5-2015

# Robo-Teaching? : Automated Essay Scoring and K-12 Writing Pedagogy

Swati Viren Chauhan
*Montclair State University*

MONTCLAIR STATE UNIVERSITY

Robo-Teaching?: Automated Essay Scoring and K-12 Writing Pedagogy

by

Swati Viren Chauhan

A Master's Thesis Submitted to the Faculty of

Montclair State University

In Partial Fulfillment of the Requirements

For the Degree of

Master of English

May 2015

The College of Humanities/Social Sciences

English Department

Thesis Committee:

[redacted]

Thesis Advisor: Emily J. Isaacs

[redacted]

Committee Member: Caroline E. Dadas

[redacted]

Committee Member: Laura E. Jones

Swati Viren Chauhan

English MA Thesis

Abstract

This paper will examine the current state, and future, of AES in secondary literacy education through a review of current research in the topic. An analysis of the history of assessment will seek to explain why AES systems have gained such popularity within high-stakes assessment, and how the use of AES in secondary education, high-stakes testing affects pedagogy. This paper will also look into reliability and validity issues that are presented when using AES as a form of scoring essays. Finally, this paper explores some ways that AES can be used effectively within the K-12 writing classroom, rather than solely in high stakes assessments. Ultimately, this paper will assert that AES is not the right tool for use in high-stakes assessments, but can be beneficial within the high school English classroom as a means of formative assessment.

# ROBO-TEACHING?: AUTOMATED ESSAY SCORING AND K-12 WRITING PEDAGOGY

A THESIS

Submitted in partial fulfillment of the requirements

For the degree of Master of English

by

SWATI VIREN CHAUHAN

Montclair State University

Montclair, NJ

May 2015

TABLE OF CONTENTS

## 1. Introduction and Background

Technology is an ever-present force in the 21st century; it affects almost every aspect of our lives, from the mundane, daily tasks to life-changing decisions. Technology has also found a new home within almost every aspect of the educational system. As a high school English teacher, I see this technology incorporation in every facet of my job—grades and attendance are done online using a system called Power School, I have an interactive SMARTBoard and a document camera in my classroom, I keep and update my own teacher website daily, I tweet homework, and I have taught a Freshman Honors English class entirely digitally using iPads. Thus, it is no surprise then that technology has also found its way into that once humanistic space of writing assessment. Prior to 2012, I assumed it was mostly limited to usage within higher education as a form of placement for students into writing courses; I myself took the Accuplacer test as a freshman in college in order to place out of Writing 101. However, Automated Essay Scoring, or AES, has also begun to gain popularity in the worlds of elementary and secondary education.

With the implementation of the national Common Core State Standards (CCSS), a new set of assessments was needed to test students' learning based on these standards. The Partnership for Assessment of Readiness for College and Careers (PARCC) is one of two consortia (the other being Smarter Balanced) that is creating these new K-12 assessments, of which the state of New Jersey is a member. As part of the incoming PARCC assessments, the exams are now to be completed on computers; the Educational Testing Service (ETS), the company that created some of the most popular Automated

Essay Scoring (AES) systems in usage, also notes that AES will most likely be used for response scoring of writing, stating:

> The interest in automated scoring of essays is not new and has recently received additional attention from two federally supported consortia, PARCC and Smarter Balanced, which intend to incorporate automated scoring into their Common Core state assessments planned for 2014. (Zhang 1)

As a high school English teacher in a New Jersey public high school who teaches and assesses student writing constantly and consistently, and who will also see the implementation of the PARCC assessments in the immediate future, I am interested in, and worried about, this influx of AES into writing assessment and pedagogy.

