final drafts. The mechanics in students' writing did improve in direct relation to the feedback they received – so they followed teachers' surface-level recommendations – but there was no statistically significant change in overall quality. The qualitative examination of the content feedback revealed that most of it was about word change, and students did follow those teacher recommendations.

Overall, teacher feedback research showed a variety of effects on student revision. Clare et al. (2000) and Yagelski (1995) showed that students actually made more surface than structural changes. Though these studies were from peer reviewed journals, Yagelski (1995) identified the small sample size as a study limitation. Clare et al. (2000) linked the type of student changes to the type of teacher feedback, in that more surface feedback from teachers led to more surface revision by students. The Sommers (1982) study characterized teacher feedback as generic and hard for students to understand. This peer-reviewed journal article also omitted pertinent descriptive sample information, such as

the definition of a draft and the writing prompts. S. Smith (1997) identified genres of end comments and mapped their usage patterns. By showing that teachers gave final comments following identifiable conventions, S. Smith (1997) suggested that teacher feedback was not as individualized as might be expected. The majority of teacher feedback was positive – supporting that teacher feedback strategy as presented in Straub (2000).

Automated Essay Scoring Feedback

Automated essay scoring vendors recommend that such applications be used in the classroom only as a supplement to teachers' feedback (Burstein, Chodorow et al., 2003; Burstein & Marcu, 2003; Ware & Warschauer, 2006). In this section, the available AES functionality is surveyed, and then the Ware and Warschauer (2006) study on the use of AES in the classroom is examined. That study also leads to the identification of students' computer gaming perceptions affecting the use of AES systems. Finally, Chen and Cheng (2006) investigated the use of an AES system in three classes of third-year English majors at a national technological university in Taiwan, China.

Summative human feedback encompasses the sophisticated, expensive protocols necessary for human, holistic scoring of large-scale testing. Human graders must be given interrater reliability training, and their grading requires reliability checks. Research consistently demonstrates a typical 97% agreement of holistic scores between human raters and the AES *Criterion*, even for the Test of English as a Foreign Language (TOEFL) exams (Chodorow & Burstein, 2004; Rudner et al., 2006; Shermis & Burstein, 2003). This interrater agreement also held true in those cases where a third human rater is required to resolve a discrepancy between the AES and human raters. Students also need

formative, trait evaluations in the classroom in order to improve their writing (Wolcott & Legg, 1998; Wolfe-Quintero et al., 1998). With classroom use, an AES system might reduce the number of hours that teachers must spend on grading essays, or students may have more writing opportunities with feedback without a corresponding increase in teachers' grading time (Ware & Warschauer, 2006).

How teachers set-up an assignment directly affects the AES feedback that students receive (Educational Testing Service, 2006b). *Criterion* provides holistic and formative feedback for both the system-provided writing prompts and those prompts created by the teacher. In order to evaluate teacher-created prompts, ETS developed a content vector analysis calculus (i.e., algorithm) to identify *unexpected topic* and *bad faith* essays (Burstein & Higgins, 2005 p. 4). The algorithm was successfully tested with 8,000 *unexpected topic* and 732 *bad faith* essays. A tutorial guides teachers through the creation of their own prompt, with the goal of providing prompts that facilitate students' writing (Educational Testing Service, 2007a). The writing prompts can be set for a specific grade level (e.g., ninth or tenth) and as several genres (e.g., persuasive or descriptive) (Educational Testing Service, 2007c). The vendor states that *Criterion*'s purpose is not to evaluate creative writing.

A new prewriting function is also available for teacher selection (Educational Testing Service, 2007b). It provides eight different strategy templates for students to choose for planning: "Outline, list, idea tree, free writing, idea web, compare & contrast, cause & effect, and persuasive" (Educational Testing Service, 2007b). The text that a student enters into the selected planning template is automatically entered into the text editing screen in the organizational hierarchy provided by the template. A split screen is

also available for viewing the filled-in planning template on the top of the screen, while working on the actual essay below.

The teachers are provided with the flexibility of selecting whether or not to provide holistic scoring and whether to use a holistic scoring range from 1-6 or 1-4. In addition to receiving a holistic score, students can access a generalized description about an essay receiving their holistic score. For example, an essay with a holistic score of 3 out of 4, "Is well organized with transitions, maintains focus;" and "contains errors in grammar and conventions that do not generally interfere with understanding" (Educational Testing Service, 2006b). Thus holistic feedback is both positive and negative. For system prompts, students may also view sample essays for each rank in the range of holistic scores.

Teachers also choose which categories of trait evaluations are available to students (i.e., grammar, usage, mechanics, style, and organization and development) (Attali, 2004). A count is kept on the number of times a student has submitted an assignment's essay. Only the first essay of an assignment and the last submission are retained in the AES, so there is no way to measure which feedback suggestions were actually followed. When students are provided with trait evaluations, they may view the evaluated essay on the top of a split screen, while working on the revision below (Educational Testing Service, 2007a). The different grade (e.g., ninth or tenth) that is selected in the AES can impact the level of error explanation text provided, and the level of the *Writers Handbook*. The *Writer's Handbook* is an extended explanation on how to correct the errors (Educational Testing Service, 2006b).

Another opportunity for teachers is the choice to include their own feedback for use on individual student essays. In addition, a library of teacher messages can even be created, which is reminiscent of the end comment repertoires examined by Smith (1997). Teachers' feedback is presented as electronic post-it notes on selected essay areas (Educational Testing Service, 2006b). Each student's essays are also collected into an individual, online portfolio.

The individualized student feedback from *Criterion* is, for the most part, surface feedback: (a) spell checking, (b) style, (d) mechanics, (e) grammar, and (f) usage (Attali, 2004; Educational Testing Service, 2006b; Ware & Warschauer, 2006). A list of the subcategories is available in Appendix A. Style feedback also includes the number of words, number of sentences, and average number of words per sentence. The student can select to see a category's errors one at a time or all at once. The exact error is highlighted within the essay and positioning the mouse on the highlighted area provides a brief explanation of what the error means (Educational Testing Service, 2006b). The trait feedback is both negative and positive – depending on whether or not errors are indicated.

The trait feedback uses natural language processing (NLP) and statistical machine learning, but the AES trait scores have not been studied as extensively as the holistic scores (Burstein, Chodorow et al., 2003). Attali (2004) conducted a study for ETS on the usefulness of *Criterion*'s formative feedback by measuring the change in feedback from the first to the last submission of an essay. Essay length was included since it has a high correlation to writing quality. The research took place during the 2002-2003 school year, but little was known about the participants except their grade level. Only the first and last

essay submissions were available, providing (a) the corresponding scores and feedback reports, (b) the number of submissions that occurred per prompt, and (c) the grade level of the prompt. By setting the limit to essays of 50 or more words, 33,171 essays from sixth through twelfth grades that used system-provided *Criterion* prompts were evaluated. Of these, 71% (23,567) were submitted only once. The remaining 9,604 were reduced to 9,275 (i.e., 97% of the population of multiple submissions) by selecting those submitted only 10 times or fewer.

Among the essays submitted multiple times, the initial essays' lengths were shorter and received lower holistic scores that those essays that were submitted only once, but the differences were not significant. However, there were other significant differences between the initial and final essays having multiple submissions. Holistic scores, based on a five paragraph model, improved by an effect size of .47, and length increased an average effect size of .39. Development scores increased by an effect size of .31, while error rates for grammar, usage, mechanics, and style decreased by an effect size of .15 to .27. Of the 33 measurements, 23 were significantly changed between the first and last essays. Overall, students found and corrected about 25% of their errors.

Criterion goes beyond surface errors by evaluating an essay's organization and development. A group of three discourse analysis programs use machine learning to identify the discourse elements (e.g., topic sentence) (Burstein, Marcu, & Knight, 2003). A large number (i.e., 989) of twelfth grade essays were scored by *Criterion* and human scorers to test *Criterion*'s coherence analysis (Higgins, Burstein, Marcu, & Gentile, 2004). The researchers found that *Criterion* was able to identify sentences' "relationship

to the topic, relationship to other discourse elements, relevance with discourse segment, and errors in grammar, usage, and mechanics" (p. 2).

Criterion's organization and development feedback are in the same format as described for the surface-level feedback, using highlighted areas and a mouse-over function (i.e., messages pop-up depending on the location of the cursor). The organization and development feedback is both positive and negative since it indicates which elements do exist, as well as those missing. The color coded presentation of the parts of the essay also enables the student to see if there are sequencing problems in the essay (e.g., conclusion sentences interspersed throughout the essay).

A recent, mixed-methods study investigated the use of AES systems in the classroom (Grimes & Warschauer, 2006). As part of a larger 1:1 laptop initiative study, Grimes and Warschauer (2006) studied the use and outcomes with the use of AES systems at three junior high schools and two high schools. Some schools were high-SES and some were low, one had a majority of European-Americans and Asian-Americans, while another had a majority of Latinos. The teachers, "selected by availability," were mostly language arts or English teachers (p. 7). Three schools used *My Access!* and two used *Criterion*. Data included semi-structured interviews of three principals, three technical administrators, and nine language arts teachers. Twenty language arts classes were observed, two focus groups were conducted, and over 2,400 *MY Access!* reports and student essays were examined. Nine teachers and 564 students in the 1:1 laptop schools responded to the *MY Access!* surveys.

Data were analyzed for usage patterns, attitudes, and social context. The data provided high opinions from teachers and administrators of the AES systems, including

support for students' increased motivation in writing and development in creative writing. However, the actual use by seventh grade students in the two 1:1 laptop classes was only 2.38 essays per student during the whole 2004-2005 school year, with even less use in the lower 1:1 grades and the non-1:1 schools. The most frequent reason for low level of use was the lack of available classroom time due to the need for preparation for state tests.

The teachers did not feel the *MY Access!* scores were always fair. Their average rating of "fair and accurate" scores was 2.71 on a scale of 1 to 5, with 3 as neutral. The teachers did feel the numerical score (i.e., holistic score) helped students improve their writing. The students had higher opinions of the numerical score, rating it as 3.44 in fairness. Another research variable was that it was the first year of AES use for the teachers, with the exception of one teacher in her third year of AES use. The experienced teacher only spot-checked students' essays, while the other teachers continued grading with a concerned focus on fairness.

The most important AES feature, all teachers agreed, was the speed of response, because it was a strong motivator. The immediate feedback was also supported as the most important AES feature by teachers in the Delphi study by Boone and Frost (2005). The Delphi study also reported that students liked seeing their score improve, and they aimed for higher scores via revision. The teachers in the Grimes and Warschauer (2006) study even reported that students responded to their holistic scores much like when receiving a computer game score – with shouts of joy or groans of dismay.

While it is questionable that AES developers sought to create a gaming environment, the AES systems do seem to meet the game definition provided by Juul (2003):

> A game is a rule-based formal system with a variable and quantifiable outcome, where different outcomes are assigned different values, the player exerts effort in order to influence the outcome, the player feels attached to the outcome, and the consequences of the activity are optional and negotiable.

Automated essay scoring systems, following the game definition, are based on English language writing rules, and different ratings (e.g., holistic score or trait errors) are provided for different performances. Some students seemed attached to their holistic score outcome, as indicated by vocal responses to the scores. Within the constraints of an essay's AES set-up, it is up to the student as to how many revisions are created, so consequences are varied.

Juul (2003) also defined a player's relationship to a computer game with three components: (a) some outcomes are positive and some are negative; automated essay scoring systems meet these criteria by providing a range of holistic scores, some positive and some negative, (b) the player must extend an effort or do something; the students must write when using an AES system, thus meeting the criteria, and (c) the player is happy if they win and unhappy if they lose the game, based on the Grimes and Warschauer (2006) study, AES students were happy with a high holistic score and unhappy with a low score. Thus, AES systems seem to meet these components of a player's relationship to a computer game.

Gee (2003) identified that human learning is based on *practice effect*, something that good video games provide (Gee, 2003). While AES systems do provide writing

practice, they are not exactly like video or computer games because they do not include graphics or a story line. AES systems, however, could be called a simulation because they are trying to model the results from human holistic scoring of essays.

Students' comfort with computer gaming is indicated by statistics from the Entertainment Software Association (2007): (a) 31% of video and computer game players are under 18-years-old, (b) 36% of the most frequent computer game players are under 18-years-old, and (c) 62% of the computer game players are male and 38% are female. It is important to note that statistics are not available on the gender distribution by age. On their own, students developed their perceptions of the AES holistic score's similarity to computer game scores (Grimes & Warschauer, 2006). Students' attachment to an AES's holistic score may indicate that students are actively and critically involved in their learning process, which is a goal for all education (Gee, 2003).

Grimes and Warschauer (2006) found it difficult to correlate actual revisions to outcomes because students could re-submit after changing one word. *MY Access*! reports indicated that 72% of the seventh grade essays were not revised at all and 28% were only revised after receiving a preliminary score and feedback. Sometimes the initial draft of an essay would be spread over three class periods, which gave students the opportunity for either three submissions or just saving and not submitting for evaluation. Therefore, the researchers discounted this AES revision counter. Further verification for the lack of importance of the revision counter was provided by a survey of 10 revised essays that only showed changes in superficial features.

The seventh graders' scores on the language arts portion of the 2005 California state tests did not show any outcome changes after AES use. However, the infrequent use

of the AES systems precluded any expected changes. In interviews, teachers noted that the AES systems assisted the writing development of all students, no matter what special learner categories existed, such as (a) English language learners, (b) gifted, (c) special education, (d) at-risk, and (e) students without any special needs. These teacher opinions were not based on any scientific measurement of students' writing development.

Chen and Cheng (2006) studied the use of an AES system in three college classes of third-year English majors in Taiwan, China. The classes were different sizes: (a) 26, (b) 18, and (c) 14. The data included (a) 53 students' responses to a questionnaire (i.e., by class, 21, 19 and 18), (b) writing samples, (c) AES feedback, and (d) three focus-group interviews with 16 participants who represented the three classes (i.e., participants by class were 5, 5, and 6). The surveys investigated the students' views and reactions on the use of the AES system to improve their writing. The focus group had students talk about how the AES system was used in their class, and what they thought of it as a writing tool (i.e., diagnostic feedback) and an essay grader (i.e., holistic score). The writing samples and *My Access!* response data was used to triangulate the student interviews.

The highest satisfaction rating (i.e., 71%) from the students was for the speed of response from the AES system. The greatest dissatisfaction with the AES system was that the grading (i.e., holistic score) was not considered fair (i.e., 63%). For example, one student wrote an essay without a conclusion but still received a high score (i.e., 5 out of a 6-point holistic rating scale). The second problem was that the AES system did not provide trait (i.e., diagnostic) feedback that was individualized enough. The participants found the AES feedback helpful for early drafts, but subsequent revisions would keep receiving the same holistic score but without changes to the trait feedback to guide

revisions. The students would depend on their instructors to get more individualized feedback. It should be noted here, again, that AES systems are not promoted as a replacement to instructors (Burstein, Chodorow et al., 2003; Burstein & Marcu, 2003; Ware & Warschauer, 2006).

Participant ratings of the individual parts of the diagnostic feedback (e.g., My Editor or Thesaurus) on a scale of 1 to 5 found only 40% of the students perceived the individual functionality as helpful (Chen & Cheng, 2006). In an overall rating of the AES system, 55% of the participants found the AES system was moderately or slightly helpful. However, 45% did not find it at all helpful. When analyzing those ratings, Chen and Chang (2006) found the pedagogical differences in the use of the program were more important than the functionality of the AES system.

There were many commonalities among the three teachers. They (a) attended the one hour AES training session, (b) had similar class objectives, (c) used the same textbook, (d) taught similar content, and (e) used a similar process-oriented curriculum. The differences in use included (a) the teachers' familiarity with the AES and technology skills (e.g., low to high), (b) the number of essays graded by the AES system (e.g., ranging from two to six), (c) teacher feedback frequency (e.g., ranging from after each essay to only at the end of the semester) and the grading policy as it related to the AES program (e.g., ranging from no importance of the AES score to the AES score counting for 40% of the final grade).

Only 14% of the students in Class A thought the AES system was of no use, compared to 72% and 58% in the other two classes. Using data from the interviews of five students from Class A, their teacher was described as (a) very familiar with the AES

system, (b) having a high level of technology skills, (c) providing detailed instructions and demonstrations on the use of the AES, (d) requiring students to have at least a holistic score of 4 before handing the paper in to the teacher to grade, (e) giving individual, written feedback on each essay and (f) holding class discussions about the feedback from the teacher and the AES system. The researcher concluded that the teachers' pedagogy influenced the students' perceptions of the usefulness of the AES system. It also shows that the use of educational technology cannot be separated from teachers' instruction.

There is no way to track exactly which AES feedback has been implemented by the students. The revision that happens with the use of an AES system may actually occur over multiple revisions. The available categories of trait feedback that students receive are dependent upon how the teacher has set-up the assignment. The fact that AES systems provide more surface than content feedback is similar to what research has found about teacher feedback to students (Clare et al., 2000; Yagelski, 1995). Teachers credited the AES with increasing writing development, but there was no specific measure of this fact.

Summary

Nevada's requirement that high school students pass the writing proficiency examination in order to graduate from high school places a focus on students producing a high quality product (Nevada Department of Education, 2007). It is agreed that while high school teachers are overwhelmed with grading student writing, more writing opportunities need to be provided in order to prepare students for high stakes testing (MacArthur, 2006). Perhaps AES systems can help teachers provide students with more

writing opportunities with skilled feedback. Research has shown that computer access needs to be 1:1 for optimum classroom use of educational technology, thereby supporting teachers' integration of the technology into the classroom.

An AES systems' analysis of mostly surface features is similar to the surface feedback frequency that has been found in research on teacher feedback. The strength of a CAI system, such as an AES system, is its immediate feedback to students. Previous research studies have initially shown that some students have valued the holistic score provided by an AES system, much like a computer or video game score, indicating their engaged learning. However, AES systems fail to account for social learning, which is considered a key component in linguistic development (Ware & Warschauer, 2006). The current research will meet the need for research to compare students' writing proficiency and writing development with and without the use of an AES system. The results from AES trait error categories will also be investigated with and without the use of an AES system. Finally, student's degree of user satisfaction will be explored, along with the impact of gender on all the research.

CHAPTER 3

METHODOLOGY

In order to become better writers, research has shown that students need to participate in the recursive processes of writing and revision (Shermis et al., 2006). Students' frequent writing and revision needs to include skilled (e.g., a teacher's) feedback (Burstein, Chodorow et al., 2003; Nippold et al., 2005; Pritchard & Honeycutt, 2006). External factors that improve students' writing include teachers' classroom instruction, writing feedback, and pedagogy (Berninger, Fuller, & Whitaker, 1996; Bruning & Horn, 2000; Nippold et al., 2005; Pritchard & Honeycutt, 2006; Scott, 1988).