While I became interested in the idea of AES though my Writing Studies courses, my interest in pursuing further research into AES was spurred by some experiences I had in teaching over the last year. The school district I work for was one that was chosen by the state to field test the English Language Arts/Literacy section of the PARCC test for the 9th graders in the 2013-2014 school year. After administering the exam to my freshmen, my students and I had a candid conversation about the test; we discussed how they felt taking a test on a computer versus on paper, how they felt about the digital tools available to them on the test and if they utilized them, and how they felt about writing their essay on the computer. When I offhandedly remarked that the essays would perhaps be scored by a machine rather than a human, the students had a collective moment of shock, horror, and outright indignation at the fact that their writing would be scored essentially by a computer system; they had assumed that there would be a teacher, a human, on the other end to evaluate and score the writing on which they had worked so

hard. This reaction from my students prompted me to do some further, informal, polling on how well known AES is in the context of the secondary education realm.

While I had expected that high school freshmen would be unaware about the existence of AES, I did not expect that professionals in the educational field would be as unaware about it. My colleagues in the English department and various district administrators also had the exact same reaction as the students to my statement that the essays on PARCC would most likely be scored by an AES system. Many of the teachers were not familiar with AES, and definitely were not aware that PARCC would utilize AES for their K-12 assessments, and unanimously, my colleagues were horrified by AES's infiltration of writing assessment. It seems that all sides involved with education on the local level-- teachers, students, and administration-- are unschooled in, and fearful of AES, as they believed that there would be a person, a human, in charge of looking over students' writing.

The reactions of the teachers and administrators are understandable; as of 2015, there is a forward movement within education to tie student standardized test scores to teacher evaluations, as a result of federal waivers from No Child Left Behind (NCLB), called Race to the Top, given to states by the US Education Department. If a state received a NCLB waiver and enrolled in Race to the Top in 2012 (New Jersey is one such state), then the rules of the waiver require the state to pilot new teacher evaluation systems in which student growth on assessments must become a significant portion of the teacher's evaluation. The timeline for this has been rapid-fire; under Race to the Top, "States must then implement the new evaluations statewide in 2014-15, when the common assessments are expected to begin and test scores are expected to plummet. The

new teacher-evaluation systems then must be tied to personnel decisions and consequences in 2015-16, just a year into the new tests" (McNeil). Thus, not only has the method for how teachers are evaluated and granted tenure changed drastically under Race for the Top, but there are new content standards and new assessments that are tied to that new evaluation system, all occurring simultaneously[1].

These movements have converged in a problematic way, generating for me, a central and important question: If education is moving towards more high-stakes assessments that require writing and use AES, and these assessments are tied to teacher evaluations, will the way we as educators teach writing change? If so, how? My thesis will examine the current state, and future, of AES in secondary literacy education through a review of current research in the topic. My history and analysis will seek to explain why AES systems have gained such popularity within high stakes assessment, and how the use of AES in secondary education, high-stakes testing affects pedagogy. I will also look into reliability and validity issues that are presented when using AES as a form of scoring essays. Finally, I will also outline some ways that AES can be used effectively within the classroom, rather than solely in high-stakes assessments. Ultimately, I will

---

[1] According to the New Jersey Education Association (NJEA): As of 2012, the time period to achieve tenure has changed from three years to four. Teachers must be evaluated and rated either as "effective" or "highly effective" on their summative evaluations for the three years before tenure is achieved. (http://www.njea.org/njea-media/pdf/TenureLawQ-A_2012.pdf?1419702298285)

Evaluations are broken down into the following percentages as per the New Jersey Department of Education: 10% of a teacher's overall evaluation rating is based on Student Growth Percentile (SGP) data from state standardized assessments. 20% is based on Student Growth Objective (SGO) data from one to two measures that teachers set with the approval of their principals. 70% is based on classroom observations. (http://www.nj.gov/education/AchieveNJ/intro/guide.pdf)

assert that AES is not the right tool for use in high stakes assessment, but can be beneficial within the high school English classroom as a means of formative assessment.

## 2. The Rising Popularity of AES in High Stakes Assessment

Standardized testing has grown exponentially across all levels of education, from elementary school placement testing to professional licensure exams. Before delving into the particulars of Automated Essay Scoring, it is important to look back at the history of standardized testing that has brought writing assessment to where it is today. A panel of experts on assessment, including Marguerite M. Clarke and George F. Madaus, look back at the history of testing in their 1999 article "Retrospective on educational testing and assessment in the 20th century." They state that one of the first forms of large-scale testing, which quickly and efficiently tested many people at once (which is the forefather to standardized testing), was the Army's intelligence tests for classifying recruited soldiers during World War I. Large-scale testing then gained traction outside of the military (due to its efficiency), and began an influx into public schools for grouping students by ability into various tracks. The key to this movement into large-scale testing was the advent of the multiple-choice item in 1914 by Frederick J. Kelly; more items could be administered within in short time period, and they could be scored quickly and objectively. These first large-scale assessments did not include any type of constructed-response (written) section, a fact that was boasted about by the test manufacturers. Testing continued to boom in the education system in the United States, with advances in

technology (such as scanners) and educational reforms spurring it on. (Clarke et al 162-163)

In the 1970's, there began a movement towards large-scale assessments at the state, national, and international levels, in which multiple-choice items were most popular. The 1970's also saw the movement towards the testing of "basic skills," in which multiple-choice testing "proliferated in the form of state-mandated minimum competency" (Clarke et al 164). The 1970's also saw developments in the collection and analysis of data from large-scale testing, such as matrix sampling and item response theory, which helped create the push for large-scale testing. Then, in the 1980's, came the push in the US for states to mandate academic content standards and assessments aligned with these standards. It was during this movement towards standard-aligned assessments that the shift from multiple-choice to performance tasks on tests occurred, "performance" being a term here that asks for the test taker to showcase knowledge through a task, rather than through choosing an answer from multiple choices. Supporters of performance tasks on assessments, "believed that this more open-ended, complex way of assessing students' knowledge could defeat negative test-preparation effects associated with multiple-choice tests, would give teachers clear models of acceptable outcomes, measure higher- order skills, and would lay bare examinees' thinking processes" (Clarke et al 168). It was this consideration that has pushed high-stakes assessments to what they have become today--a mix of both multiple choice items and performance tasks. (Clarke et al 164-168)

Most exams now require a performance of knowledge along with the requisite multiple-choice items. Proponents of performance-tasks on assessments suggest that constructing responses rather than simply selecting responses from a predetermined set

(as in multiple choice tasks) is a more authentic means to gauge learning as rarely in life is one presented with a set of predefined choices (Darling-Hammond, ch. 1). Written response is one such performance that is being demanded, with constructed-response items being the most prevalent form of this performance task. Exams such as the Graduate Record Examination (GRE), Test of English as a Foreign Language (TOEFL), and most recently the SAT, have incorporated these constructed-response items. Chaitanya Ramineni and David M. Williamson discuss this move towards the incorporation of constructed-response items in their article, "Automated Essay Scoring: Psychometric Guidelines and Practices," stating:

> Constructed-response items, in which an examinee produces a response in a more naturalistic manner rather than selecting a response from a set of pre-defined options (as in multiple-choice items), represent an effort to provide some connection between the naturalistic characteristics of real world performance and the controlled circumstances of assessment. (26)

Constructed-response performance tasks have the ability to more wholly gauge student understanding through a demonstration that requires the student to combine many major skills and knowledge sets into one cohesive writing task. Constructed-response can take a myriad of forms from short answer responses of a few sentences, to essays, to creating a laboratory report based on a hands-on investigation (Darling-Hammond, ch.1).

However, while these writing tasks are obviously beneficial to showcase the strengths and weaknesses of each individual test-taker, they are not without their own set of weaknesses. They engage students, and can provide a real-world example of usage of a particular skill, and allow for unconstrained responses on the part of the test-taker.

However, as Darling-Hammond notes, "That freedom permits performance tasks to include more varied, complex, and novel stimuli than are typically used in multiple-choice assessments" (Part One). Another issue is that the directions for the tasks must be explicit enough to be understood by various test-takers so that they may have an understanding of what they are being asked to do. Therefore, the language of the directions or queries must be specific enough to generate a similar understanding amongst test-takers, but not so vague as to be misinterpreted (Darling-Hammond, ch. 1).