The impact of educational technology on students' learning, cognitive development, and linguistic development is difficult to separate from other external factors (Berninger et al., 1996; Schrum et al., 2005). This study employed the use of an automated essay scoring (AES) system in combination with teacher-led writing instruction. The AES system was used as a classroom intervention to provide additional skilled feedback opportunities for students.

The AES measurements (i.e., holistic score and trait feedback categories) were calculated for both the treatment and control groups, though only the treatment group used the AES instructionally. Definitions are available in Appendix B. Writing from both groups was scored using the AES system holistic score, human rater's holistic

score, and human rater's words per t-units (W/T). See Appendix B for definitions. This study was guided by the following research questions:

1. Is there a significant difference in the writing proficiency improvement of students who use an AES system in combination with teacher-led writing instruction compared to students who receive only teacher-led writing instruction, with assessment based on holistic scores from human raters and an AES system? Is gender a significant factor in the results?
2. Is there a significant difference in the writing development of students who use an AES system combined with teacher-led instruction compared to students who receive only teacher-led instruction, as measured by words per t-units (W/T)? Is gender a significant factor in the results?
3. Is there a significant difference between pre- and post-test AES trait error feedback categories for those students who use an AES system combined with teacher-led instruction when compared to those students who had only teacher-led instruction? Is gender a significant factor in the results?
4. What was the degree of user satisfaction for the students who used the AES system as measured by a survey and semi-structured interviews? Is gender a significant factor in the results?

Research Design

Quasi-experimental

In order to meet evidence standards, a scientific study's design must be a randomized controlled trial (RCT) or quasi-experimental. An RCT design is very difficult for an educational technology study to achieve outside of a clinical setting. In this case,

the research method is constrained by the use of educational technology as an intervention.

Research has shown that sufficient computer access must be assured in order for educational technology to possibly be a successful intervention (Grant et al., 2005; Russell et al., 2004; Smerdon et al., 2000). Teachers are more likely to integrate technology into their classroom curriculum when the student to computer access ratio is 1:1, thus leading to more opportunities for students' use of educational technology (Smerdon et al., 2000). Therefore, experimental random selection was not possible for this study due to aforementioned constraints. The focus was on factors to consider for a quasi-experimental design.

The Nonequivalent Comparison Control Group (NCCG) design may be the most common of all quasi-experimental designs (Beins, 2004; Campbell & Stanley, 1963; Committee on Scientific Principles for Education Research, 2002; Creswell, 2002; McMillan, 2004; Mertens, 1998). It is often used, as is the case here, where the participants are in pre-existing groups, such as classrooms. A pre- and post-essay was used to measure the performance-based outcomes of writing proficiency, development rate of change, and AES trait scores of both the treatment and control groups.

This was a mixed-methods study of quantitative and qualitative data, with the choice of several of the instrument measures and data types being controlled by the choice of the AES system, *Criterion* (Educational Testing Service, 2007a). The quantitative data included human raters' and AES holistic scores, AES trait scores, W/T (i.e., a writing development ratio), and a student satisfaction survey. See Appendix B for definitions. Qualitative data were collected from teacher interviews about the classroom

use of the AES system and student interviews about their satisfaction with the use of an AES system.

Potential Threats to Validity

A research study has potential threats to internal and external validity that need to be accounted for in its design. Internal validity means that the study has been designed so that the causal relationship between the dependent and independent variables is not compromised by interference of unrelated variables (Beins, 2004; Campbell & Stanley, 1963; Committee on Scientific Principles for Education Research, 2002; Creswell, 2002; McMillan & Schumacher, 2006; Mertens, 1998). Several issues are threats to the internal validity of this study. The major weakness of quasi-experimental design is the assignment bias, meaning that the participants may differ in some unexplained way. The pre-test helps address this issue by defining possible issues in the beginning of the study. In addition, several factors were used to help match the participants in the treatment and control groups. The teachers were from the same school, the same English department, and the same student and teaching teams. The participants were from the same grade level, ninth.

Another internal threat to the NCCG design is that maturation may occur at different rates for individuals. A strength of this study is that the writing development measure helped identify class level (i.e., standard or honors) differences, thereby minimizing individual differences.

It is known that the effects of educational technology use on student learning are difficult to separate from other variables that may also affect learning (Schrum et al., 2005). Other strengths to this study are that the school district is committed to providing

the 1:1 level of students' computer access, and the treatment teacher had used the AES treatment system for three years. The level of access increases the potential for students' comfort with the intervention and the potential level of teachers' integration into the classroom. The teachers' experience means there is less risk of poor implementation or compromised fidelity from the educational technology intervention.

The different characteristics of the participants (e.g., only students with a low fluency rate drop out of the study) may negatively affect the internal validity threat of mortality. This study could be impacted by the mortality rate because the urban school research setting had a high transience rate. The statistical regression threat occurs when extreme (e.g., only honors or only remedial) groups of participants are used in the research. This study used both standard and honors groups for the ninth grade participants. There may be other threats to validity that are as yet unknown.

Educational Scientific Research

Emphasis on scientific educational research has resulted from the No Child Left Behind (NCLB) Act of 2001 (2002). The following scientific research constructs, not necessarily in order, were followed by this study (Committee on Scientific Principles for Education Research, 2002; Creswell, 2002; Kingsley, 2005; North Central Regional Educational Laboratory, 2004; Phye, Robinson, & Levin, 2005; What Works Clearinghouse, 2006):

1. Empirical methods are to be appropriate, systematic, uniform, and followed in detail.
2. The design method should be experimental or quasi-experimental.

3. The data are to be provided by measurement methods that are reliable and valid.
4. The method should provide enough detail to enable replication.
5. Data analysis should use methods that examine the problem and justify the conclusions.

This study followed the guidelines for educational scientific research as detailed by the NCLB, and the What Works Clearinghouse sponsored by the Institute of Education Sciences and the U.S. Department of Education (No Child Left Behind Act of 2001 et al., 2002; What Works Clearinghouse, 2006).

Participants

Setting

The research setting was a large urban high school in a large southwestern school district during the fall and spring semesters of the 2006-2007 school year. The technology leadership of a large school district selected the research site. Aggregated data are available to characterize the school for the research year (Nevada Department of Education, 2008). There were 3029 students in the school and 947 ninth grade students. The school was divided equally between males and females. The graduation rate was 42.2% and the transiency rate was 36.4%.

The ethnicity of the specific class groups from which the research sample population was drawn is available from a teacher survey. The ethnicities are shown in Table 1. The survey included the entire class' students, more than just the research participants, encompassing 48 from the treatment standard classes, 53 from the treatment

honors classes, and 29 from the control standard class. It was from this available group that the participants volunteered.

Table 1

Ethnicity of Class Levels from Which Participants Joined Study

	Treatment	Treatment	Control
Ethnicity	Standard	Honors	Standard
Caucasian	6	11	14
African American	10	2	7
Asian/Pacific Islander	8	11	3
Hispanic	69	72	66
Other	6	4	10

Note. Numbers are percentages

Teachers

The technology leadership of a large school district selected a large urban high school research site and the teachers whose students served as the treatment participants. The treatment group's teacher was selected by the school district to use the AES to supplement her classroom writing instruction. This treatment teacher was selected from a population of prior participants in a AES research study (Boone & Frost, 2005). The selected treatment teacher, in turn, chose the control group teacher whose class most closely corresponded to the treatment classes. The teachers were from the same school, department, and class teaching teams and did their lesson planning together. The

treatment teacher had 3 years of AES annual school district training and classroom AES use experience. The control teacher had taught for 8 years and the treatment teacher had taught for 10 years.

Students

The participants were 9th grade composition students in either treatment standard, treatment honors, or control standard classes with one of two teachers. Student gender was tracked for two reasons. The first was to verify the current research about computer access that no longer deems gender an issue (Day et al., 2003; Parsad & Jones, 2005). The second was to identify possible gender developmental issues (Berninger & Swanson, 1994; Day et al., 2003; Santrock, 2005).

Protocol

Teacher Interviews

Qualitative data from the semi-structured teacher interviews were used to provide descriptions of the settings that were used in collecting the test data. This assisted with the study's ability to be scientifically replicated. The teachers' interview questions, available in Appendix C, were formulated to enable a comparison to previous research on teachers' use of an AES system (Chen & Cheng, 2006; Grimes & Warschauer, 2006).

Writing Prompts

Both the control and treatment groups received writing instruction from their classroom teachers, who had all the essays first drafted by hand. The same persuasive writing prompts were given to the treatment and control groups. This research had the teachers choose persuasive essay prompts, either system- or teacher-provided. It was expected that the experienced teachers would know what is best for their students and

what best fit into their curriculum. The pre-test prompt was as follows: "Construct two paragraphs supporting your opinion of whether Odysseus was or was not a hero in the space provided. Make sure you are supporting your opinion with examples from the textbook." The post-test prompt was as follows: "Teenagers don't know what true love really feels like. Agree of disagree? Persuade with strong support."

The effects of writing prompts on student proficiency outcomes are called "prompt effects" (P. LaMahieu, personal communication, January 19, 2007). Prompts are impacted by writers' (i.e., students') interpretations, which may differ from that of the writing prompt creator/teacher (Ruth & Murphy, 1984). Differences exist because the students and teachers have different knowledge and background experiences. Students may also differ in the way they construct the task, depending on whether they are skilled or novices in the writing genre.

Writing in different genres can also be impacted by writing development (Nippold, 2000). A meta-analysis determined that while adolescents' syntax development is "gradual and subtle," it was more evident with persuasive writing than descriptive or narrative genres (2000, p. 6; Scott, 1988). In addition to being more revealing of writing development, persuasive prompts are one of the genres used in the state writing proficiency exam (Nevada Department of Education, 2006-2007).

Developmental Index

Words per t-unit (W/T) is a writing development measure that is not under the conscious control of the writer (Hunt, 1970; Loban, 1976; Nippold et al., 2005; Wolfe-Quintero et al., 1998). It is one of the writing development ratios that measure fluency, accuracy, and complexity (Wolfe-Quintero et al., 1998). A minimal terminal unit (i.e., t-

unit) is defined as one independent clause plus all associated dependent clauses (Hunt, 1965a, 1965b; Nippold et al., 2005; Polio, 1997; Wolfe-Quintero et al., 1998). A t-unit is a little different than a sentence in that a compound sentence would be measured as two t-units. An independent clause consists of a subject, a main verb, and expresses a complete thought (Nippold et al., 2005). Dependent clauses also include a subject and a verb, but need to be linked to an independent clause to complete an idea. The three types of dependent clauses include: (a) a relative/adjective clause that describes a preceding noun; (b) an adverbial clause that expresses condition, time, or manner; and (c) a nominal clause that acts as the subject. Definitions are also available in Appendix B.

In order to determine the W/T, the t-units were calculated by the researcher and verified by a masters' student with a 99% agreement on a random sample of 20% of the essays. The number of words in each essay was calculated by the AES system.

Treatment Group

Automated Essay Scoring System. The technology intervention, an AES system, is designed to affect student writing and is targeted at the 9-12 grade population (Educational Testing Service, 2007a). According to the vendor's online materials, *Criterion* Online Writing Evaluation (Educational Testing Service, 2007a), is a Web-based writing system that gives teachers and students individualized evaluations on submitted essays almost immediately. The immediate feedback is a characteristic of computer-assisted instruction (CAI) (Christmann et al., 1997). Automated essay scoring systems also simulate the summative holistic scores used by high-stakes tests to measure writing proficiency. Such systems also provide formative data (e.g., trait analysis and spell checking) so students can improve their writing by revision, in a self-paced manner.

The AES's trait categories include grammar, usage, mechanics, style, and organization and development (Burstein, Chodorow et al., 2003; Burstein & Marcu, 2003; Educational Testing Service, 2007a; Ware & Warschauer, 2006). The trait scoring categories and subcategories are available in Appendix A. The writing prompts for the AES can be system- or teacher-provided, and teachers may create their own feedback messages. The immediate and personalized feedback from the use of an AES, according to Educational Testing Service, should be considered only as a supplement to teachers' feedback (Burstein, Chodorow et al., 2003; Burstein & Marcu, 2003; Ware & Warschauer, 2006).

The optimal 1:1 ratio for Internet-connected computers was provided through mobile carts of laptops in the classroom (Grant et al., 2005). Each class had a cart of 30 laptop computers. It was up to the treatment teacher to decide how and when to use the AES system in the classroom. The teacher was also responsible for training her students how to utilize the AES in use.

The procedures for creating the pre- and post-essay treatment samples were almost the same. The teacher of the treatment participants set-up the AES so students received (a) individualized holistic scores on a 6-point scale, (b) all the available trait scores, (c) the persuasive prompt, (d) the prompt's grade level of ninth grade, and (e) the number of possible submissions. The pre-test submissions were limited to five, but the post-test submissions were unlimited.

Student Interviews. Question Four about the degree of user satisfaction for the students who used the AES system was answered by a combination of survey and semi-structured interview. The data were modeled to extend previous research on students' perceptions of the helpfulness of AES systems (Chen & Cheng, 2006; Grimes &

Warschauer, 2006). The questions were modified to address the functionality of the AES. The survey and interview questions are provided in Appendix D.

Control Group

The control group used Microsoft Word to publish their essays, either at home or school. The home use version is not known, but 2003 Microsoft Word was used at school. The electronic files were not available to the researcher, so the essays were re-created electronically.

Data Collection

Pre-essay samples were gathered during the Fall, 2006 semester and post-essay samples toward the end of the Spring, 2007 semester. Each test sample (i.e., pre- and post-test) consisted of the final draft of one essay. Standards for the protection of research participants have been met for the University of Nevada, Las Vegas and the participants' school district. Table identifies the data being measured for each question and provides the timing of the data collection, after which the NWP scoring, the treatment group, and control group are addressed in order to further explain the data collection procedures.

Table 2

Data Measurement and Collection Timing for Each Research Question

Research Question	Data measurement	Data collection timing	
		First or pre-test	Second or post-test
Question One	AES holistic score	November, 2006	May, 2007
	NWP holistic score	November, 2006	May, 2007
Question Two	W/T	November, 2006	May, 2007
Question Three	AES Grammar errors	November, 2006	May, 2007
	AES Usage errors	November, 2006	May, 2007
	AES Mechanics errors	November, 2006	May, 2007
	AES Style errors	November, 2006	May, 2007
	AES Organization and development	November, 2006	May, 2007
Question Four	Student survey and interview	May, 2007	
Setting	Teacher interviews	November, 2006	May, 2007

National Writing Project Holistic Score Collection

The writing prompts and the pre- and post-essay samples from both the treatment and control groups were sent to the National Writing Project (NWP) for scoring by human raters. The control group essays had to be typed so they did not appear any differently to the human scorers from the AES typed essays of the treatment groups. Hard

copies (i.e., typed) of all the essays were prepared according to the (National Writing Project) NWP instructions for their summer scoring institute, where they scored multiple papers with a variety of writing prompts. This scoring readiness included anonymous coding (i.e., matching for the pre- and post-essays) to identify each paper and sanitizing any location information (i.e., blacking it out).

Treatment Group Data Collection

Automated Essay Scoring System Data Collection. The tracked AES data from both the pre- and post-essay samples included their holistic score, trait feedback errors (i.e., grammar, usage, mechanics and style) and an organization and development structure measure identifying which essay structures exist. In order to provide information to duplicate this research, additional data collected from the treatment group was the total number of AES writing prompts and frequency of submissions for the corresponding school year. The treatment teacher also may have her own procedures for the classroom use of the AES, so available functionality was collected from the AES for the pre- and post-essays.

Student Survey and Interview Data Collection. The survey and interview data from the ninth grade treatment students, toward the end of the Spring, 2007, semester, were recorded to determine the degree of student satisfaction with the use of the AES. The questions provided in Appendix E were based on two previous research studies (Chen & Cheng, 2006; Grimes & Warschauer, 2006).

Control Group Automated Essay Scoring System Data Collection

An AES model class area was set-up to match the AES set-up for the instructional classes. There is not any difference between the Web-based software used for the

instructional class and that used for the model area, except the students did not have any access to the secure model area. The AES teacher options used for the instructional class were duplicated for the AES model area. The AES model area calibration was checked by re-scoring essays that had been already scored by the AES system in the instructional class.

The AES model area results did not exactly match what had been done in the instructional classes, possibly due to a known AES update by the vendor. Therefore, to create the analysis data, the AES model area was used to both re-score the treatment essays and score the control essays. The researcher copied the AES electronic treatment files and submitted them to the AES model area. Though the control students used a word processor, their electronic files were not available, so the researcher electronically re-created the control files to mirror the hard copies, including all errors, and submitted them for AES scoring to the model area.

Teacher Semi-structured Interview Data Collection

The teachers were interviewed twice with semi-structured interviews in order to further describe the essay samples and their collection for both treatment and control groups in order to assist in the replication of this research. This included information on the total number of writing assignments given during the school year.

Developmental Index Data Collection

All pre- and post-essay writing samples were measured using the words per t-unit (W/T) developmental index (Hunt, 1970; Loban, 1976; Nippold et al., 2005; Wolfe-Quintero et al., 1998). A t-unit is an independent clause and all its subordinate clauses and modifiers, which express a complete thought. A t-unit is a little different than a

sentence in that a compound sentence would be measured as two t-units. Words per t-unit (W/T) are calculated by dividing the total number of words by the total number of t-units. Interrater reliability of the researcher's calculation of t-units was verified by a master student's calculations on a randomly selected 20% sample of the pre- and post-essays from the treatment and control groups for each class level (i.e., standard or honors). Appendix E provides advisory guidelines for calculating t-units and clauses (Polio, 1997). The AES system provided the word counts on the pre-and post-test essays. Microsoft® Excel 2003 was then used to calculate W/T.

Data Analysis

Data were entered into the Statistical Product and Software Solutions (SPSS) 15.0 for Windows computer program for statistical analyses between pre- and post-samples of the treatment and control groups. The ninth grade classes included treatment standard (TS), treatment honors (TH), and control standard (CS).

Automated Essay Scoring System Data Analysis

The AES system data was analyzed for both the treatment and control students. The outcome analysis was a 2 (i.e., male, female) X 3 (i.e., TS, TH, and CS) repeated measure ANOVA, for each measurement outcome, which served as a dependent variable (see Table 3). The ANOVAs were used for each of the following AES dependent variables: (a) holistic score, (b) the grammar errors, (c) the usage errors, (d) the mechanics errors, (e) the style errors, and (f) the organization and development structures.

Table 3

Analysis of Variance Comparisons of Outcomes

		Test	
Class Level	Gender	Pre	Post
TS	Female		
	Male		
TH	Female		
	Male		
CS	Female		
	Male		

Note. TS = treatment standard; TH = treatment honors; CS = control standard; Pre = pre-test; Post = post-test.