One of the major concerns when it comes to effectively utilizing constructed-response tasks is the task of scoring of these responses reliably and within a specific time frame. Constructed-response tasks, in order to be considered a reliable form of assessment, "must prompt responses that can be scored according to a clear set of standards," most often through a rubric (Darling-Hammond, Part One). Multiple-choice was popular in its onset because of its efficiency in being scored, but performance tasks such as constructed-response bring about their own difficulties in terms of their grading because, as Ramineni suggests,

> When compared to their multiple-choice counterparts, constructed-response items take longer to administer and provide less psychometric information per unit time of assessment. They also reduce the reliability of test scores (holding total test time constant) and delay score reporting. They also tend to be more expensive to develop, administer and score than multiple-choice items. (26)

Much of the challenge of using constructed-response items stems from the use of human raters. It takes considerable effort and expense to recruit enough human raters in order to score large volumes of responses in a timely manner, to train such raters to norm

themselves into scoring based on the rubric, and to monitor the human raters during scoring to ensure consistency of the scores. Therefore, it is common for high stakes assessments that use constructed-response tasks to have two human raters score each task, and if there is a notable discrepancy between the scores to have additional human raters evaluate the discrepancy (Ramineni 27). The human raters are generally considered experts within the field upon which the assessment is based; for example, the Advanced Placement (AP) exams use teachers who teach the AP-level courses, and have a Master's in the field for that specific subject area. Human raters must be paid for their time, and also must be trained to score writing, all of which costs more than using a machine to do the same work. This makes AES a logical option for writing assessment, as Condon notes, "Large-scale testing is big business, with huge profits in the offing, especially if tests can be constrained to the point that computers can provide a large portion of the labor of scoring essays" (Condon 101). It is ironic then, when the history of testing is tracked to this current state. It was the critique of multiple-choice testing that led to the advent of constructed-response items, which in turn resulted in the usage of AES for scoring methods. AES now brings with it many of the same criticisms as multiple-choice tests, which originally had that created the move towards constructed-response in the first place.

It is at this crossroads that Automated Essay Scoring emerges as an appealing choice for assessing constructed-responses as, "Automated essay scoring (AES) systems hold the potential for greater use of essays in assessment while also maintaining the reliability of scoring and the timeliness of score reporting desired for large-scale

assessment" (Ramineni 26). However, research is split on the validity of using AES to score constructed responses on these standardized tests versus using human scorers.

What does not seem to be split, however, is the visceral reaction of people who are not familiar with AES systems when they consider the use of machines to score writing. In my experience as an English teacher and Writing Studies student, given the conversations I have had with people on all levels of education-- students, teachers, administrators, professors-- the reaction has been negative. People do not like the idea of machines scoring human because it seems to fly in the face of the basic idea that writing at its base is a humanistic thing-- we write to communicate to each other at the human level. It seems to be something that separates us from animals and machines alike. If machines are able to score writing as a human can, then we have given something human up.

Human scorers have many points in their favor as assessors for constructed-response tasks--indeed they seem to be the obvious choice. Mo Zhang, a researcher for the Educational Testing Service (ETS), discusses how human scorers are normed to follow a rubric to gauge the quality of a piece of writing. Some of the strengths that human raters bring over AES are that they can:

> (a) cognitively process the information given in a text, (b) connect it with their prior knowledge, and (c) based on their understanding of the content, make a judgment on the quality of the text. Trained human raters are able to recognize and appreciate a writer's creativity and style (e.g., artistic, ironic, rhetorical), as well as evaluate the relevance of an essay's content to the prompt. A human rater

can also judge an examinee's critical thinking skills, including the quality of the argumentation and the factual correctness of the claims made in the essay. (2)

Essentially, human scorers' biggest strength comes from their ability to assess the higher-level, humanistic elements of writing that machines may not be able to fully judge and score. However, there are issues with using human scorers as well. Zhang refutes many of the humanistic positives of using real scorers when he outlines the limitations of human scorers by stating that humans carry a set of cognitive biases that could undermine the validity of scores as these biases are near impossible to quantify (3). Some of these biases include the halo effect (when a rater takes one characteristic of an essay and generalizes it to the entire essay), inconsistency (erratic grading due to different interpretations of the grading rubric), or general scoring fatigue. As these biases negate any sense of reliability for the given assessment, human limitations for assessing writing is evident.

It is because of this human bias that Zhang believes the tide turned in favor of using automated scoring. Zhang's list of positives of using AES for writing assessment generally comes down to one main strength that automated scoring has over humans: reliability. Zhang states,

> The primary strength of automated scoring compared to human scoring lies in its efficiency, absolute consistency in applying the same evaluation criteria across essay submissions and over time, as well as its ability to provide fine-grained, instantaneous feedback. Computers are neither influenced by external factors (e.g., deadlines) nor emotionally attached to an essay. Computers are not biased

> by their stereotypes or preconceptions of a group of examinees. Automated scoring can therefore achieve greater objectivity than human scoring. (3)

Machines simply will not fall prey to the same issues in reliability that human scorers can because they alone are wholly and completely objective in their scoring.

So, these are all marks in favor of AES. But, will an AES system actually be as valid as a human scorer in every aspect? One issue can be seen as both favoring and refuting the use of AES in writing assessment: AES machine scoring requirements are modeled upon those of human raters. Paul Deane states, "The limitations of an AES engine are a direct consequence of the way that engine is built. For instance, AES models are trained against the judgments of human raters, and might therefore replicate biases present in the original human raters" (15). We cannot know what kinds of biases are present in the machine and therefore built into the scoring. And if biases are present, they negate the reliability of the assessment. Some proponents of AES believe that the use of "experts" upon which to model the machine's scoring algorithms is the answer to this issue of bias, stating,

> By using a representative committee of experts it is assumed that undesirable aspects of human scoring (e.g., bias, halo effects, etc.) are eliminated in much the same way classical test theory assumes the random error component of test scores cancels out to produce the 'true score' over repeated administrations. (Williamson, Bejar and Hone 161)

Zhang acknowledges that biases in human scorers used for development in algorithms exist, but offers no commentary on this issue other than echoing Williamson, Bejar and

Hone's point about using experts for development of scoring algorithms without offering any further support (8).

Advocates of AES cite numerous studies that demonstrate that various AES systems and human scorers have the same reliability- that the algorithms in the systems have come to a point where they essentially assess writing just as a human scorer does, which is another reason why AES is such an attractive option for utilization in high-stakes assessments. Ideally, in order to win the argument that AES systems should be used in high-stakes assessments, the AES systems would have to be proven to quickly and cheaply score writing in the same way that human raters would, but without the biases. A recent study conducted in 2014 by Mark Shermis, who is an outspoken advocate of AES and an important figure in Writing Studies, used student essays written for the tests given by the two consortia, PARCC and Smarter Balanced, and scored them using human scorers and AES (Shermis 57). The implications of this study may drive the discussion as to whether AES should indeed be used when scoring writing in the K-12 high-stakes assessment arena. The results of the experiment indicated that, "in a high stakes testing environment machine-predicted scores came close to matching the distributional and agreement characteristics of scores assigned by humans;" that is, the scores between the AES systems and human raters were similar, seemingly lending AES the winning draw (75). However, Shermis also acknowledges that his study is limited in scope, and is not meant to be used as a deciding factor in the usage of AES in K-12 educational writing assessments, but is meant to showcase "the extent to which the current state-of-the-art of machine scoring can replicate the scores of human graders on existing writing prompts, regardless of the extent to which these are thought to be optimal

measures of the construct of writing" (Shermis 56). Therefore, while this study seems to originally serve as a mark in favor of utilizing AES in high-stakes assessment, the limitations that Shermis makes note of brings to light the bigger issues in AES that go above and beyond any machine-- the difference between the writing education and writing assessment. Shermis makes it a point to note that "The construct of writing is a rich, nuanced, and sophisticated domain," and that summative writing assessments are generally short in length, timed, and do not allow for drafts (Shermis 56). He maintains that "there are quite legitimate concerns that this common practice of summative assessment does not adequately reflect the construct of writing as it is taught in the K-12 American educational system," and simply distances himself from this by stating that this study is not meant to definitively answer if AES is the answer to high-stakes assessment, but only shows that AES and human scores have come to a point in high-stakes assessments where they agree (Shermis 56). Shermis in the end notes that in order for the results of the study to be taken seriously, further research on validity must be done (75).