National Writing Project Holistic Scoring

The NWP scored papers in comparison to anchor papers that demonstrated the values of a six- point scale (Buchanan, Eidman-Aadahl, Friedrich, LeMahieu, & Sterling, 2006). This would be similar to the holistic scoring method described in Chapter One (Wolcott & Legg, 1998). The six-point scale was used so the scores could be compared to those provided by the AES system. The pre- and post-test NWP holistic scores, serving as dependent variables, were then analyzed with an ANOVA just like the AES scores (see Table 3).

Other Scoring

A repeated measures ANOVA was also used to analyze the W/T pre- and post-test scores (see Table 3). The treatment student surveys and interviews were analyzed with descriptive statistics. Some questions were coded according to identified themes in order to report group averages.

Conclusion

It has been shown that 1:1 computer access increases both teachers' classroom technology integration (of computer applications in general) and the impact of those applications on students' learning (Warschauer, 2006). Differences in teachers' computer skills and pedagogy may be partially reduced by selecting teacher's that are (a) from the same school, (b) experienced and trained in using an AES system, and (c) veterans from prior research with an AES system. Educational research supports the fact that in order to improve their writing, students must write more and receive feedback on their writing (Burstein & Marcu, 2003; Nippold et al., 2005; Page, 2003; Pritchard & Honeycutt, 2006). With large class sizes in high schools, an AES system can provide students with more opportunities to write and receive skilled feedback than teachers alone could make available. An AES's summative feedback, a holistic score, has had many research comparisons that significantly correlate the score to human scorers. However, the use of AES as a classroom intervention has received less research attention to date (Warschauer & Ware, 2006).

The focus of this research was on measurable improvement in the proficiency and development of student writing with the use of an AES system as an intervention. Writing proficiency was measured by the AES holistic and NWP holistic. Gender was

included in analyses to see if there was any difference in the use of technology and to identify any writing development rate of change differences. Students' position of writing development in the different class levels was measured by the W/T index. Increases in writing development are more likely to be revealed with the persuasive genre that was used for the samples. In addition, students' perceptions about using an AES system were examined. Teacher interviews will provide the setting of how many writing assignments were done in the classroom and how much teacher help was provided for the test samples.

CHAPTER 4

RESULTS

This study examined student writing in the beginning of the school year and toward the end of the school year to explore the effects of the use of an automated essay scoring system (AES) to assist student competence in the process of writing.

The study was guided by four research questions:

1. Is there a significant difference in the writing proficiency improvement of students who use an AES system in combination with teacher-led writing instruction compared to students who receive only teacher-led writing instruction, with assessment based on holistic scores from human raters and an AES system? Is gender a significant factor in the results?
2. Is there a significant difference in the writing development of students who use an AES system combined with teacher-led instruction compared to students who receive only teacher-led instruction, as measured by words per t-units (W/T)? Is gender a significant factor in the results?
3. Is there a significant difference between pre- and post-test AES trait error feedback categories for those students who use an AES system combined with teacher-led instruction when compared to those students who had only teacher-led instruction? Is gender a significant factor in the results?

4. What was the degree of user satisfaction for the students who used the AES system as measured by a survey and semi-structured interviews. Is gender a significant factor in the results?

The data were first analyzed for differences between the class levels (i.e., treatment standard, treatment honors, and control standard), and then differences between gender (e.g., male standard group versus female standard group, female honors group, and female control group) and class levels. Gender differences were examined only between male and female, not between persons of the same gender from different class levels. Pre- and post-test essay results were examined for the first three questions with tests in the following order: (a) mixed design analysis of variance (ANOVA), with the between-group variables of class levels or class levels and gender and the repeated measure (i.e., pre- and post-test) as the within-group variable, (b) an ANOVA on the pre-test with the between-group variables of class levels or class levels and gender, (c) if the pre-test showed a significant difference between the groups, an analysis of covariance (ANCOVA) on the post-test, with the pre-test as the covariate and class level or gender and class level as the between group factors, and (d) the post hoc analysis, where necessary, was a Tukey or a Least Significant Difference (LSD) in order to determine which groups were significantly different.

In the event that the pre-test was significant, the mixed design ANOVA was no longer the appropriate analysis choice and its results were not reported, but those of the ANCOVA were reported. However, the multi-factor (i.e., gender and class levels) ANCOVA results were not reported because the small sample size makes them inconclusive. The degrees of freedom and the sample size population do not match across

all the tests due to the AES system results. The AES system did evaluate all the pre- and post-tests of the sample population, but it did not provide a score for all those evaluated. The software would provide a message that there were too many errors to evaluate, but it would not specify what the errors were. The pre- and/or post-tests which did not receive a score varied across the test measurements.

The survey results for question four were analyzed in two ways. The first was the question results were coded and analyzed by frequency descriptive statistics for class levels and class levels and genders. Other survey questions were answered on a scale of 1 to 100, with 100 being best. These answers were then averaged according to the pertinent group analysis (i.e., class levels or gender and class levels).

Though the sample sizes of this research are small, the statistical analyses are valid. However, as an exception, multi-factor ANCOVA results were not reported as they were deemed to be inconclusive due to small sample size. Otherwise, an experiment with a small sample size that produces an F that is significant at $p = .05$ can have a stronger effect than a larger sample size that produces the same level of significance (Keppel, 1991). "In view of the fact that power and sample size are positively correlated, we simple cannot use significance level alone as an index of the strength of an experimental effect" (p. 64). If significant differences are not observed in this research, the conclusion can only be that the research design was not sensitive enough to detect them if they did exist.

First the participant teachers, students, and test setting will be described. Next, the results will be presented, organized by research questions. Descriptive and analytical

statistics were calculated through the use of the Statistical Product and Software Solutions (SPSS) computer program, version 15.0 for Windows.

Participants

The participants were 9th graders in an urban high school in a large southwestern school district during the fall and spring semesters of the 2006-2007 school year. The participants had one of two teachers, one for the treatment group and one for the control group. Thirty-four percent of the participants who began the research did not complete it. The treatment group participants were in two levels, standard and honors, while the control group was only standard level, as shown in Table 4. Each treatment group was made up of two standard classes or two honors classes, but the control group was only one standard class. The gender and numbers for each class level are also shown in Table 4.

Table 4

Participants by Gender and Class Levels

Class Level	N Female	N Male	N All
Treatment Standard	10	5	15
Treatment Honors	13	10	23
Control Standard	8	3	11

Test Setting

The treatment students, according to that teacher, only used the AES system with their major writing assignments, but not all major assignments. All the essays, for both treatment and control groups, were first drafted by hand before being entered into a computer. In order to replicate the test setting, it is important to know how many scored writing assignments and persuasive genre (i.e., the test sample genre) assignments were provided to the students. The persuasive genre was chosen because research indicated it would more likely reveal writing development differences (Scott, 1988). According to online AES tracking, the treatment classes submitted seven writing assignments, four of which were persuasive genre. Based on teacher interviews, the treatment classes and control classes had 9 or 10 major assignments. Both treatment and control classes received the same writing prompts (i.e., writing topics) for the major assignments (including the pre- and post-test) and spent the same amount of classroom time on them.

The pre-test, collected during the month of December, 2006, was the second major assignment using the AES system and the participants' first persuasive prompt, according to the AES tracking. The pre-test prompt was based on a classroom literature assignment: "Construct two paragraphs supporting your opinion of whether Odysseus was or was not a hero in the space provided. Make sure you are supporting your opinion with examples from the textbook." The pre-test took a period of 3 to 4 weeks between initial assignment of the topic and last submission of the essay. This period included the reading of the literature. The teachers provided verbal feedback in the classroom and written feedback, after which participants could re-submit their work for a higher grade.

The post-test, collected during May, 2007 was the seventh major assignment using the AES and the fourth persuasive prompt. The post-test prompt was based on personal experience: "Teenagers don't know what true love really feels like. Agree or disagree? Persuade with strong support." Teachers provided verbal classroom feedback over the period of 3 to 4 days spent on this assignment.

Question One

Question one investigated whether there was significant difference in the writing proficiency improvement of students who use an AES system in combination with teacher-led writing instruction compared to students who received only teacher-led writing instruction, with assessment based on the holistic scores of the pre- and post-test essay. The holistic scores from the pre- and post-test essays were provided by the AES software system and National Writing Project (NWP) human raters. The research also investigated whether gender was a significant factor in the results.

Holistic Scores and Class Levels

The final analyses of the AES holistic score post-test and NWP holistic score post-test results for class levels showed mixed results. The AES holistic score was significant, with treatment honors having a higher mean than the control standard group, but the NWP holistic score showing no statistical significance. Preliminary analysis indicated that the AES holistic score pre-test showed significant differences between the treatment standard group, having the higher mean, and the control standard group.

The AES holistic score ANCOVA was conducted with the post-test as the dependent variable, the pre-test as the covariate, and class levels as the factor. The AES holistic scores post-test ANCOVA was statistically significant ($F_{(2, 43)}$ 3.426, $p = .042$),

with the treatment honors level having a higher mean than the control standard group. Because the NWP holistic score did not show any pre-test significance, the mixed design ANOVA, using the independent variable of class levels, was conducted. No statistical significance was found for the NWP holistic score. The class levels means for the AES holistic and the NWP holistic scores are shown in Tables and 5 and 6, respectively. The pre-test ANOVA of the AES holistic score between the three class level groups was statistically significant ($F_{(2, 46)}$ 3.750, $p = .031$), with the Tukey post hoc analysis showing the significant difference of the treatment standard mean higher than the control standard mean. There was no statistical significance for the pre-test AES holistic score analysis between treatment honors and the other two groups.

Table 5

Automated Essay Scoring's Holistic Score by Class Levels

	Treatment Standard			Treatment Honors			Control Standard		
Test	Mean	SD	N	Mean	SD	N	Mean	SD	N
Pre	4.53	1.125	15	3.70	1.396	23	3.38	0.916	8
Post	3.90	0.799	15	4.17	0.887	23	3.13	1.356	8

Table 6

National Writing Project's Holistic Score by Class Levels

	Treatment						Control		
	Standard			Honors			Standard		
Test	Mean	SD	N	Mean	SD	N	Mean	SD	N
Pre	3.03	0.935	15	2.67	1.202	23	2.55	1.369	11
Post	3.00	1.000	15	2.67	0.806	23	2.64	1.002	11

Holistic Scores for Gender and Class Levels

Question one results had no statistical significance for gender and class levels in the mixed design analysis of the AES holistic score. The analysis of the AES holistic score, examining a potential interaction between gender and class levels, was conducted using a mixed design ANOVA, where gender and class levels were the between-group variables and the repeated measures (i.e., pre- and post-test) were the within-group measure. The NWP holistic score by gender and class levels had significant pre-test results, so the post-test analysis of the NWP holistic scores was a multi-factorial post-test ANCOVA. The small size of the sample makes the results from the multiple-factor ANCOVA inconclusive, so they are not reported.

The gender and class levels groups showed no statistical significance in writing quality improvement with or without the use of the AES system. The results were based on the AES holistic mixed design ANOVA ($F_{(2, 40)}$ 1.041, n.s.), in which the gender and class levels served as the between-group variables and the repeated measure (i.e., pre-

and post-test) served as the within-group variable. Tables 7 and 8 display the descriptive gender and class levels means for each group, respectively.

The NWP holistic scores post-test ANCOVA used the pre-test as the covariate and the gender and class levels as the multiple-factors. Too small of a sample size makes the results inconclusive, so the results are not reported. The ANCOVA was conducted for the NWP holistic post-test because the NWP holistic pre-test score ANOVA ($F_{(2, 43)}$ 4.006, $p = .025$) for gender and class levels had a statistically significant interaction. The Fisher Least Significant Difference (LSD) post hoc showed the significant group interactions were between (a) the male treatment standard group, having the higher mean, and the female treatment standard group, (b) the male treatment standard group, having the higher mean, and the female control standard group, and (c) the male control standard group, having the higher mean, and the female control standard group.

Table 7

Automated Essay Scoring's Holistic Score by Gender and Class Levels

		Treatment Standard			Treatment Honors			Control Standard		
Gender	Test	Mean	SD	N	Mean	SD	N	Mean	SD	N
Female	Pre	4.30	1.160	10	3.92	0.760	13	3.17	0.753	6
	Post	4.10	0.876	10	4.38	1.044	13	2.83	0.753	6
Male	Pre	5.00	1.000	5	3.40	1.955	10	4.00	1.414	3
	Post	3.60	0.548	5	4.17	0.887	10	3.13	1.356	3

Table 8

National Writing Project's Holistic Score by Gender and Class Levels

		Treatment Standard			Treatment Honors			Control Standard		
Gender	Test	Mean	SD	N	Mean	SD	N	Mean	SD	N
Female	Pre	2.65	0.747	10	3.00	1.225	13	2.25	0.886	8
	Post	3.10	1.101	10	2.92	0.886	13	2.38	0.694	8
Male	Pre	3.80	0.837	5	2.25	1.087	10	3.33	2.309	3
	Post	2.80	0.837	5	2.35	0.580	10	3.33	1.528	3

Question Two

Question two investigated the difference in the writing development of students who use an AES system combined with teacher-led writing instruction compared to students who receive only teacher-led writing instruction, as measured by words per t-units (t-units) in the pre- and post-essays. It was also investigated whether gender was a significant factor in the results.

Words per T-unit and Class Levels

Question two showed no class levels significant differences for the post-test words per t-unit (W/T) results for any of the groups, with or without treatment. The pre-test results did show class level differences, at the beginning of the research period, between the following groups: treatment standard, with the higher men, and control

standard and treatment standard, with the higher mean, and treatment honors. This pre-test difference necessitated the use of the post-test ANCOVA.

The post-test ANCOVA was conducted with the pre-test as the covariate and class levels as the between-group measure. The W/T results showed no statistically significant interaction between the groups ($F_{(2, 46)}$.053, n.s.). The class levels means are shown in Table 9. A class levels ANOVA for the pre- test W/T scores was statistically significant ($F_{(2, 46)}$ 3.637, $p = .034$), with the LSD post hoc revealing the significance between the treatment standard and control standard groups and also between the treatment standard and treatment honors groups.

Table 9

Words per T-unit by Class Levels

	Treatment Standard			Treatment Honors			Control Standard		
Test	Mean	SD	N	Mean	SD	N	Mean	SD	N
Pre	16.945	3.5659	15	14.152	3.220	23	13.552	4.451	11
Post	14.569	2.2963	15	14.311	4.780	23	13.847	2.315	11

Words per T-unit for Gender and Class Levels

Since the pre-test W/T showed a significant difference between the gender and class level groups at the beginning of the research period, an ANCOVA was conducted on the post-test. The small size of the sample makes the results from the multiple-factor

ANCOVA inconclusive. The significant differences in the pre-test ANOVA were between the following groups: (a) the female treatment standard group and the male treatment honors group, (b) the male treatment standard group and the female control honors group, (c) the male control standard group and the female treatment honors group, and (d) the male control standard group and the female control standard group.

Potential interaction between gender and class levels were examined with a post-test ANCOVA, where the pre-test was the covariate and gender and class levels were the multiple independent factors. The small size of the sample makes the results from the multiple-factor ANCOVA inconclusive, so they are not reported. The gender and class levels means are shown in Table 10. The W/T pre-test ANOVA ($F_{(2, 43)}$ 6.696, $p = .003$) had a statistically significant interaction for gender and class levels. The post hoc LSD revealed the significant gender and class levels interactions were between (a) the female treatment standard group, having a higher mean, and the male treatment honors group, (b) the male treatment standard group, having the higher mean, and the female control standard group, (c) the male control standard group, having the higher mean, and the female treatment honors group, and (d) the male control standard group, having the higher mean, and the female control standard group.

Table 10

Words per T-unit by Gender and Class Levels

		Treatment			Treatment			Control		
		Standard			Honors			Standard		
Gender	Test	Mean	SD	N	Mean	SD	N	Mean	SD	N
Female	Pre	17.29	3.955	10	14.26	2.203	13	11.28	1.676	8
	Post	15.20	2.542	10	14.65	5.168	13	12.80	1.656	8
Male	Pre	16.26	2.903	5	14.02	4.339	10	19.62	3.641	3
	Post	13.30	.951	5	13.87	4.508	10	16.65	1.019	3

Question Three

Question three investigated if there was a significant difference between pre- and post-test AES trait error feedback categories for those students who use an AES system combined with teacher-led instruction when compared to those students who had only teacher-led instruction. It also investigated if gender was a significant factor in the results. The error categories consist of errors in (a) grammar, (b) usage, (c) mechanics, and (d) style. The errors were expected to decrease with improvement in writing. The organization and development structure in the trait category was addressed separately because it evaluated the existence of various essay parts, therefore increasing with improvement.

Grammar Errors

Grammar errors are made up of ten sub-categories (see Appendix A). Grammar errors analysis for the post-test results between the class levels did not show significant differences between the groups. The post-test ANCOVA was used because pre-test grammar errors for class levels showed significant differences between the treatment standard and control standard groups. The mixed design ANOVA by gender and class levels did not show any significant difference.

There were no statistically significant group interactions for the post-test ANCOVA with the pre-test as the covariate and class levels as the between group measure for the grammar errors category ($F_{(2, 45)}$ .719, n.s.). The grammar errors pre-test ANOVA by class level showed significance ($F_{(2, 46)}$ 6.281, $p = .004$), which was the reason for the post-test ANCOVA instead of the mixed design ANOVA. The Tukey post hoc analysis identified that the grammar errors pre-test significance was between the treatment standard group, having the higher mean, and the control standard group and the treatment standard group, having the higher mean, and the treatment honors group. The pre-test and post-test class levels means are shown in Table 11. The grammar errors mixed design ANOVA by gender and class levels ($F_{(2, 42)}$ .564, n.s.) did not show any significance. Table 12 displays the gender and class levels means.

Table 11

Automated Essay Scoring Grammar Errors by Class Levels

	Treatment			Treatment			Control		
	Standard			Honors			Standard		
Test	Mean	SD	N	Mean	SD	N	Mean	SD	N
Pre	6.67	4.451	15	3.39	2.840	23	2.60	2.221	10
Post	3.33	2.664	15	3.74	2.767	23	2.70	1.636	10

Table 12

Automated Essay Scoring's Errors by Gender and Class Levels

		Treatment			Treatment			Control		
		Standard			Honors			Standard		
Gender	Test	Mean	SD	N	Mean	SD	N	Mean	SD	N
Female	Pre	7.70	3.773	10	4.23	3.032	13	3.00	2.309	7
	Post	3.50	2.718	10	4.38	2.959	13	2.14	1.574	7
Male	Pre	4.60	5.413	5	2.30	2.263	10	1.67	2.082	3
	Post	3.00	2.828	5	2.90	2.378	10	4.00	1.000	3

Usage Errors

Usage errors are comprised of seven sub-categories (see Appendix A). The usage errors post-test analysis results of class levels did not show any significant differences for any of the groups, with or without treatment. The usage errors pre-test did reveal a significant difference between the treatment standard group, having the higher mean, and the treatment honors group, and the treatment standard group, having the higher mean, and the control standard group. The pre-test class level differences were no longer evident by the end of the research period since the post-test class levels analysis did not have any significant differences. The usage errors mixed design analysis results for usage gender and class levels did not show any significant differences for any of the groups, with or without treatment.