This issue on the validity of human scorers versus AES is an important one to look at, but one that has a serious flaw in terms of research itself, in that is there is a lack of research done on human versus machine scoring for writing assessment that has not been done in a biased manner. Indeed, in doing the research for this paper, I was surprised to see that the Educational Testing Service and Pearson, corporations that coincidentally create the more popular AES systems for high stakes writing assessment, have ties to many of the articles that are in favor of using AES over human raters. The potential problem here is that this research can be taken as gospel to push the use of AES, but there is a conflict of interest here-- those backing the studies are the ones who stand

the most to gain on a monetary level. In terms of the PARCC and Smarter Balanced assessments, the monetary value for pushing the use of AES is a significant one as the estimates for the cost of assessing the Common Core Standards run in the billions of dollars (Condon 101). One article published by the *Journal of Technology, Learning, and Assessment* notes this discrepancy in the amount of research done with the backing of these large companies that create AES systems, stating,

> So far, very few studies have been conducted by independent researchers and users of AES. Institutions that are in the process of making decisions about whether or not to adopt AES for the benefit of its efficiency are left with little impartial research on which to base their decisions. (Wang and Brown 5)

And yet the use of AES for high stakes assessment in K-12 education is happening in the immediate future, without the backing of necessary and unbiased research needed to intelligently make the decision to use it, giving it the red flag from many members of the writing studies community. In 2013, The National Council of Teachers of English (NCTE) put forward their position statement on machine scoring, setting themselves against it by stating a host of issues with an attached bibliography of research to support their points ("NCTE Position Statement on Machine Scoring"). In the same vein, many researchers in the field of writing studies under the moniker of "Professionals Against Machine Scoring Of Student Essays In High-Stakes Assessment" put forth a petition with attached bibliography to put an end to the usage of AES on high-stakes assessments, stating "current machine scoring of essays is not defensible, even when procedures pair human and computer raters. It should not be used in any decision affecting a person's life or livelihood and should be discontinued for all large-scale assessment purposes"

("Human Readers"). So while proponents of AES cite its validity, the detractors do legitimately question AES and how valid it can be as a means of assessing and are actively seeking an end to its usage the realm of assessment.

3. Gaming the System

One major consideration that comes into play when using AES in standardized testing is that the test takers may be able to manipulate the computer into a higher score by "gaming" written responses to incorporate specific requirements, known to be searched for by the machine, for a higher score, as "automated methods may be unduly influenced by extraneous features of examinees' writing (and therefore ''tricked'' into awarding higher-than- deserved scores)" (Powers 105). The NCTE echoes this point about gaming in their position statement against the use of AES, and take it one step further by stating that the use of AES not only weakens assessments, but also separates students "not on the basis of writing ability but on whether they know and can use machine-tricking strategies" ("NCTE Position Statement on Machine Scoring"). If AES systems can be gamed, the real issue that is raised is whether the use of AES weakens the writing construct itself. Picture a room full of bright, motivated students, who are highly aware that their scores on the writing assessment in front of them may determine whether they graduate high school or track their level of class (as the PARCC assessments will do in the future), or determine where they will go to college (as the SATs do now). Do you believe these students would change the way they write to satisfy the requirements of the algorithm of the AES machine scoring their writing? Would they play up certain aspects

of writing that they knew the AES system would score them more highly on? And do you believe that this negates the point of writing to begin with? These are only some of the issues that gaming raises.