The post-test ANCOVA analysis used the pre-test as the covariate and class levels as the between-group factor of the usage errors and showed no statistical significance ($F_{(2, 44)}$ ..152, n.s.). Table 13 shows the usage error means for the class levels. It was because the pre-test ANOVA of class level for the usage errors was statistically significant ($F_{(2, 45)}$ 8.569, $p = .001$) that the ANCOVA was used for the post-test. The Tukey post hoc on the class level pre-test analysis showed interaction between the treatment standard, having the higher mean, and treatment honors groups and between the treatment standard, having the higher mean, and control standard groups. The mixed design ANOVA of the usage errors by gender and class levels was not statistically significant ($F_{(2, 42)}$ .564, n.s.). Table 14 shows the usage error means for gender and class levels.

Table 13

Automated Essay Scoring Usage Errors by Class Levels

	Treatment			Treatment			Control		
	Standard			Honors			Standard		
Test	Mean	SD	N	Mean	SD	N	Mean	SD	N
Pre	4.13	3.226	15	1.17	1.749	23	1.44	1.130	9
Post	3.27	3.535	15	1.91	3.088	23	1.44	1.014	9

Table 14

Automated Essay Scoring's Usage Errors by Gender and Class Levels

		Treatment			Treatment			Control		
		Standard			Honors			Standard		
Gender	Test	Mean	SD	N	Mean	SD	N	Mean	SD	N
Female	Pre	4.80	3.458	10	1.62	2.063	13	1.33	1.211	6
	Post	4.00	4.055	10	2.38	3.709	13	1.50	1.225	6
Male	Pre	2.80	2.490	5	0.60	1.075	10	1.67	1.155	3
	Post	1.80	1.643	5	1.30	2.058	10	1.33	0.577	3

Mechanics Errors

The mechanics errors are composed of eleven sub-categories (see Appendix A). The mechanics errors mixed design ANOVA results of class levels and gender and class levels did not show any significant differences, with or without treatment. The mixed design ANOVA of mechanics errors ($F_{(2, 45)}$ .304, n.s.), with class levels as the between-group variable and repeated measures (i.e., pre- and post-test) as the within group variable, showed no statistical significance. The class levels means are displayed in Table 15. The mixed design ANOVA of the mechanics errors (F (2, 42) .102, n.s.), with gender and class levels as the between-groups variable and repeated measures (i.e., pre- and post-test) as the within group variable, was also not statistically significant. Table 16 shows the gender and class levels means.

Table 15

Automated Essay Scoring's Mechanics Errors by Class Levels

	Treatment			Treatment			Control		
	Standard			Honors			Standard		
Test	Mean	SD	N	Mean	SD	N	Mean	SD	N
Pre	2.20	2.396	15	1.48	3.013	23	0.60	0.843	10
Post	1.67	2.193	15	1.70	3.535	23	0.50	0.527	10

Table 16

Automated Essay Scoring's Mechanics Errors by Gender and Class Levels

		Treatment Standard			Treatment Honors			Control Standard		
Gender	Test	Mean	SD	N	Mean	SD	N	Mean	SD	N
Female	Pre	0.50	0.527	10	5.62	8.559	13	3.71	3.773	7
	Post	2.30	2.791	10	2.92	4.112	13	0.71	0.951	7
Male	Pre	4.00	2.236	5	3.80	3.967	10	1.67	0.577	3
	Post	4.60	5.814	5	8.40	22.401	10	1.67	0.577	3

Style Errors

The style errors are made up of nine sub-categories (see Appendix A). The style errors mixed design ANOVA results of class levels and gender and class levels did not show any significant differences, with or without treatment. Neither of the style errors mixed design ANOVAs, with the between-groups variable of class levels ($F_{(2,45)}$ .683, n.s.) or gender and class levels ($F_{(2,46)}$ .824, n.s.) and the within group variable of repeated measures (i.e., pre- and post-test), showed statistical significance. Tables 17 and 18, respectively, show the class level means and the gender and class levels means.

Table 17

Automated Essay Scoring's Style Errors by Class Levels

	Treatment			Treatment			Control		
	Standard			Honors			Standard		
Test	Mean	SD	N	Mean	SD	N	Mean	SD	N
Pre	31.20	17.358	15	29.39	16.997	23	32.90	12.206	10
Post	38.73	22.864	15	39.52	23.833	23	31.10	9.689	10

Table 18

Automated Essay Scoring's Style Errors by Gender and Class Levels

		Treatment			Treatment			Control		
		Standard			Honors			Standard		
Gender	Test	Mean	SD	N	Mean	SD	N	Mean	SD	N
Female	Pre	29.40	12.851	10	33.54	17.101	13	35.14	12.615	7
	Post	43.10	25.701	10	42.00	27.009	13	31.43	10.998	7
Male	Pre	34.80	25.665	5	24.00	16.097	10	27.67	11.590	3
	Post	30.00	14.160	5	36.30	19.883	10	30.33	7.638	3

Organization and Development

The organization and development evaluation provided by the AES was analyzed by counting the AES's stated existence of the essay's organization and development structure parts (see Appendix A). In addition, any stated error in the thesis statement was counted as a negative (i.e., -1). The organization and development post-test ANCOVA results for class levels did not show any significant differences for any of the groups, with or without treatment. The organization and development structure analysis indicated that the pre-test showed significant differences between the treatment standard and control standard groups and the treatment honors and control standard groups. This is the reason that ANCOVA was used for the post-test analysis. The organization and development mixed design analysis by gender and class levels did not indicate any significance.

The post-test ANCOVA for development and organization structure ($F_{(2, 45)}$ .191, n.s.) was not statistically significant, whereby the covariate was the pre-test and the between-group measure was the class levels. The class levels means are shown in Table 19. The class levels ANOVA of the pre-test score was statistically significant ($F_{(2, 45)}$ 4.372, $p = .018$), with the post hoc Tukey indicating significance with the treatment standard group having a higher mean than the control standard group and the treatment honors having a higher mean than the control standard group. This pre-test significance was the reason for using the ANCOVA on the post-test.

The organization and development mixed design ANOVA, whereby the between-group measures were gender and class levels and the within-group measure was the repeated measure (i.e., pre- and post-test score), did not show any statistical significance. Table 20 shows the gender and class levels means.

Table 19

Automated Essay Scoring's Development and Organization by Class Levels

	Treatment			Treatment			Control		
	Standard			Honors			Standard		
Test	Mean	SD	N	Mean	SD	N	Mean	SD	N
Pre	6.33	1.397	15	4.78	2.467	23	4.10	1.449	10
Post	4.53	1.552	15	4.87	1.660	23	2.70	2.710	10

Table 20

Automated Essay Scoring's Development and Organization by Gender and Class Levels

		Treatment			Treatment			Control		
		Standard			Honors			Standard		
Gender	Test	Mean	SD	N	Mean	SD	N	Mean	SD	N
Female	Pre	6.40	1.350	10	5.77	2.127	13	4.14	1.464	7
	Post	4.60	1.647	10	5.00	2.041	13	2.00	1.000	7
Male	Pre	6.20	1.643	5	3.50	2.369	10	4.00	1.732	3
	Post	4.40	1.517	5	4.70	1.059	10	4.33	4.933	3

Question Four

Question four investigated the level of user satisfaction for the students (n = 36) who used the AES system. The significance of gender on the results was also investigated. The response rate for the standard level class was 100% and 91% for the honors class level. The results are reported for the received responses and rounding errors result in some totals not equaling 100%. The participant response rate was 100% for female standard, 100% for male standard, 92% for female honors, and 90% for male honors. The survey questions were analyzed by frequency according to the categories of (a) participant experiences and self-perceptions, (b) participants' AES preferences, (c) AES's usability, writing improvement, and effectiveness, and (d) frequency of AES's use.

Participant Experiences and Perceptions

The survey questions in this section, item numbers 21 through 25, 27, and 28, described the participants': (a) school computer experience, (b) home computer access, (c) preference for writing with a computer, (d) self-perceptions on their writing quality, and (e) language(s) spoken at home. All participants, except for one of the male honors treatment participants, had taken the school district's required computer class. The participants in the standard group averaged 4.37 years of classroom computer experience while the honors group averaged 5.14 years. From the highest to the lowest, by gender and class levels, the participants' years of school computer experience follow: (a) male honors averaged 6.78 years, (b) female standard averaged 5.2 years, (c) female honors averaged 3.92 years, and the male standard averaged 2.7 years.

Table 21 contains the percentages of participants' having home computers and Internet connections. Participants' writing modality preferences were (a) computer, (b) hand, or (c) both, also shown in Table 21. The percentages of participants' self-perceptions of writing quality, being a good writer, are shown by in the same table.

Table 21

Treatment Participant's Home Computer Access, Modal Writing Preference, and Perceived Writing Ability by Class Levels and Gender

	Standard			Honors		
Categories	Female	Male	All	Female	Male	All
Home computer	100.0	100.0	100.0	84.6	77.8	85.7
Home Internet	80.0	100.0	86.7	75.0	55.6	66.7
Preferring writing by computer	80.0	100.0	87.6	75.5	100.0	85.7
Preferring writing by hand	20.0	0.0	13.3	16.7	0.0	9.5
Preferring both for writing	0.0	0.0	0.0	8.3	0.0	4.8
Think self a good writer	60.0	60.0	60.0	91.7	100.0	95.2

Note. Numbers are percentages.

Participants' responses about language spoken at home were divided into three categories: (a) Spanish or other languages, (b) bilingual, and (c) English. For the category

of Spanish or other languages spoken at home, the standard group included Tagalog (i.e., Filipino), in addition to Spanish. For the bilingual category, the standard group included Spanish and Tagalog (i.e., Filipino) and the honors group included Spanish, Tagalog (i.e., Filipino), and Molikese (i.e., Polynesian). Table 22 shows the percentage results.

Table 22

Treatment Participant's Percentage of Languages Spoken at Home by Class Levels and Gender

	Standard			Honors		
Language	Female	Male	All	Female	Male	All
Spanish or other	40.0	20.0	33.4	50.0	33.3	42.9
Bilingual	10.0	20.0	13.4	23.1	33.3	26.6
English	50.0	60.0	53.3	25.0	33.3	26.1

Note. Numbers are percentages.

Preferences for the Automated Essay Scoring System

The survey questions in this section, item numbers 26, 19, 20, and 30, covered general participant preferences for the AES: (a) what they thought about writing feedback from a computer, (b) the best thing about the AES, (c) the worst thing about the AES and (d) how they responded to the holistic score. Responses about participants' perceptions of receiving feedback on their writing from a computer were (a) positive, (b) negative, or (c) neutral. Table 23 shows the results.

There were four responses given for the most help in improving participant's writing during the research period: (a) the teacher, (b) the AES, (a) writing more, and (d) other opinions. Table 23 also shows these percentage results. For the survey question requesting participant's responses to the holistic score, multiple answers resulted in numbers greater than 100%. The five responses were as follows: (a) liked the score, (b) made them want or need to improve, (c) do not understand the holistic score, (d) do not like the score, and finally, (e) no opinion. Those who did not like the AES holistic score still felt it made them want to improve. However, liking the score did not mean the participants said it was motivational. Table 23 provides the percentage results.

Table 23

Treatment Participants' Preferences for Computer Feedback, Most Help to Writing and Holistic Score Perception by Class Levels and Gender

	Standard			Honors		
Categories related to writing	Female	Male	All	Female	Male	All
Computer feedback positive	70.0	80.0	73.3	66.7	44.4	57.1
Computer feedback negative	20.0	20.0	13.3	16.7	22.2	19.0
Computer feedback neutral	10.0	0.0	13.3	16.7	33.3	27.0
Teacher most help	40.0	60.0	46.7	45.9	75.0	57.5
AES was most help	30.0	40.0	33.3	33.3	12.5	25.0
Writing more most help	30.0	0.0	20.0	11.9	12.5	12.5
Others most help	0.0	0.0	0.0	0.0	12.5	5.0
Liked holistic score	100.0	100.0	100.0	75.0	77.8	76.2
Holistic score made them want or need to improve	80.0	100.0	86.7	66.6	77.8	71.4
Did not like holistic score	0.0	0.0	0.0	8.3	11.1	9.5
Neutral to holistic score	0.0	0.0	0.0	16.7	11.1	14.3
Did not understand holistic score	0.0	0.0	0.0	8.3	0.0	4.8

Note. Numbers are Percentages.

Five categories were included in the comments about the best thing about the AES: (a) error feedback, (b) having the help the AES provided, (c) using the computer for writing, (d) using spell check, and (e) revising more than would be done otherwise. Table 24 provides the result percentages.

Six categories of comments about the worst thing about the AES included: (a) nothing, (b) not understanding specific trait error messages or categories, (c) holistic score, (d) students losing work from not saving or system crash, (e) inaccurate results, and (f) problems with spell check. The percentage results are also shown in Table 24.

Table 24

Treatment Participants Regarding the Best and Worst Things about the Automated Essay Scoring by Class Levels and Gender

		Standard			Honors		
Rating	Category	Female	Male	All	Female	Male	All
Best	Error feedback	40.0	40.0	40.0	75.0	22.2	52.4
	AES online help	30.0	20.0	26.7	8.3	22.2	14.3
	Writing on computer	20.0	20.0	20.0	0.0	33.3	14.3
	Spell check	10.0	0.0	6.7	8.3	22.2	14.3
	Revising more	0.0	20.0	6.7	8.3	0.0	4.8
Worst	Nothing	50.0	60.0	53.3	25.0	55.6	38.1
	Not understanding feedback message(s)	10.0	20.0	13.3	25.0	22.2	23.8
	Holistic score	10.0	0.0	8.7	33.3	11.1	23.8
	Losing work	10.0	20.0	13.3	8.3	0.0	4.8
	Inaccurate results	10.0	0.0	6.7	8.3	0.0	4.8
	Spell check	10.0	0.0	6.7	0.0	11.1	4.8

Note. Numbers are percentages.

Usability, Improvement from, and Effectiveness of the Automated Essay Scoring System

The survey questions in this section addressed various aspects of the use of the *AES*: (a) usability, (b) writing improvement, and (c) effectiveness. The usability questions, item numbers 1 through 4 and 10, covered how easy the participants thought that the AES was to use. Writing improvement questions, including item numbers 5 though 9, had participants identify if they thought the AES helped improve their writing. The effectiveness portion of the survey, encompassing question items 11 through 18, questioned whether the participants thought the functionality sections of the AES were effectively helpful. Table 25 gives the percentage results.

Table 25

Treatment Participants' Usability, Writing Improvement, and Effectiveness of the Automated Essay Scoring System by Class Levels and Gender

	Standard			Honors			Both
AES Category	Female	Male	All	Female	Male	All	
Usability	81.3	77.0	79.8	80.0	77.2	78.8	78.8
Writing improvement	85.3	80.2	83.6	83.1	82.1	82.7	83.1
Effectiveness	82.9	80.0	82.2	84.8	74.1	80.2	81.0

Note. Numbers are percentages.

Frequency of Automated Essay Scoring System's Use

The AES tracked how many times and which essays a participant submitted for review. Overall, the average percent of treatment participants submitting their pre- and post-test essays multiple times was 65% for an average of 2.6 times. The results are displayed in Tables 26 and 27 respectively.

Table 26

Treatment Participants' Multiple Automated Essay Scoring Submissions by Class Levels and Gender

	Standard			Honors			Both
Category	Female	Male	All	Female	Male	All	
Submitting multiple pre-tests	60.0	50.0	53.3	60.0	92.3	78.3	68.4
Submitting multiple post-tests	70.0	60.0	66.7	46.2	70.0	56.5	60.5

Note. Numbers are percentages.

Table 27

Treatment Participant's Test Submissions by Class Levels and Gender

	Standard			Honors			Both
Averages Category	Female	Male	All	Female	Male	All	
Pre-test submissions	1.6	2.6	1.9	2.7	2.5	2.6	2.3
Post-test submissions	3.1	2.4	2.9	2.9	2.7	2.8	2.8

Note. Numbers are averages.

Summary

Question One's research on the writing proficiency rate of change for participants using an AES system and having teacher-led instruction compared to participants only having teacher led instruction showed mixed results. Question One also investigated whether gender was a factor in the results. The analysis results for the AES holistic score by class level showed post-test significance between the treatment honors and control standard groups, but the mixed design analysis for NWP holistic score did not show any significance. The AES holistic score post-test significance by class levels was between the honors treatment group, which had the higher mean, and the control standard group. The AES holistic pre-test significance by class levels was between the standard treatment, which had a higher mean, and the standard control group. This pre-test significance necessitated the use of the post-test ANCOVA instead of the mixed design ANOVA.

The AES holistic mixed design analysis for gender and class levels was not significant. The NWP holistic post-test analysis for gender and class level was not reported due to the small sample size. The NWP holistic score pre-test significance necessitated the use of the post-test ANCOVA. The NWP holistic pre-test by gender and class levels showed significance between (a) male treatment standard group, which had the higher mean and the female treatment standard group, (b) the male treatment standard group, which had the higher mean, and the female control standard group, and (c) the male control standard group, which had the higher mean, and the female control standard group.

Question Two's research on the writing maturity rate of change for participants using an AES system and having teacher-led instruction compared to participants only having teacher led instruction did not show class level significance for W/T. Question Two also investigated whether gender was a factor in the results. The pre-test class level differences were between the treatment standard, having the higher mean, and control standard groups and the treatment standard, having the higher mean, and treatment honors groups. The gender and class levels post-test on W/T was not reported due to small sample size. The post-test analysis was conducted because the W/T pre-test analysis for gender and class levels did show significance. The gender and class level differences were between (a) female treatment standard, having the higher mean, and male treatment honors groups, (b) male treatment standard, having the higher mean, and female control standard groups, (c) male control standard, having the higher mean, and female treatment honors groups, and (d) male control standard, having the higher mean, and female control standard groups.

Question Three's research was on the rate of change in AES writing trait error scores and organization and development structure for participants using an AES system and having teacher-led instruction compared to participants only having teacher led instruction. Question Three also investigated whether gender was a factor in the results. Post-test analysis of (a) grammar errors, (b) usage errors, and (c) organization and development structure by class levels were not significant. Neither was the class level mixed design analysis of usage and mechanics. The post-test analysis was necessitated by pre-test significance in the different measures. The pre-test grammar errors for the class levels did show a significant difference between treatment standard, having the higher mean, and treatment honors groups and treatment standard, having the higher mean, and control standard groups. The pre-test usage errors for the class levels did show a significant difference between the treatment standard, which had the higher mean, and the treatment honors groups and the treatment standard, which had the higher mean, and the control standard groups. The pre-test analysis of organization and development did show a significant difference between the treatment standard, which had the higher mean, and the control standard groups and the treatment standard, which had the higher mean, and the treatment honors groups. These pre-test differences between the class levels in grammar errors, usage errors or organization and development structure were no longer evident by the end of the research because significance was no longer evident in the post-test analysis.

Question Four's research on treatment participant's degree of user satisfaction with the AES system defined various characteristics of the treatment participants and their preferences for the AES system. Question Four also investigated whether gender

was a factor in the results. Almost all the participants had taken the school district required computer class. The standard participants were more likely to have a home computer and internet access than were honors participants. More honors than standard participants were likely to speak a language other than English or to be bilingual. Despite any shortcomings noted by the participant's preferences, they overwhelmingly liked the AES system.

CHAPTER 5

DISCUSSION OF FINDINGS

This study investigated multiple measurement differences in pre- and post-essay samples for students who used an automated essay scoring (AES) system plus teacher-led instruction for approximately 6 months during an academic school year compared to students who only received teacher-led instruction. The discussion in the following section relates the findings in this study to professional literature on technology and writing in the classroom. That discussion is followed by the limitations of the findings and implications and recommendations for future research.

Discussion of Research

The data of the first three questions were analyzed with either a mixed design analysis of variance (ANOVA) for pre- and post-test differences between the class levels (e.g., treatment standard, treatment honors, and control standard) or a post-test analysis of covariance (ANCOVA) with the pre-test as the covariate and the class levels as the factor. A significant pre-test required the use of the ANCOVA in order to account for the individual differences found in the pre-test. In addition, the pre- and post-test differences between gender and class levels (e.g., male treatment standard group compared to female treatment standard group, female treatment honors group, and female control standard

group) were analyzed with a mixed design ANOVA or a post-test ANCOVA with the pre-test serving as the covariate and the gender and class levels serving as the multiple-factors. The gender and class level multi-factor ANCOVA results were not reported because the small sample size made the results inconclusive. Again, a significant pre-test difference necessitated the use of a post-test ANCOVA to account for the pre-test differences. Gender differences were examined only between male and female, not between persons of the same gender from different class levels.

If the pre- and post-test analysis (i.e., mixed design) ANOVA was not significant, further analyses were conducted, but only reported if there was significance. The pre-test ANOVAs independent variable was either class levels or gender and class levels. In the event that significance was found, a post hoc Tukey or Least Significant Differences (LSD) was performed to determine which groups had the significant difference. The discussion of research findings will be in the order of the four research questions.

Question One

Question One asked: Is there a significant difference in the writing proficiency improvement of students who use an AES system in combination with teacher-led writing instruction compared to students who receive only teacher-led writing instruction, with assessment based on holistic scores from human raters and an AES system? Is gender a significant factor in the results?

The post-test analysis of AES holistic scores by class levels indicated significant differences for the writing quality rate of change between the treatment honors and the control standard groups, with the treatment honors mean being higher. In contrast, the National Writing Project (NWP) holistic scores showed no significance for the mixed

design ANOVA for class levels. The pre-test analysis of the AES score for class levels did show significance, which was why an ACOVA was used for analysis. The AES holistic score pre-test significance was between the treatment standard and the control standard groups, with the treatment standard group mean being higher. No class level significance was shown in the pre-test for the treatment honors group with either of the other two group.

The pre-test to post-test AES holistic means only increased for the treatment honors group, which was apparent by their significance in the post-test analysis. It is unclear why the post-test AES holistic means decreased for the treatment standard and control standard groups. There may have been a prompt effect (P. LaMahieu, personal communication, January 19, 2007). Though both prompts were persuasive genres, the pre-test was based on literature and the post-test was based on student's personal experience. In addition, the assignments were not equal in length, with the pre-test assignment being about 3 weeks in length, while the post-test assignment was less than 1 week in length. The pre-test assignment was longer, in part, because it was the student's first introduction to the persuasive genre with teacher-led instruction.

The treatment honors group improved in proficiency based on the ANCOVA on the AES post-test holistic score, but there was no significance from the NWP holistic score analysis to corroborate the AES outcome. The disparate results from the analyses for the two proficiency outcomes from the current research seems to support prior research on the use of 16 different educational technology products that showed no improvement in student outcomes with their use (Dynarski et al., 2007). The disparate outcomes also seems to support prior research that showed no significant improvement in

student's writing proficiency with the use of AES systems (Grimes & Warschauer, 2006). The conclusion in the Grimes & Warschauer (2006) research was based on standardized test results that were independent of the AES system scoring, similar to the human scored NWP holistic score used in this research in that it was also independent of the AES system.

The AES holistic score's mixed design analysis by gender and class levels did not show any significance, so the AES holistic scoring did not show any gender effect. The post-test analysis of the NWP holistic score (i.e., a human score) by gender and class levels was not reported due to the small sample size making the analysis results inconclusive. The post-test analysis was used because the NWP holistic had a significant pre-test. The NWP holistic pre-test score significance by gender and class levels revealed the significant differences between the following groups: (a) the male treatment standard, having the higher mean, and the female treatment standard group, (b) the male treatment, having the higher mean, and the female control standard groups, (c) the male control standard, having the higher mean, and the female control standard groups. No NWP holistic score pre-test significance by gender and class levels was shown for the male honors or the female honors groups. While previous research has shown that gender was no longer an issue with computer access in an educational setting, the AES proficiency scoring in this research also shows there are no significant gender differences for outcomes with the use of computers in an educational setting (Day et al., 2003; Parsad & Jones, 2005).

The speed of the scoring feedback is a feature that has been considered a strong positive of AES systems (Boone & Frost, 2005; Chen & Cheng, 2006; Educational

Testing Service, 2007a; Grimes & Warschauer, 2006; Waxman et al., 2003). However, it is unclear whether the AES system was beneficial for proficiency improvement. The treatment honors group improved their AES holistic score mean, but the other class level groups did not, and there was no corroboration of the AES holistic scoring by the NWP holistic scoring. The majority of both of the groups surveyed said they were motivated by the AES holistic score to improve their writing, but the current research outcomes cannot support the student opinions.

All the groups could be considered to have used word processing, in that the control group used Microsoft® Word and the treatment groups used a text editor that is part of the AES system. Based on previous research, using word processing should increase writing proficiency, but current research did not support that outcome for either the treatment standard or the control standard groups since their post-test AES-scored holistic means decreased from the pre-test (Bangert-Drowns, 1993; Graham & Perin, 2006).

In summary, the use of an AES system plus teacher-led instruction showed post-tests significantly higher for the AES holistic scores for only the treatment honors group, when compared to the use of a word processor and teacher-led instruction of the control standard group. The lack of significance for the NWP holistic scores does not provide the data to support a proficiency improvement for participants with the use of the AES plus teacher-led instruction. Gender and class levels also had no AES-scored holistic score significance, so no technology gender benefit was evident.

Question Two

Question Two asked: Is there a significant difference in the writing development of students who use an AES system combined with teacher-led instruction compared to students who receive only teacher-led instruction, as measured by words per t- unit (W/T)? Is gender a significant factor in the results?

Words per t-unit is a writing development measure that is not under the conscious control of the writer (Hunt, 1970; Loban, 1976; Nippold et al., 2005; Wolfe-Quintero et al., 1998). A minimal terminal unit (i.e., t-unit) is defined as one dependent clause plus all associated dependent clauses (Hunt, 1965a, 1965b; Nippold et al., 2005; Polio, 1997; Wolfe-Quintero et al., 1998). It is somewhat different than a sentence because a compound sentence would equal two t-units.

The three class levels showed no significant difference over the research period for the post-test ANCOVA on words per t-unit. The post-test ANCOVA on words per t-unit between the gender and class levels is not reported because the small sample size makes the test results inconclusive. The post-test ANCOVA was used for class levels and gender and class levels because the pre-test analyses were significant. The pre-test analysis for class levels showed significance between the treatment standard and control standard groups and the treatment standard and treatment honors groups, with the treatment standard mean being higher in both instances. The W/T pre-test significance may only indicate writing development differences with the class levels at the beginning of the research period, because the differences were no longer evident at the end of the period. The gender and class levels words per t-units pre-test analysis also showed significance for W/T with (a) the male control standard mean significantly higher than

the female control standard mean, (b) the male control standard mean significantly higher than the female treatment honors mean, (c) the male treatment standard mean significantly higher than the female control standard mean, and (d) the female treatment standard mean significantly higher than the male treatment honors mean.

In conclusion, the post-test W/T results analysis were required because both class levels and gender and class levels had significant pre-tests. There was no significant writing development rate of change for W/T for either the treatment or control groups by class levels. The significance that was evident on the W/T pre-test for the class levels was no longer evident by the post-test. No results were reported for the W/T post-test analysis by gender and class levels due to the small sample size making the results inconclusive.

Question Three

Question Three asked: Is there a significant difference between pre- and post-test AES trait error feedback categories for those students who use an AES system combined with teacher-led instruction when compared to those students who had only teacher-led instruction? Is gender a significant factor in the results? The feedback results were investigated for the AES's individual error categories of (a) grammar, (b) usage, (c) mechanics, and (d) style errors. The organization and development category was analyzed separately from the error categories because the means are expected to increase, while the error means are expected to decrease.

The AES post-test analysis for class levels, used due to pre-test significance, showed no significant differences for the error categories of (a) grammar, (b) usage and the category of (c) organization and development structure. Neither did the mixed design AES error analysis by class levels show any significant differences for the categories of

(a) usage, (b) mechanics, and (c) style. No significance was shown for gender and class levels for the mixed design analysis of any of the AES error categories or the organization and development structure category.

The class level pre-test analysis showed significance for the following: (a) grammar errors with the treatment standard mean higher than control standard and the treatment standard mean higher than treatment honors, (b) usage errors with the treatment standard mean higher than treatment honors and the treatment standard mean higher than control standard, and (c) organization and development structure with the treatment standard mean higher than control standard and the treatment honors mean higher than control standard. None of the pre-test significance by class levels was evident by the post-test, so it may have indicated class level differences at the beginning of the research period.

The AES system can also be considered computer assisted instruction (CAI). The current research does not support previous research that provided evidence that CAI student outcomes improved from the 50.0 percentile to the 57.2 percentile since no significance in the rate of change was shown for any of the error categories or the organization and development category (Christmann et al., 1997). These results also did not show any benefit of the immediate feedback of the AES system (Boone & Frost, 2005; Chen & Cheng, 2006; Educational Testing Service, 2007a; Grimes & Warschauer, 2006; Waxman et al., 2003) since there was no significance between the treatment or control groups.

This study measured the pre- and post-essays of two different topics, unlike previous research. However, the overall lack of significance for all the error

measurements and the organization and development structure did not seem to support the previous research finding that students using the AES corrected about 25% of the trait errors between the pre- and post-essays of one topic (Attali, 2004). Since the majority of AES feedback is formative, it was expected that students' revision would focus on formative errors rather than organizational or development content. More formative corrections would also support prior research (Yagelski, 1995). However, none of the AES trait error scores or organization and development structure showed any significance in the rate of change for treatment or control groups.

Question Four

Question Four asked: What was the degree of user satisfaction for the students who used the AES system as measured by a survey and semi-structured interviews? Is gender a significant factor in the results? The survey questions were analyzed by one or two methods. The answers were coded and analyzed by frequency descriptive statistics for class levels and gender and class levels. Other survey questions were answered on a scale of 1 to 100, with 100 being best. The answers were then averaged according to the relevant group (i.e., class levels or gender and class levels) analysis. Due to rounding, the percentages may not equal 100%.

Participant Descriptions and Perceptions

Treatment participants' demographics about (a) school computer experience, (b) home computer access, (c) home Internet access, (d) modal writing preference, (e) self-perceived writing quality, and (f) the language spoken at home may help explain their perceptions about using the AES system in the classroom.

School computer training and home computer access. The results on school computer training and home computer access were not matched for both groups. The standard and honors treatment participants had similar lengths of school experience with computers and all but one male honors participant had taken the school district's required class. Of the 100% of the standard class participants who computers at home, 86% had Internet access, which about equaled the 87% of honors class participants who had computers at home (but not necessarily Internet access). Thirty-three percent honors participants did not have home Internet access, but only 14% treatment standard participants did not have home Internet access. The group with the most classroom computer experience, male honors, was the least likely of the treatment participants to have home computer or Internet access.

Modal writing preference and self-perceived writing quality. The modal writing preference and self-perceived writing quality survey questions had either gender and class level or class level differences. The hand writing modality was preferred by only female participants, more female standard (i.e., 20%) than female honors participants (i.e., 17%). All of the male participants and the majority of female participants (i.e., 80% standard and 75% honors) preferred writing by computer. Almost all (i.e., 95%) of the honors participants considered themselves to be good writers, compared to only 60% of standard participants. The good writer self-evaluation was given by all of the male honors group and 91% of the female honors group. There was no gender difference in the self-evaluations of good writer for the standard groups, both male and female groups being just 60%.

The language spoken at home. The language spoken at home differed by class level and by gender and class level. English was spoken at home by 53% of standard participants, considerably more than the 29% of honors participants who spoke English at home. The honors participants spoke Spanish or other languages at home (i.e., 43%) and were almost twice as likely to be bilingual (i.e., 27%) as the standard participants. The gender and class levels results showed English was spoken at home, from highest to lowest, by 60% of the male and 50% of the female standard participants and by 33% of the male and 25% of the female honors participants. Spanish and other languages were much more likely to be spoken at home by female standard (i.e., 40%) and female honors (i.e., 50%) participants than the male standard (i.e., 20%) or male honors (i.e., 33%) participants. About 20% of male standard participants and 23% of female honors participants were bilingual at home, with half less (i.e., 10%) for standard females participants and half more (i.e., 33%) for male honors participants.

Summary. For participant descriptions and perceptions, more standard level participants had computers and Internet access at home, yet the majority of all students preferred to write by computer instead of by hand. The definitive gender difference was that only the female honors and female standard participants had any preference for writing by hand. The male honors participants had the most school experience with computers and yet, they were the least likely to have computers or Internet access at home. The honors level classes had more participants who spoke non-English or were bilingual at home than the standard level classes, yet more honors participants considered themselves good writers than standard participants. More female honors and female standard participants spoke Spanish or other languages at home than either the male

standard or male honors participants, but it was the honors group who considered themselves better writers.

Preferences for the Automated Essay Scoring System

Computer writing feedback and automated essay scoring system's helpfulness and holistic score. The participant preferences for computer writing feedback, AES helpfulness, and the AES holistic score were disparate. Almost 75% of standard participants liked receiving writing feedback from a computer, but only 33% felt the AES system was the most help, yet 87% of the group liked the AES holistic score and felt it motivated them. Just over half (i.e., 57%) of the honors participants liked receiving computer feedback for their writing and only 25% felt that the AES system was the most help, but 71% liked the AES holistic score and felt it made them want to improve.

The gender and class levels group preferences were just a diverse as the class levels. While 44% the male honors group liked receiving writing feedback from a computer and only 25% thought that the AES was the most help to their writing, the group still had 78% who liked the holistic score and felt it motivated them. Sixty-seven percent of female honors participants liked the computer feedback, yet only 33% felt that the AES was the most help to their writing, and 67% said the AES holistic score made them want to improve. Eighty percent of the female standard group felt the AES holistic score motivated them, 70% liked the computer feedback for their writing, and 30% thought that the AES system was the most help. The highest ratings for all the categories were held by the standard male with 89% liking the computer feedback and 40% thinking the AES system was most important for their writing, while the entire group was motivated by the AES holistic score.

Even responders who did not like the AES holistic score said it motivated them to improve. Despite the participant's response to writing feedback from the computer or whether the AES system was the most help to their writing, the majority of all groups liked the AES holistic score and felt it motivated them to improve. The majority of participants' responding as wanting to improve their AES holistic scores supports previous research (Boone & Frost, 2005; Grimes & Warschauer, 2006).

Teacher's importance. Preferences about the AES system supported the importance of the teacher's help in learning to write. The largest percentage of the participants, 67.5% honors participants and 47% standard participants, considered the teacher the most help to their writing during the research period. The female standard (i.e., 40%) and female honors (i.e., 46%) participants considered the teacher the most important factor to improving their writing, but the male standard (i.e., 60%) and male honors (i.e., 75%) participants felt even more strongly about the teacher's importance. These results support the AES system's purpose to supplement, not replace, the teacher (Burstein, Chodorow et al., 2003; Burstein et al., 2004; Burstein & Marcu, 2003; Ware & Warschauer, 2006).

Importance of writing more. Writing more was considered most helpful most in improving their writing during the research period by 20% of the standard participants and fewer honors participants (i.e., 13%). The importance of writing more to improve their writing was evaluated as important as the AES system by 30% of the female standard participants and fewer female (i.e., 13%) honors participants, but not at all by the male standard and honors participants. The research results indicate a gender difference in the preference of writing more as the best way to improve writing. In

support of these increased writing opinions, it is known that increased writing with feedback will increase the quality of writing (Pritchard & Honeycutt, 2006).

Preferences regarding the best thing about the automated essay scoring system. The preferences for the best and worst thing about the AES system had definite foci. The AES system was considered the most help due to either its feedback of specific errors or overall help by 67% of both class levels. This was a higher rating than the 55% of the students who found an AES system helpful in previous research (Chen & Cheng, 2006). The genders did show some differences in their opinions of the AES system's greatest benefit. The female honors participants had the largest percentage (i.e., 83%) considering the AES system's feedback or overall help most beneficial and the male honors participants (i.e., 44%) had the smallest percentage. There was little gender difference in the standard group's perceptions of the AES error feedback or overall help as the best feature of the AES system. The 70% for standard female participants and 40% for standard male participants fell between the two honors gender groups' percentages.

A few participants in the current research mentioned that the AES system helped them to revise more. In support of this student opinion, one of the benefits of the AES system was thought to be to provide was more writing opportunities with feedback for students, without the corresponding increase in teacher's grading time (MacArthur, 2006; Warschauer & Ware, 2006). The other benefits of using the AES system that participants selected were more general, writing with a computer and using spell check, both of which are available with word processors. None of the female honors participants considered writing with a computer the most important benefit of using the AES system, while the other groups with that opinion ranged from 20% to 33%. The opinion of spell check

being the most important part of using the AES system varied most by gender, from 22% by the male honors group to none by the male standard group, with the female groups (i.e., standard female group 10% and honors female group 8%) midway between the male groups.