Anne Herrington and Charles Moran further discuss the concept of gaming in their 2012 article, "Writing to a Machine Is Not Writing At All." Herrington and Moran focus on the Educational Testing Service's (ETS) *Criterion* AES system and outline the various ways they played around with *Criterion's* holistic scoring design after they were able to use a trial version of the AES, specifically looking at its application in the classroom rather than as a true assessor. Herrington and Moran look at the feedback provided by *Criterion* on student essays, and changed the essays as per the feedback of the AES system in order to "game" *Criterion* into a giving a higher score. While limited in terms of truly showcasing the concept of gaming, one of the main conclusions that Herrington and Moran draw as a result of their "experiment" is that having students write to AES is an exercise in futility for the students because they know they are writing to an imaginary audience, which makes them feel as if their individual writing does not matter; they are simply writing to get the best scores from the AES, not writing to showcase their own personal strengths. As Herrington and Moran note of one student in their study, "Her focus was not on what she was aiming to get across--indeed, the revised essay did not represent her view, rather on adding the favored structural features" (230). One result of this type of "writing for nobody" is that, "*Criterion* and its competitors 'turn writing into academic gamemanship' (White, 1969, p. 168), a game of learning to write the kind of essays that are used for ETS testing programs as they are the source for the prompts used by *Criterion* in order to provide the norming capability for the holistic ratings" (229).

This "gamemanship" has a wide range of consequences, including breaking writing down into a beatable assessment where one only needs to understand the limitations of the AES system to conquer it-- an issue that can be easily exploited by teachers and students.

In 2001, Powers, Burstein, Chodorow, Fowles and Kukich conducted an experiment that showcased this issue by inviting various writing experts, as well other professionals well-versed in the field of writing, to compose GRE writing assessment essays with the intention of undermining *e-rater* into scores that were either higher or lower than deserved, comparing these scores to those of human raters. The results of the experiment proved that it was in fact possible to easily trick *e-rater* into scoring higher than deserved. The study discusses the essay that showed the highest score discrepancy between the e-rater and human scorers:

> The clear ''winner'' in our contest was an issue essay submitted by a professor of computational linguistics. His principal strategy was simply to write several paragraphs and to repeat them (37 times, in fact!). This strategy did indeed fool e-rater, resulting in a maximum discrepancy of 5 points. E-rater assigned the essay the highest possible score (6), while both study readers awarded it the lowest possible score (1). (Powers 111)

The results of the study corroborated that it was in fact possible to trick that particular incarnation of *e-rater* into a higher score. It is important to note that this study is somewhat limited; having only highly trained and education professionals in the writing community try to game an AES system is a little unfair as they obviously are at an advantage. However, consider the earlier scenario of the room full of bright students with the knowledge of gaming the writing assessment in front of them. Who taught them the

methods for gaming the system? Could is possibly be the teachers, the professionals, who also have a stake in the success of their students?

Powers, Burstein, Chodorow, Fowles and Kukich's 2001 study ends with a note that the algorithms of the *e-rater* machine was subsequently fixed to be able to identify "off-topic" responses and score them appropriately because of their study (116). However, a 2012 experiment conducted by Les Perelman, MIT director of writing and outspoken critic of AES, showed that despite the fix from the 2001 study, *e-rater* can still be effectively gamed. Perelman studied the algorithms and biases of the *e-rater* system, and created a long essay that was completely on topic— but filled with nonsensical sentences and complete fabrication of facts. This essay was given a perfect score of 6, while a well-written and well-argued essay scored a 5 (Winerip). Perelman is well known in the Writing Studies community for showcasing methods of gaming AES systems because he believes that AES systems demean the construct of writing, which "cannot be judged like the answer to a math problem or GPS directions" ("Critique" 2). Perelman's successful gaming experiments seem to bolster the critique that AES systems break writing down to its most formulaic parts, thereby negating the true purpose of assessing writing.