Preferences regarding the worst thing about the automated essay scoring system. The focus for the participant preferences for the worst thing about the AES system was consistent for the class levels and gender and class levels. The largest percentages of both class levels (i.e., 53% standard and 38% honors group) said "nothing" was the worst thing about the AES system. That was 50% to 60% of the female standard, male standard and male honors groups, but only 25% of the female honors group. After the answer of "nothing," the participant's answer with the next highest frequency for being the worst thing about the AES system was participant's not understanding specific trait error messages or categories, which was given by more honors (i.e., 24%) than standard participants (i.e., 13%). In comparison, research by Sommers (1982) reported in student interviews that students had trouble understanding what the teacher's comments meant for them to do with their writing. So while the current research shows that participants had trouble understanding the AES system's feedback, prior research shows that students may also have trouble understanding teacher feedback. Some participants from each of the gender groups also did not understand the AES trait error messages, with the 25% of female honors participants having the highest percentage, closely followed by male honors participants at 22%, and male standard participants at 20%. The female standard group's percentage who said they did not understand the AES trait error messages was at least one-half less than the other groups (i.e., 10%), and yet, that group had the second

highest percentage (i.e., 40%, with honors females being highest at 50%) of participants who spoke Spanish or another language at home. Therefore, language spoken at home may have no relationship to the lack of understanding the error trait messages.

More than twice of the honors participants (i.e., 24%) as standard participants (i.e., 10%) had complaints about the holistic score being the worst thing about the AES system. Despite this fact, 71% of the honors participants still reported in an earlier survey question that they liked the holistic score and it made them want to improve. Thirty-three percent of the female honors participants, which was three times as many as male honors (i.e., 11%) or female standard (i.e., 10%) participants, thought the holistic score was the worst thing about the AES system, while the male standard participants had no reports of this problem. Female honors participants were also the group who reported the largest percentage (i.e., 8%) of participants who did not understand the holistic score in an earlier survey question. Again, 66% of the female honors participants in an earlier survey question still reported that they were motivated by the holistic score.

More than twice as many standard participants (i.e., 13%) as honors participants (i.e., 5%) said the worst thing about the AES system was losing their work from lack of saving or computer crashes. Twenty percent of the standard male group had this complaint, with the female standard (i.e., 10%) and female honors (i.e., 8%) groups having complaint the complaint of losing their work half as frequently as the male standard group, but there were no such complaints from the male honors group. These mixed results do not support gender-based technology differences. The standard groups had an identical percentage (i.e., 7%) of complaints about inaccurate results or spell check issues, as did the honors group (i.e., 5%). The male standard group had no

complaints for either issue, but it was the male honors group who had no complaints for inaccuracies and the female honors group who had no complaints about spell check. The female standard group had the same percentage (i.e., 10%) of complaints for both issues.

Summary. Participant preferences for the AES system, both treatment class levels felt that (a) computer feedback on their writing was positive, (b) the teacher more important to their writing than the AES system, (c) the holistic score was liked and motivating, (d) the help and feedback provided by the AES system was the best thing from its use, and (e) "nothing" was the worst about the use of the AES. The male honors group was the least likely to consider computer feedback on writing positive and the male honors and standard groups had the highest percentages for the teacher being the most help. The gender and class level groups of participants liked the holistic score and felt it made them want to improve. Overall, both class levels and gender and class levels groups felt the best thing about the AES systems was the error feedback and overall help and "nothing" was the worst thing about the AES system. Only the female honors group did not have any participants who thought that writing on the computer was the best thing. The lowest percentage for "nothing" as the worst thing about the AES system was from the female honors group, who also had the highest percentage of participants who reported not understanding the AES feedback as the worst thing about the AES system. More honors participants than standard participants thought the holistic score was the worst thing about the AES system. No responses about the worst thing about the AES system were included by (a) the male honors groups about losing work, (b) the male standard group and the male honors group about inaccurate results, and (c) the male standard group and the female honors groups about spell check.

The Automated Essay Scoring System's Usability, Improvement, and Effectiveness

The participants' ratings of the AES system's (a) usability, (b) improvement, and (c) effectiveness were very consistent across the groups and also very close between the groups of questions, all within a range of 5% and none lower than 79%. Usability covered survey question items 1 through 4 and 10 and had the lowest average (i.e., 79%). The survey questions asking whether the AES system helped improve participants writing encompassed question items 5 through 9 and its average was highest (i.e., 83%). This high improvement rating, similar for all groups, is somewhat in contrast to the survey question that showed the participants considered the teacher of more help than the AES system. It shows that participants considered the AES system helpful to improving their writing, even if the teacher was more important.

The effectiveness survey question items were 11 through 18 and dealt with the different trait error sections, the organization and development section, and the AES system's online help. Even though earlier survey questions may have displayed problems with some sections of the AES system, the average (i.e., 81%) in this section provides resounding support for the assistance provided by the AES system's feedback. Overall, the participants considered the AES system very helpful to improving their writing.

The Automated Essay Scoring System's Frequency of Use

The frequency of use of the AES system by the participants is important because if an essay was only submitted once, it is unclear as to whether the participant was acting upon the system's feedback in order to improve the essay before handing it in for a grade, and, therefore, using the system as was expected. Together, the treatment class levels had an average of 2.60 submissions. This is higher than the average 2.38 submissions found

for 7th graders in previous research (Grimes & Warschauer, 2006). The honors participants had more multiple (i.e., more than one) pre-test submissions (i.e., 78%) than the standard participants (i.e., 53%). Compared to the number of pre-test submissions, the post-test submissions for the standard group (i.e., 68%) increased and those for the honors group (i.e., 57%) decreased, leaving the standard group with more multiple post-test submissions. Overall, there were still almost two out of five participants with only one submission for a pre- or post-test, thus possibly not using the system's feedback despite the submission. However, we do not know how many revisions were done by the treatment group using the AES system compared to those done by the control group.

In this study, as the previous research, the teacher cited classroom time limitations as the reason for limited use. The Boone and Frost (Boone & Frost, 2005) research documented that access impacted the use of an AES system. The frequency of AES use in this research was facilitated with the use of laptop carts in the classrooms. The surveys of the participants did indicate that the computers crashed and work was lost. Therefore, the current research results were also affected by problems with robustness of the network, Internet connectivity, and robustness of the AES application (treatment teacher, personal communication, fall, 2006).

The availability of the AES system was facilitated with the use of lap top carts. Despite the survey results showing the overwhelming majority of students liked the AES system and thought it was helpful and motivating, less than three-fourths of the participants actually used its feedback more than once for an essay. However, there was no frame of reference for comparison to the number of revisions that were done by the control group.

Limitations of the Study

The very fact that this study took place in the classroom created limitations to this study. The treatment teacher was selected by the school district from the voluntary teacher group that was using the AES system. That treatment teacher selected the control teacher and, thus the students, who would most closely match the treatment classes. Therefore, this quasi-experimental study did not use a random selection method for the teachers or the participants, thus compromising the generalizability of the study results.

It was expected that there would be a 10% dropout rate in the number of participants who would write pre-test but not post-essay samples, but the actual dropout rate was almost three times the expected rate (i.e., 28%), thus creating a very small sample. The treatment participants were taken from four classes, two standard and two honors, while the control participants were only from one standard class. It is not known what criteria were used to place the students in the honors level class. There was no initial measurement of the populations to establish their beginning writing skill level. There also was no measurement on how many revisions were done by the control class.

The school district provided the treatment classroom with a cart of laptops dedicated to their use, which is not standard classroom availability within the school district. The selected school had wireless access and network capacity to use the AES system on the provided laptops. There were, however, network, connectivity, and AES issues that periodically limited access. The beginning of the use of the AES program was delayed by a nationwide laptop battery recall.

The teachers' goals did not match the five paragraph evaluation expectation of the AES system. The teachers were focused on writing longer single paragraphs, not five

paragraph essays. Sometimes the errors found by the AES system precluded it from providing any holistic and/or error evaluation score.

The pre- and post-tests were not identical. A longer period of classroom time was spent on the pre-test assignment than on the post-test assignment. Both teachers gave verbal and written feedback on the pre-test, but only verbal feedback on the post-test. Students were allowed to re-submit the pre-test after receiving teacher feedback. Both assignments were persuasive genre, but the pre-test assignment was based on literature and the post-test assignment was based on the student's personal experience.

Implications and Future Research

This research adds to the body of knowledge on outcomes from the use of educational technology and the supplemental use AES systems in the classroom. The Nonequivalent Comparison Control Group (NCCG) design was used for this quasi-experimental design where the participants were in pre-existing groups from classrooms (Beins, 2004; Campbell & Stanley, 1963; Committee on Scientific Principles for Education Research, 2002; Creswell, 2002).

When the pre-test analysis showed significance, the mixed design analysis on the pre- and post-test was not reported, but rather the analysis of the post-test analysis, with the pre-test as a covariate, was reported. However, the post-test results by gender and class levels were not reported because the small sample size made the results inconclusive. Pre-test significance was shown for several measurements by class levels and gender and class levels, as shown in Table 28. The class levels pre-test significance for usage errors and organization and development structure was no longer evident by the post-test analysis, so perhaps the significance indicated group differences at the

beginning of the test period. The pre-test significant for the AES holistic by class levels did not include the treatment honors group, which showed significance in the post-test analysis.

Table 28

Pre-test Measurements with Significance for Class Levels or Gender and Class Levels

Class levels	Gender and Class Levels
AES holistic	NWP holistic
W/T	W/T
AES Usage	
AES Organization and development	

Note: AES = Automated essay scoring system; NWP = National Writing Project; W/T = words per t- unit.

The outcomes from the supplemental use of the AES system were mixed, revealing the need for more research. There was a significant increase in proficiency rate of change for treatment honors participants as measured by the AES holistic, but there was no significance for any of the treatment or control groups as measured by the NWP holistic score. None of the AES error measurements or the organization and development measurement showed any significance. None of the AES system measurements indicated any significant differences by gender and class levels, so educational technology outcomes do not seem to be effected by gender. The only gender and class levels analyses

that were not reported, NWP holistic and W/T, were related to human scoring. There is no doubt that the AES system was liked by almost all of the participants.

The results of this research point to possibilities for future research. The next study could include a larger pool of participants and a greater variety of high school grade levels than just ninth grade. The pre- and post-test could both be planned as similar length assignments within the curriculum. Research needs to investigate how many revisions are done by the control group. Qualitative research also needs to further investigate why such a large percentage of participants only had one AES submission for the pre- and/or post-test, even though the participants liked the AES. Such research may help determine how to increase the participation rate of multiple submissions by the research population. One suggestion from research in a college class that was considered successful with the use of an AES system was that the teacher required the students to have a holistic score of 4 before handing the paper into the teacher to grade (Chen & Cheng, 2006).

The treatment population did not significantly improve their writing development, trait errors, or organization and development structure with the use of the AES system. However, the results for the proficiency outcome were mixed, with the treatment honors improving on the AES holistic but not on the NWP holistic. Based on the significant outcome, the preferences of the participants to write with a computer, and participant beliefs that the AES system helped their improve their writing, the AES system's use in the classroom should be supported while more research is conducted.

APPENDIX A

CRITERION'S SCORING CATEGORIES AND SUBCATEGORIES

Following is the list of categories and subcategories that Criterion uses for trait scoring (Educational Testing Service, 2006a). The main categories are indicated in bold.

Grammar Errors

Fragment or Missing Comma

Run-on Sentences

Garbled Sentences

Subject Verb Agreement

Ill-Formed Verbs

Pronoun Errors

Possessive Errors

Wrong or Missing Word

Proofread This!

Usage Errors

Wrong Article

Missing or Extra Article

Confused Word

Wrong Form of Word

Faulty Comparisons

Preposition Error

Nonstandard Word or Verb Form

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

Style

Repetition of Words

Inappropriate Words or Phrases

Sentences Beginning with Coordinating Conjunctions

Too Many Short Sentences

Too Many Long Sentences

Passive Voice

Number of Words

Number of Sentences

Average Number of Words per Sentence

Organization and Development

Introductory Material

Thesis Statement

Main Ideas

Supporting Ideas

Conclusion

Transitional Words and Phrases

Other

APPENDIX B

DEFINITIONS

adverbial clause – A dependent clause that begins with a subordinating conjunction and describes a verb in the main clause (Benner, 2007; Nippold et al., 2005). It answers the question of where, why, how, when, or to what degree. Common subordinating conjunctions include: after, before, until, while, because, since, as, so, that, in order that, if unless, whether, though, although, even though, and where.

clause – A structure with a subject and a main verb (Hunt, 1965a; Nippold et al., 2005). This includes independent clauses, adverbial clauses, adjective/relative clauses, and nominal clauses. It does not include phrases.

independent clause – It contains a subject and a main verb, and it expresses complete thought (Nippold et al., 2005).

mixed design ANOVA – An analysis of variance with repeated measures for one factor and independent groups for the other factors (Keppel, 1991).

nominal clause – A subordinate clause that names a person, place or thing (Benner, 2007; Nippold et al., 2005).

relative clause – A subordinate clause that begins with the words which, that (for things), or who, whose, whom (for people), or when, where, or why (Benner, 2007; Simmons, 2007). Also known as an adjective clause, it describes a noun and will answer the questions: What kind? Which one?

t-unit – An independent clause with a subject, a main verb, and all the supporting clauses (Hunt, 1965a; Nippold et al., 2005). The supporting clauses include: adverbial, relative, and nominal.

writing development – Characteristics of individual writing development located at some point along a continuum; a part of language development (Wolfe-Quintero et al., 1998).

writing proficiency – An overall evaluation of an essay, a holistic score, which is greater than the sum of the evaluation of specific writing traits like grammar (Wolcott & Legg, 1998).

writing prompt – The topic to be used for the writing the essay.

W/T – Words per t-unit is calculated by dividing the total number of words by the total number of t-units.

APPENDIX C

SEMI-STRUCTURED TEACHER INTERVIEWS

Following is a list of opening questions for the participants' teachers:

Describe the assignments used to collect the essay samples.

What would differentiate this essay-sample assignment from others given in your classes?

How many different, graded writing assignments were given for the year?

How many times would a specific assignment be graded?

For the treatment classrooms, additional questions would define teachers' assignment methodologies and the integration of the AES system into the classroom. Following are the preliminary questions for the teachers whose classes will use the intervention (Grimes & Warschauer, 2006):

How do you teach the use of the AES system?

Why would you or would you not recommend this program to other teachers?

How do you utilize the AES system within your classes?

Do you feel that the AES's scores are fair?

APPENDIX D

TREATMENT STUDENTS' INTERVIEWS

Following are guiding questions for interviews of the treatment participants (Grimes & Warschauer 2006; Chen & Cheng 2006).

Section 1: Directions: On a scale from 0 (no chance) to 100 (completely certain), select a number that indicates how confident are you about the use of *Criterion* as described in the following statements (Schiffman, Reynolds et al. 1981).

1. I want to use *Criterion* next year.
2. I use *Criterion* at home.
3. I find *Criterion* easy to use.
4. I sometimes have trouble using *Criterion*. Can you give an example of a problem you might have?
5. I revise my writing more when I use *Criterion*.
6. Writing with *Criterion* has increased my confidence in my writing.
7. *Criterion* has good suggestions for improving my writing.
8. The essay scores *Criterion* gives are fair.
9. *Criterion* helps improve my writing. Can you give an example of how Criterion has improved your writing?

10. *Criterion*'s response is fast enough.
11. *Criterion*'s report of grammar errors (for example, subject verb agreement or run-on sentences) is helpful.
12. *Criterion*'s report of usage errors (for example, missing word or confused words) is helpful.
13. *Criterion*'s spell checker is helpful.
14. *Criterion*'s report of mechanics errors (for example, missing final punctuation or missing capital letter) is helpful.
15. *Criterion*'s report of style errors (for example, too many short sentences or sentences beginning with coordinating conjunction) is helpful.
16. *Criterion*'s report on essay length (for example, number of words or number of sentences) is helpful.
17. *Criterion*'s organizational report identifying an essay's parts or missing parts (for example, topic sentence or supporting sentence) is helpful.
18. I use *Criterion*'s Writers Handbook to help me correct errors.

The remaining questions are short answer or completion.

19. The best thing about *Criterion* is _____
20. The worst thing about *Criterion* is _____
21. Have you taken the required computer class, usually taken in ninth grade or middle school?
22. Approximately what grade did you start using computers in the classroom?
23. Do you have a computer at home?
24. Do you have an internet connection at home?

25. Do you prefer writing by hand or on the computer? Why is this your preference?
26. How do you feel about having a computer respond to your writing instead of a person?
27. What language do you speak at home?
28. Are you a good writer?
29. What helped you most with your writing this year? For example, practice, *Criterion*, or teacher.
30. Do you like receiving the essay score from 1 – 6? Does the score make any difference to you? What do you do if you receive a low score?
31. NOTE the student's gender.

APPENDIX E

T-UNIT AND CLAUSE SCORING GUIDELINES

Guidelines for measuring t-units and clauses by Polio (1997, p. 139-140):

T-units

a. A t-unit is defined an independent clause and all its dependent clauses.

b. Count run-on sentences and comma splices as two t-units with an error in the first t-unit.

   ex: My school was in Saudi Arabia, it was the best school there.

   t-unit / t-unit

   1 error error-free

   If several comma-splices occur in a row, count only the last as error free.

c. The following rules pertain to sentence fragments.

   If the verb or copula (i.e., linking verb such as to be) is missing, count the sentence as 1 t-unit with an error (*The American Heritage Dictionary of the English Language*, 2000).

   If a noun phrase is standing alone, attach it to the preceding or following t-unit as appropriate and count as an error.

   If a subordinate clause is standing alone, attach it to the preceding or following sentence and count it as 1 t-unit with an error.

d. When there is a grammatical subject deletion in a coordinate clause, count the entire sentence as 1 t-unit.

   ex: First we went to our school and then went out with our friends.

e. Count both "so" and "but" as coordinating conjunctions. Count "so that" as a subordinating conjunction unless "so" is obviously meant.

f. Do not count tag-questions as separate t-units.

g. Count a sentence with a deleted subordinating conjunction as a subordinate clause as in: I believe that A and (that) B = 1 t-unit.

h. But, direct quotes should be counted as:

   John said, "A and B."

   1 T-unit 1 t-unit

i. Assess the following type of structures on a case-by-case basis:

   If A, then B and C.

   As a result, A or B.

j. Count t-units in parentheses as individual t-units.

Clauses

a. A clause equals an overt subject and a finite verb. The following are only one clause each:

   He left the house and drove away.

   He wanted John to leave the house.

b. Only an imperative does not require a subject to be considered a clause. For example:

   Go away!

c. In a sentence that has a subject with only an auxiliary verb, do not count that subject and verb as a separate clause or as a separate t-unit (e.g., John likes to ski and Mary does too; John likes to ski, doesn't he?; John is happy and Mary is too).

APPENDIX F

PERMISSIONS

There are multiple permissions shown in this appendix: (a) informed consent from the parent (non-intervention), (b) informed consent from the student (non-intervention), (c) informed consent from the parent (intervention), (d) informed consent from the student (intervention), (e) UNLV modification approval, and (f) informed consent from the teacher.

Informed Consent of Parent (Non-Intervention)

Dr. Boone, a professor in the Curriculum and Education department at the University of Nevada at Las Vegas, is the primary investigator of a study entitled, "Criterion Writing." Please read the following information, and if you agree to have your child included in this study, please sign at the bottom. The research is sponsored by officials of the Clark County School District.

Description:

In this study, examples of your child's writing will be anonymously analyzed and the results will be compared to the results from students at another school. The students at the other school are using a computer software program as part of their writing instruction. The study wants to find out if the software at the other school is helpful in improving student writing skills. The study will include approximately 100 students from each school. There will be no additional tests or graded class activities associated with this project. The analysis of your child's writing is not a test, and you child will not be graded based on this analysis. Participation in this study will not affect your child's grade.

Risks and Benefits:

Risks involved in doing this study are minimal. Your child may be nervous about having his/her writing analyzed by the research team. Concerns about study participation may be discussed with your child's teacher, the people administering the study, or Dr. Boone at 702-895-3233. If you have a question about the rights of research subjects, you can contact the UNLV Office for the Protection of Research Subjects at (702) 895-2794.

Costs and Payments.

There are no costs for participating in this study.

Confidentiality.

All information obtained during the course of this study is strictly confidential and will be available only to authorized study staff members. Reports in scientific journals will not include any information that identifies participants in this study. All data will be kept in a locked filing cabinet on the UNLV campus for a minimum of three years and then destroyed.

Right to Withdraw at Any Time:

Your child is free to refuse participation in this study, or to withdraw at any time. Withdrawal in this study will in no way negatively affect your child.

I have read the information above, and I agree to its contents. All of my questions concerning this research have been answered. If I have any questions in the future about this study, they will be answered by Dr. Boone. A copy of this form will be given to me.

Note: This form can only be signed by a legal parent. Nevada law requires a court approval for wards to be allowed to participate in research

Signature of Parent: _____ Date:_____

Printed name of child: _____

Informed Assent for Minors (Non-intervention)

Dr. Boone, a professor in the Curriculum and Instruction department at the University of Nevada at Las Vegas, is doing a research study called, "Criterion Writing," in which students who use a computer program called Criterion, will have their writing samples compared to writing samples from students who do not use the Criterion computer software. You have been offered a chance to be in this study because you are not using the Criterion computer software. Please read this page and, if you want to be in this study, sign your name at the bottom. The research study is sponsored by officials of the Clark County School District.

Right to Withdraw at Any Time: You do not have to be part of this study if you don't want to. If you decide to be in the study and then change your mind, you can tell your teacher or the researcher, and they will not use your information. If you decide not to be in this study, it will not affect your grade or anything else about your schoolwork.

What you will be asked to do: If you decide to be in this study, some examples of your writing will be analyzed. This will not affect your grade and you will not be given any extra writing assignments.

Risks and Benefits: There is a risk that you might be nervous about using your information as part of this study. If you have any questions at any time during the study, you can call Dr. Boone at 702-895-3233. If you have a question about the rights of research subjects, you can contact the UNLV Office for the Protection of Research Subjects at (702) 895-2794.

This study may be good for everyone who takes writing classes by seeing if the Criterion software used at another school is a good way to teach people your age. The analysis of your writing will help us make that decision.

Costs and Payments: There are no costs or payments for participating in this study.

Confidentiality. We will keep your information in a safe place where it will be seen only by people who are part of the research team and by people whose job it is to make sure this is a safe and fair study. We will keep your information a minimum of three years, and then destroy it.

Talk to your parents: You should talk to your parents about being in this study before you sign this form. Your parents will also get a form to sign saying that you can be in the study.

You will get to keep a copy of this form. If you don't get a copy of the form, please ask for one. If you have any questions at any time during the study, you can call Dr. Boone at 702-895-3233.

I have read this form and agree to be in the study. I know I can choose not to be in the study at any time. I will ask Dr. Boone or any of the researchers if I have any questions during the study.

Printed Name of Student: _____

Signature of Student: _____ Date:_____

Signature of Research Assistant: _____Date: _____

Informed Consent of Parent (Intervention)

Dr. Boone, a professor in the Curriculum and Education department at the University of Nevada at Las Vegas, is the primary investigator of a study entitled, "Criterion Writing." Please read the following information, and if you agree to have your child included in this study, please sign at the bottom. The research is sponsored by officials of the Clark County School District.

Description:

In this study, your child will be interviewed two times during the semester, regarding a piece of writing software (Criterion), which your child uses at school. Also, examples of your child's writing will be anonymously analyzed to find out if using the Criterion software has an effect on your child's writing skills. There will be no additional tests or graded class activities associated with this project. The interview will include questions about your child as a writer and about the Criterion software. The interview is not a test, and you child will not be graded on his/her answers to the questions. Participation in this study will not affect your child's grade.

Risks and Benefits:

Risks involved in doing this study are minimal. Your child may be nervous about having an adult ask him/her about how he/she uses the computer. Concerns about study participation may be discussed with your child's teacher, the people administering the study, or Dr. Boone at 702-895-3233. If you have a question about the rights of research subjects, you can contact the UNLV Office for the Protection of Research Subjects at (702) 895-2794.

Costs and Payments.

There are no costs for participating in this study, but there is the cost of your child's time (about 20 minutes for each interview—40 minutes total).

Confidentiality.

All information obtained during the course of this study is strictly confidential and will be available only to authorized study staff members. Reports in scientific journals will not include any information that identifies participants in this study. All data will be kept in a locked filing cabinet on the UNLV campus for a minimum of three years and then destroyed.

Right to Withdraw at Any Time:

Your child is free to refuse participation in this study, or to withdraw at any time. Withdrawal in this study will in no way negatively affect your child.

I have read the information above, and I agree to its contents. All of my questions concerning this research have been answered. If I have any questions in the future about this study, they will be answered by Dr. Boone. A copy of this form will be given to me.

Note: This form can only be signed by a legal parent. Nevada law requires a court approval for wards to be allowed to participate in research

Signature of Parent: _____ Date:_____

Printed name of child: _____

Informed Assent for Minors (Intervention)

Dr. Boone, a professor in the Curriculum and Instruction department at the University of Nevada at Las Vegas, is doing a research study called, "Criterion Writing." You have been offered a chance to be in this study because you are using the Criterion computer software, and Dr. Boone wants to investigate two things: (1). He wants to get student reactions to using the software. (2). He wants to see if the software helps students improve their writing. Please read this page and, if you want to be in this study, sign your name at the bottom. The research study is sponsored by officials of the Clark County School District.

Right to Withdraw at Any Time: You do not have to be part of this study if you don't want to. If you decide to be in the study and then change your mind, you can tell your teacher or the researcher, and they will not use your information. If you decide not to be in this study, it will not affect your grade or anything else about your schoolwork.

What you will be asked to do: If you decide to be in this study, you will be interviewed two times by a researcher from UNLV. The interview will include questions about you as a writer and about the Criterion software. The interview is not a test, and you will not be graded on your answers to the questions. Also, some examples of your writing will be analyzed to find out if using the Criterion software has an effect on your writing skills.

Risks and Benefits: There is a risk that you might be nervous about using your information as part of this study. There is also a risk that you might be uncomfortable during an interview. If you have any questions at any time during the study, you can call Dr. Boone at 702-895-3233. If you have a question about the rights of research subjects, you can contact the UNLV Office for the Protection of Research Subjects at (702) 895-2794.

This study may be good for everyone who takes writing classes by seeing if the software you are using is a good way to teach people your age. Your answers to the interview questions may help make the software easier to use and more clear.

Costs and Payments: There are no costs or payments for participating in this study, although there is the cost of your time (about 20 minutes for each interview).

Confidentiality. We will keep your information in a safe place where it will be seen only by people who are part of the research team and by people whose job it is to make sure this is a safe and fair study. We will keep your information a minimum of three years, and then destroy it.

Talk to your parents: You should talk to your parents about being in this study before you sign this form. Your parents will also get a form to sign saying that you can be in the study.

You will get to keep a copy of this form. If you don't get a copy of the form, please ask for one. If you have any questions at any time during the study, you can call Dr. Boone at 702-895-3233.

I have read this form and agree to be in the study. I know I can choose not to be in the study at any time. I will ask Dr. Boone or any of the researchers if I have any questions during the study.

Printed Name of Student: _____

Signature of Student: _____ Date:_____

Signature of Research Assistant: _____Date: _____

UNLV
UNIVERSITY OF NEVADA LAS VEGAS

Social/Behavioral IRB – Expedited Review
Modification Approved

NOTICE TO ALL RESEARCHERS:

Please be aware that a protocol violation (e.g., failure to submit a modification for any change) of an IRB approved protocol may result in mandatory remedial education, additional audits, re-consenting subjects, researcher probation suspension of any research protocol at issue, suspension of additional existing research protocols, invalidation of all research conducted under the research protocol at issue, and further appropriate consequences as determined by the IRB and the Institutional Officer.

DATE: April 27, 2007

TO: **Dr. Randall Boone**, Curriculum and Instruction

FROM: Office for the Protection of Research Subjects

RE: Notification of IRB Action by Dr. J. Michael Stitt, Chair
Protocol Title: **Criterion Writing**
Protocol #: 0509-1695

The modification of the protocol named above has been reviewed and approved. Modifications reviewed for this action include:

- Use of data from a prior study conducted by teachers whose students are the participants in the Criterion Writing research.
- Addition of population of teachers whose students are participating in the Criterion Writing research. They will now complete two semi-structured interviews.

This IRB action will not reset your expiration date for this protocol. The current expiration date for this protocol is September 20, 2007.

PLEASE NOTE:
Attached to this approval notice is the **official Informed Consent/Assent (IC/IA) Form** for this study. The IC/IA contains an official approval stamp. Only copies of this official IC/IA form may be used when obtaining consent. Please keep the original for your records.

Should there be *any* change to the protocol, it will be necessary to submit a **Modification Form** through OPRS. No changes may be made to the existing protocol until modifications have been approved by the IRB.

Should the use of human subjects described in this protocol continue beyond September 20, 2007, it would be necessary to submit a **Continuing Review Request Form** *60 days* before the expiration date.

If you have questions or require any assistance, please contact the Office for the Protection of Research Subjects at OPRSHumanSubjects@unlv.edu or call 895-2794.

UNLV
UNIVERSITY OF NEVADA LAS VEGAS

INFORMED CONSENT

Department of Curriculum & Instruction

TITLE OF STUDY: Criterion Writing

INVESTIGATOR(S): Dr. Randall Boone

CONTACT PHONE NUMBER: 895-3233

Purpose of the Study
You are invited to participate in a research study. The purpose of these teacher interviews is to identify the teaching processes to collect sample (either control or intervention) essays and integrate Criterion's use into the classroom.

Participants
You are being asked to participate in this interview study because your students are either control or intervention participants in the Criterion Writing research.

Procedures
If you volunteer to participate in this teacher interview portion of the Criterion study, you will be asked to do the following: provide descriptive information about the assignments that resulted in essay samples. In addition, the teachers of the students using the intervention will also be asked how they taught and integrated Criterion into their classroom.

Benefits of Participation
There *may not* be direct benefits to you as a participant in this study. You will be able to provide the description of your teaching strategies for participating in this study. We hope to learn whether Criterion is beneficial to student writing outcomes.

Risks of Participation
There are risks involved in all research studies. This teacher interview portion of the Criterion Writing study may include only minimal risks. You may be uncomfortable in describing the process used to collect writing samples or the process of integrating Criterion into the classroom.

Cost /Compensation
There *will not* be financial cost to you to participate in this study. The interview study will take approximately 15- 30 *minutes* of your time during two interviews. You *will not* be compensated for your time. *The University of Nevada, Las Vegas may not provide compensation or free medical care for an unanticipated injury sustained as a result of participating in this research study.*

Contact Information
If you have any questions or concerns about the study, you may contact Dr. Randall Boone at **895-3233.** For questions regarding the rights of research subjects, any complaints or comments regarding

UNLV
UNIVERSITY OF NEVADA LAS VEGAS

INFORMED CONSENT

Department of Curriculum & Instruction

TITLE OF STUDY: Criterion Writing

INVESTIGATOR(S): Dr. Randall Boone

CONTACT PHONE NUMBER: 895-3233

the manner in which the study is being conducted you may contact **the UNLV Office for the Protection of Research Subjects at 702-895-2794.**

Voluntary Participation

Your participation in these teacher interviews for the Criterion Writing study is voluntary. You may refuse to participate in this teacher interview part of this study. You may withdraw at any time without prejudice to your relations with the university. You are encouraged to ask questions about this teacher interview study at the beginning or any time during the research study.

Confidentiality

All information gathered in this study will be kept completely confidential. No reference will be made in written or oral materials that could link you to this study. All records will be stored in a locked facility at UNLV for at least 3 years after completion of the study. After the storage time the information that was gathered will be destroyed.

Participant Consent:

I have read the above information and agree to participate in this study. I am at least 18 years of age. A copy of this form has been given to me.

Signature of Participant _____ Date _____

Participant Name (Please Print) _____

Participant Note: Please do not sign this document if the Approval Stamp is missing or is expired.

REFERENCES

The American Heritage Dictionary of the English Language. (2000). Retrieved April 5, 2007, from http://www.bartleby.com/61/23/C0632300.html

Attali, Y. (2004). *Exploring the feedback and revision features of Criterion*^{SM}. Paper presented at the National Council on Measurement in Education (NCME), San Diego, CA.

Bangert-Drowns, R. L. (1993). The word processor as an instructional tool: A meta-analysis of word processing in writing instruction. *Review of Educational Research, 63*(1), pp. 69-93.

Batali, J. (2006). *Natural Language Processing*. Retrieved July 15, 2006, from http://cogsci.ucsd.edu/%7Ebatali/108b/lectures/natlang.txt

Beins, B. C. (2004). *Research methods: A tool for life*. Boston: Pearson Education, Inc.

Benner, M. L. (2007). *Online writing support.* Retrieved March 15, 2007, from http://wwwnew.towson.edu/ows/index.htm

Berninger, V. W., Fuller, F., & Whitaker, D. (1996). A process model of writing development across the life span. *Educational Psychology Review, 8*(3), 193-217.

Berninger, V. W., & Swanson, H. L. (1994). Modifying Hayes and Flowers Model of Skilled Writing to Explain Beginning and Developing Writing. In E. C. Butterfield & J. S. Carlson (Eds.), *Children's writing: Toward a process theory of development* (Vol. 2, pp. 57-81). Greenwich, Conn.: JAI Press Inc.

Blok, H., & de Glopper, K. (1992). Large scale writing assessment. In L. Verhoeven & J. H. A. L. De Jong (Eds.), *The construct of language proficiency: Applications of psychological models to language assessment* (pp. 101-111). Amsterdam, Netherlands: John Benjamins Publishing Company.

Boone, R., & Frost, K. (2005). *Criterion^{SM} research project: Final report to CCSD (Technical Report)* (Unpublished). Las Vegas: Department of K-12 Mathematics, Science, and Instructional Technology, Clark County School District.

Brooks, D. W., & Crippen, K. J. (2001). Learning difficult content using the web: Strategies make a difference. *Journal of Science Education and Technology, 10*(4), 283-283.

Bruning, R., & Horn, C. (2000). Developing motivation to write. *Educational Psychologist, 35*(1), 25-37.

Buchanan, J., Eidman-Aadahl, E., Friedrich, L., LeMahieu, P., & Sterling, R. (2006). *Part I - Summary Report of National Results*. Berkeley, CA: National Writing Project.

Burstein, J., Chodorow, M., & Leacock, C. (2003). *Criterion^{SM} Online Essay Evaluation: An Application for Automated Evaluation of Student Essays.* Paper presented at the Fifteenth Annual Conference on Innovative Applications of Artificial Intelligence, Acapulco, Mexico.

Burstein, J., Chodorow, M., & Leacock, C. (2004). Automated essay evaluation: The Criterion online writing service. *AI magazine, 25*(3), 27-36.

Burstein, J., & Higgins, D. (2005). *Advanced capabilities for evaluating student writing: Detecting off-topic essays without topic-specific training*. Retrieved July 25, 2006, from http://www.ets.org/Media/Research/pdf/erater_burstein_higgins_CR.pdf

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

Burstein, J., & Marcu, D. (2003). Developing technology for automated evaluation of discourse structure in student essays. In M. D. Shermis & J. Burstein (Eds.), *Automated essay scoring: A cross-disciplinary perspective* (pp. 209-230). Hillsdale, NJ: Lawrence Erlbaum Associates.

Burstein, J., Marcu, D., & Knight, K. (2003). Finding the WRITE Stuff: Automatic Identification of Discourse Structure in Student Essays. *IEEE Intelligent Systems, 37*(4), 455-467.

Campbell, D. T., & Stanley, J. C. (1963). Experimental and quasi-experimental designs for research on teaching. In N. L. Gage (Ed.), *Handbook of Research on Teaching*. Chicago: Rand McNally & Company.

Chen, C.-F. E., & Cheng, W.-Y. (2006). *The use of a computer-based writing program: Facilitation or frustration?* Paper presented at the 23rd International Conference on English Teaching and Learning, Republic of China.

Chodorow, M., & Burstein, J. (2004). *Beyond essay length: Evaluating e-raters performance on TOEFL essays*. Princeton, NJ: Educational Testing Service.

Christmann, E., Badgett, J., & Lucking, R. (1997). Progressive comparison of the effects of computer-assisted instruction on the academic achievement of secondary students. *Journal of Research on Computing in Education, 16*(3), 281-296.

Chung, G. K. W. K., & O'Neil, H. F., Jr. (1997). *Methodological approaches to online scoring of essays*. Retrieved July 26, 2006, from http://www.eric.ed.gov/ERICDocs/data/ericdocs2/content_storage_01/0000000b/80/25/0d/d9.pdf

Cizek, G. J., Page, B. A., Keith, T. Z., Shermis, M. D., Daniels, K. E., Ponisciak, S., et al. (2003). Psychometric issues in automated essay scoring. In M. D. Shermis & J. Burstein (Eds.), *Automated essay scoring: A cross-disciplinary perspective* (pp. 123-192). Mahwah, New Jersey: Lawrence Erlbaum Associates.

Clare, L., Valdes, R., & Patthey-Chavez, G. G. (2000). *Learning to write in urban elementary and middle schools: An investigation of teachers' written feedback on student compositions* (No. 526). Los Angeles: University of California.

Cohen, J. (1977). *Statistical power analysis for the behavioral sciences* (Rev. ed.). New York: Academic Press.

Committee on Scientific Principles for Education Research. (2002). *Scientific Research in Education*. Retrieved April 17, 2007, from http://books.nap.edu/catalog.php?record_id=10236

Connors, R. J., & Lunsford, A. A. (1988). Frequency of formal errors in current college writing, or Ma and Pa Kettle do research. *College Composition and Communication, 39*(4), 395-409.

Coxhead, P. (2001). *An introduction to Natural Language Processing (NLP)*. Retrieved July 15, 2006

Creswell, J. W. (2002). *Educational research: Planning, conducting, and evaluating quantitative and qualitative research*. Upper Saddle River, New Jersey: Merrill Prentice Hall.

Crippen, K. J. (2000). *Analysis of learning at an Advanced Placement descriptive chemistry web site.* Unpublished Dissertation, University of Nebraska - Lincoln, Lincoln, NE.

Dale, R., & Douglas, S. (1997). Two investigations into intelligent text processing. In M. Sharples & T. van der Geest (Eds.), *The new writing environment: Writers at work in a world of technology* (pp. 123-145). Berlin: Springer-Verlag.

Day, J. C., Janus, A., & Davis, J. (2003). *Computer and internet use in the United States: 2003*. Washington, D.C.: United States Census Bureau.

Dynarski, M., Agodini, R., Heviside, S., Novak, T., Carey, N., Campuzano, L., et al. (2007). *Effectiveness of reading and mathematics software products: Findings from the first student cohort.* Retrieved April 15, 2007, from http://ies.ed.gov/ncee/pdf/20074005.pdf

Educational Testing Service. (2006a). *Criterion^{SM} Online Tour*. Retrieved October 7, 2006, from http://www.ets.org/criterion/tour/

Educational Testing Service. (2006b). *ETS.org*. Retrieved July 23, 2006, from http://www.ets.org/portal/site/ets/menuitem.3a88fea28f42ada7c6ce5a10c3921509/?vgnextoid=85b65784623f4010VgnVCM10000022f95190RCRD

Educational Testing Service. (2007a). *Criterion[SM] online writing evaluation*. Retrieved March 1, 2007, from http://www.ets.org/portal/site/ets/menuitem.435c0b5cc7bd0ae7015d9510c3921509/?vgnextoid=b47d253b164f4010VgnVCM10000022f95190RCRD

Educational Testing Service. (2007b). *Criterion[SM] online writing evaluation version 7.1 prewriting tools and other enhancements*. Retrieved March 1, 2007, from http://criterion28.ets.org/news/Criterion%20v_7%201%20%20Prewriting%20Tools%20Release.pdf

Educational Testing Service. (2007c). *FAQ (Frequently Asked Questions) About Criterion[SM] Writing Evaluation*. Retrieved April 15, 2007, from http://www.ets.org/portal/site/ets/menuitem.1488512ecfd5b8849a77b13bc3921509/?vgnextoid=f128af5e44df4010VgnVCM10000022f95190RCRD&vgnextchannel=37ed253b164f4010VgnVCM10000022f95190RCRD

Entertainment Software Association. (2007). *Facts and research*. Retrieved March 15, 2007, from http://www.theesa.com/facts/index.php

Fitzgerald, C. A., Graham, S., & Fitzgerald, J. (2006). *Handbook of Writing Research*. New York: Guilford Press.

Gee, J. P. (2003). *What video games have to teach us about learning and literacy* (First ed.). New York: Palgrave Macmillan.

Graham, S., & Perin, D. (2006). *Writing next: Effective strategies to improve writing of adolescents in middle and high schools*. Retrieved October 1, 2006, from http://www.all4ed.org/publications/WritingNext/WritingNext.pdf

Grant, M. M., Wang, W., & Potter, A. (2005). Computer on wheels: an alternative to 'each one has one'. *British Journal of Educational Technology, 36*(6), 1017-1034.

Grimes, D., & Warschauer, M. (2006). *Automated Essay Scoring in the Classroom.* Paper presented at the American Educational Research Association (AERA) Symposium on Technology and Literacy, San Francisco, CA.

Herrington, A., & Moran, C. (2001). What happens when machines read our students' writing? *College English, 61*(4), 480-498.

Higgins, D., Burstein, J., Marcu, D., & Gentile, C. (2004). *Evaluating multiple aspects of coherence in student essays*. Retrieved March 25, 2006, from http://www.ets.org/Media/Research/pdf/erater_higgins_dis_coh.pdf

Hokanson, B., & Hooper, S. (2004). *Integrating technology in classrooms: We have met the enemy and he is us.* Paper presented at the Association for Educational Communications and Technology, Chicago, IL.

Hunt, K. W. (1965a). *Grammatical structures written at three grade levels*. Urbana, IL: The National Council of Teachers of English.

Hunt, K. W. (1965b). Synopsis of clause-to-sentence length factors. *The English Journal, 54*(4), 300, 305-309.

Hunt, K. W. (1970). Syntactic maturity in school children and adults. *Society for Research in Child Development Monographs, 35*(Serial No. 134).

Juul, J. (2003). *The game, the player, the world: Looking for a heart of gameness.* Paper presented at the Level Up: Digital Games Research, Utrecht, Netherlands.

Keith, T. Z. (2003). Validity of automated essay scoring systems. In M. D. Shermis & J. Burstein (Eds.), *Automated essay scoring: A cross-disciplinary perspective* (pp. 147-167). Mahwah, New Jersey: Lawrence Erlbaum Associates.

Kelly, P. A. (2001, April 11-13). *Computerized scoring of essays for analytical writing assessment: Evaluating score validity.* Paper presented at the Annual Meeting of the National Council on Measurement in Education, Seattle, WA.

Keppel, G. (1991). *Design and Analysis:*

A Researcher's Handbook (Third ed.). Upper Saddle River, NJ: Prentice-Hall, Inc.

Kincaid, J. P., Fishburne, R. P., Jr., Rogers, R. L., & Chissom, B. S. (1975). *Derivation of new readability formulas (Automated Readability Index, Fog Count and Flesh Reading Ease Formula) for Navy enlisted personnel* (No. ED108134).

Kingsley, K. V. (2005). *A quantitative investigation of American history software on middle school student achievement scores.* Unpublished Dissertation, University of Nevada, Las Vegas.

Kukich, K. (2000). Beyond automated essay scoring. *IEEE Intelligent Systems, 15*(5), 22-27.

Liddy, E. D. (2001). *Natural Language Processing*. Retrieved 24 April, 2007, from http://www.cnlp.org/publications/03NLP.LIS.Encyclopedia.pdf

Likert, R., Roslow, S., & Murphy, G. (1934). A simple and reliable method of scoring the Thurston attitude scales. *Journal of Social Psychology* (5), 228-238.

Lipsey, M. W., & Wilson, D. B. (2000). *Practical meta-analysis*. Thousand Oakes, CA: Sage.

Loban, W. (1976). *Language development: Kindergarten through grade twelve.* Urbana, IL.

Luger, G. F. (2001). *AI: Early history and applications.* Retrieved April 20, 2007, from http://www.cs.unm.edu/~luger/ai-final/chapter1.html

MacArthur, C. A. (2006). The effects of new technologies on writing and writing processes. In C. A. MacArthur, S. Graham & J. Fitzgerald (Eds.), *Handbook of writing research* (pp. 248 - 262). New York: The Guilford Press.

Macdonald, N. H., Frase, L. T., Gingrich, P. S., & Keenan, S. A. (1982). The Writer's Workbench: Computer aids for text analysis. *IEEE Transactions on Communications, Com-30*(1), 105-110.

Matsumura, L. C., Patthey-Chavez, G. G., Valdes, R., & Garnier, H. (2002). Teacher feedback, writing assignment quality, and third-grade students' revision in lower- and higher-achieving urban schools. *The Elementary School Journal, 103*(1), 3-25.

McCanne, L. (2004). *The relationships among computer literacy, computer access, and achievement in high school students.* Unpublished Dissertation, Boston University, Boston.

McMillan, J. H. (2004). *Educational research: Fundamentals for the consumer* (4th ed.). Boston: Pearson Education, Inc.

McMillan, J. H., & Schumacher, S. (2006). *Research in Education.* Boston: Pearson Education, Inc.

Mertens, D. M. (1998). *Research methods in education and psychology.* Thousand Oakes, CA: Sage Publications, Inc.

Nevada Department of Education. (2006-2007). *HSPE in Writing*. Retrieved March 30, 2007, from http://www.doe.nv.gov/statetesting/npep.attachment/307395/TAB_5_-_HSPE_in_Writing.pdf

Nevada Department of Education. (2007). *Writing Assessment*. Retrieved March 30, 2007, from http://www.doe.nv.gov/statetesting/writingassess.html

Nevada Department of Education. (2008). *Nevada Report Card*. Retrieved June 16, 2008, from http://www.nevadareportcard.com/

Nilsson, N. J. (2005). *Introduction to machine learning*. Retrieved April 20, 2007, from http://robotics.stanford.edu/people/nilsson/mlbook.html

Nippold, M. A. (2000). Language development during the adolescent years: Aspects of pragmatics, syntax, and semantics. *Topics in Language Disorders*, 15-28.

Nippold, M. A., Ward-Lonergan, J. M., & Fanning, J. L. (2005). Persuasive writing in children, adolescents, and adults: A study of syntactic, semantic, and pragmatic development. *Language, Speech, and Hearing Services in Schools, 36*, 125-138.

No Child Left Behind Act of 2001, Public Law 107-110, & 115 Stat. 1425. (2002). Retrieved March 30, 2007, from http://www.ed.gov/policy/elsec/leg/esea02/107-110.pdf

North Central Regional Educational Laboratory. (2004). *Scientifically based research: Understanding the No Child Left Behind Act of 2001*. Retrieved April 20, 2007, from http://www.ncrel.org/csri/tools/qkey7/qkey7.pdf

Olson, M. W., & Raffeld, P. (1987). The effects of written comments on the quality of student compositions and the learning of content. *Reading Psychology, 8*(4), 273-293.

Page, E. B. (1966). The imminence of Grading Essays by Computer. *Phi Delta Kappan, 47*, 238-243.

Page, E. B. (1994). Computer grading of student prose, using modern concepts and software. *Journal of Experimental Education, 62*(2), 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, New Jersey: Lawrence Erlbaum Associates.

Page, E. B., & Paulus, D. H. (1968). *The analysis of essays by computer. Final report* (No. ED028633). Washington, D.C.: Office of Education, Bureau of Research.

Page, E. B., & Petersen, N. S. (1995). The computer moves into essay grading. *Phi Delta Kappan, 76*(7).

Page, E. B., Poggio, J. P., & Keith, T. Z. (1997). *Computer analysis of student essays: Finding trait differences in student profile.* Paper presented at the Annual Meeting of the American Educational Research Association, Chicago, IL.

Parsad, B., & Jones, J. (2005). *Internet access in U.S. public schools and classrooms: 1994-2003* (No. NCES-2005-015). Washington, D.C.: U.S. Department of Education.

Phye, G. D., Robinson, D. H., & Levin, J. R. (Eds.). (2005). *Empirical methods for evaluating educational interventions.* San Diego: Elsevier, Inc.

Polio, C. G. (1997). Measures of linguistic accuracy in second language writing research. *Language Learning, 47*(1), 101-143.

Pritchard, R. J., & Honeycutt, R. L. (2006). The process approach to writing instruction: Examining its effectiveness. In C. A. MacArthur, S. Graham & J. Fitzgerald

(Eds.), *The Handbook of Writing Research* (pp. 275 - 290). New York: The Guilford Press.

Reid, S., & Findlay, G. (1986). Writer's workbench analysis of holistically scored essays. *Computers and Composition, 37*, 6-12.

Reingold, E., & Nightingale, J. *Artificial intelligence tutorial review*. Retrieved April 25, 2007, from http://psych.utoronto.ca/~reingold/courses/ai/Welcome.html

Rudner, L. (1992). *Reducing errors due to the use of judges*. Retrieved July 31, 2006, from http://pareonline.net/getvn.asp?v=3&n=3

Rudner, L., & Gagne, P. (2001). *An overview of three approaches to scoring written essays by computer*. Retrieved July 16, 2006, from http://pareonline.net/getvn.asp?v=7&n=26

Rudner, L., Garcia, V., & Welch, C. (2006). An evaluation of the IntelliMetric[SM] essay scoring system. *Journal of Technology, Learning, and Assessment, 4*(4).

Russell, M. (1999). *Testing on computers: A follow-up study comparing performance on computer and on paper*. Retrieved February 6, 2006, from http://epaa.asu.edu/epaa/v7n20/

Russell, M., Bebell, D., & Higgins, J. (2004). Laptop learning: A comparison of teaching and learning in upper elementary classrooms equipped with shared carts of laptops and permanent 1:1 laptops. *Journal of Educational Computing Research, 30*(4), 313-330.

Ruth, L., & Murphy, S. (1984). Designing topics for writing assessment: Problems of meaning. *College Composition and Communication, 35*(4), 410-422.

Santrock, J. W. (2005). *Adolescence* (Vol. 10th). New York: McGraw-Hill.

Schrum, L., Thompson, A., Sprague, D., Maddux, C., McAnear, A., Bell, L., et al. (2005). Advancing the field: Considering acceptable evidence in educational technology research. *Contemporary Issues in Technology and Teacher Education, 5*(3/4), 202-209.

Scott, C. M. (1988). Later language development: Ages 9 through19. In C. M. Scott (Ed.), *Spoken and written syntax* (pp. 49-95). Boston: College-Hill.

Shermis, M. D., & Burstein, J. (2003). *Automated essay scoring: A cross-disciplinary perspective*. Mahwah, New Jersey: Lawrence Erlbaum Associates.

Shermis, M. D., Burstein, J., & Leacock, C. (2006). Applications of computers in assessment and analysis of writing. In C. A. Fitzgerald, S. Graham & J. Fitzgerald (Eds.), *Handbook of Writing Research* (pp. 403 - 416). New York: Guilford Press.

Shermis, M. D., & Daniels, K. E. (2003). Norming and scaling for automated essay scoring. In M. D. Shermis & J. Burstein (Eds.), *Automated essay scoring: A cross-disciplinary perspective* (pp. 147-167). Mahwah, New Jersey: Lawrence Erlbaum Associates.

Shermis, M. D., Koch, C. M., Page, E. B., Keith, T. Z., & Harrington, S. (1999). *Trait ratings for automated essay grading*. Retrieved June 12, 2006, from http://www.eric.ed.gov/ERICDocs/data/ericdocs2/content_storage_01/0000000b/80/11/9d/38.pdf

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., Mzumara, H. R., Olson, J., & Harrington, S. (2001). On-line grading of student essays: PEG goes on the World Wide Web. *Assessment & Evaluation in Higher Education, 26*(3), 247-259.

Simmons, R. L. (2007). *Grammar bytes!* Retrieved March 15, 2007, from http://www.chompchomp.com/

Smerdon, B., Cronen, S., Lanahan, L., Anderson, J., Iannotti, N., & Angeles, J. (2000). *Teachers' tools for the 21st century: A report on teachers' use of technology. Statistical analysis report.* (No. NCES 2000-102). Washington, DC: U.S. Department of Education.

Smith, C., & Kiefer, K. (1983). *Using the Writer's Workbench programs at Colorado State University.* Paper presented at the 6th International Conference on Computers and the Humanities, North Carolina State University.

Smith, S. (1997). The genre of the end comment: Conventions in teacher responses to student writing. *College Composition and Communication, 48*(2), 249-268.

Sommers, N. (1982). Responding to student writing. *College Composition and Communication, 33*(2), 148-156.

Straub, R. (2000). The student, the text, and the classroom context: A case study of teacher response. *Assessing Writing, 7*(1), 23-55.

Stuebing, S., Celsi, J. G., & Consineau, L. K. (1998). *Modes of learning: The results of two interactive design workshops: Apple Computer, Inc.* Retrieved June 12, 2006, from http://images.apple.com/euro/pdfs/acotlibrary/rpt19.pdf

Valenti, S., Neri, F., & Cucchiarelli, A. (2003). An overview of current research on automated essay grading. *Journal of Information Technology Education, 2*, 319-330.

Ware, P., & Warschauer, M. (2006). Electronic feedback and second language writing. In K. Hyland & F. Hyland (Eds.), *Feedback and Second Language Writing*. Cambridge: Cambridge University Press.

Warschauer, M. (2006). *Laptops and literacy.* Paper presented at the Annual Meeting of the American Educational Research Association, San Francisco, CA.

Warschauer, M., & Ware, P. (2006). Automated writing evaluation: Defining the classroom research agenda. *Language Teaching Research, 10*(2), 1-24.

Waxman, H. C., Lin, M.-F., & Michko, G. M. (2003). *A meta-analysis of the effectiveness of teaching and learning with technology on student outcomes.* Retrieved June 5, 2006, from http://www.ncrel.org/tech/effects2/intro.htm

What Works Clearinghouse. (2006). *A trusted source of what works in education.* Retrieved May 5, 2006, from http://www.w-w-c.org/default.asp

Williams, R. (2001). *Automated essay grading: An evaluation of four conceptual models.* Paper presented at the Expanding Horizons in Teaching and Learning 10th Annual Teaching and Learning Forum, Perth, Australia.

Wolcott, W., & Legg, S. M. (1998). *An overview of writing assessment: Theory, research, and practice*. Urbana, Ill.: National Council of Teachers of English.

Wolfe-Quintero, Inagaki, S., & Kim, H.-Y. (1998). *Second language development in writing: Measures of fluency, accuracy, & complexity*. Honolulu, HI: University of Hawai'i Press.

Yagelski, R. P. (1995). The role of classroom context in the revision strategies of student writers. *Research in the Teaching of English, 29*(2), 216-238.

Yancy, K. B. (1999). Looking back as we look forward: Historicizing writing development. *College Composition and Communication, 50*(3), 483-503.

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

VITA

Graduate College
University of Nevada, Las Vegas

Kathie L. Frost

Local Address:
2030 S. Red Rock St.
Las Vegas, NV 89146

Degrees
Bachelor of Science
University of Arizona

Masters of Business Administration
University of Nevada, Las Vegas

Dissertation Title: The Effects of Automated Essay Scoring as a High School Classroom Intervention

Dissertation Examination Committee:
Chairperson: Dr. Randall Boone, Ph. D.
Committee Member: Dr. Greg Levitt, Ph. D.
Committee Member: Dr. Marilyn McKinney, Ph. D.
Committee Member: Dr. Kent Crippen, Ph. D.
Graduate Faculty Representative: Dr. Kyle Higgins, Ph. D.