In another study done in 2014, the researchers again used the *e-rater* AES as a means to study the vulnerabilities in AES as far as gaming. In this study, the participants used what the researchers refer to as a "construct-irrelevant response strategy" (CIRS); in this case the CIRS used is a word substitution (Bejar et al 49). The study notes that *e-rater* uses "the frequency of words and their length as indicators of writing quality" (Bejar et al 49). Therefore, the participants in this study attempted to game the AES

system simply by substituting common words with longer and more infrequent words in their writing, with the words being completely nonsensical to their placement. The study provides a table of words targeted for substitution in this study, such as switching out "ask" for "enquire," or "salary" for "remuneration" (Bejar et al 53). The results showed that the *e-rater* was fooled into giving a higher score through this methodology; "there is a tendency for the e-rater score to increase as a result of the substitution strategy. For example, 47 essays (2%) that received a "5", received a "6" after substitution" (Bejar et al 52). While the researchers (writing for the Educational Testing Service) state that this study was conducted on purpose eliminating other features that would effectively have lowered the score of the essay, such as argumentation, the fact remains that the AES was once again tricked into scoring a piece of writing highly by means of a gaming method (Bejar et al 58).

The concept of "gaming" AES is an important factor to consider in how students interact with a piece of writing when they know it is being scored by a machine. In both studies discussed, those who were "gaming" the AES were professionals in their fields, or students who had many years of writing experience. In the realm of K-12 education, students will not have the knowledge or expertise in writing to significantly "game" the system in the manner in which the participants in the listed studies were able, at least not right away. However, as the Common Core State Standards assessments go on and the tasks become clear, students will eventually be able to glean some tips and tricks for manipulating the system. Ramineni and Williamson's article notes these various issues, stating "If examinees are motivated (e.g. test scores matter to them) and can improve AES scores by employing test taking strategies that would be similarly effective with

human scoring then such behavior can disrupt the intended appropriate use of AES" (36). It is important to add that not only would it disrupt the appropriate use of AES, but would also disrupt the point of writing assessment, and writing itself.

As AES becomes more prevalent in K-12 education, students will be introduced to the concepts surrounding AES and may be able to utilize them in their writing if they are aware that AES will be used. AES systems are already being marketed toward the K-12 sector; systems such as ETS's *Criterion* and Vantage Learning's *My Access!* are marketed towards use in the secondary literacy classroom. Beth Ann Rothermel notes that these K-12 targeted AES systems are, "reaching more directly into classrooms to shape the learning and teaching process," thereby seemingly influencing writing at the basic classroom level (Rothermel 199). Indeed, as *Criterion* is used most prevalently at the higher education level, its influx into the K-12 market seems to show a vertical integration of both the system and the way the system dictates writing pedagogy in the classroom, vertical integration here meaning that the systems are used throughout all levels, bringing in business for the creators. Take ETS's *Criterion* system, which is marketed towards the classroom first, but is also marketed towards usage in high-stakes assessment by using the AES system *e-rater* to provide a holistic score. It would make sense that the classroom-based AES system will prioritize all of the features, and use the same if not a similar algorithm as the high-stakes version of the system, allowing students and teachers alike the opportunity to practice "gaming" writing to make it more in tune with the AES system's requirements for high scores before sitting for the assessment. This issue, known as "washback," will be discussed in the next section in more detail. At

its crux, question in issue is this: how will AES's algorithm-based writing assessment affect writing pedagogy in the K-12 literacy classroom?

## 4. Effects of High Stakes AES on Writing Pedagogy

High-stakes assessments are labeled as "high-stakes" because their results are scrutinized and lead to educational decisions not only for the individual student taking the exam, but also affect the teachers, administrators, curriculum, districts and overall educational policies. Under No Child Left Behind (NCLB) for example, these high-stakes tests had the power to determine whether or not a school would be closed down. This power that high-stakes assessment has over all members of the local and higher levels of education is what makes it potentially so influential upon pedagogy.

An issue that has been long prevalent in high-stakes assessment is that of the impact of the assessment on teaching and learning within the classroom, which is an issue that needs revisiting with the upcoming new assessments, which will be unified across the nation for the first time in the United States education system's history. High-stakes testing impacts on the student are often the only ones considered, as the students are the ones being tested. However, for the first time on a national stage, teachers will also be directly impacted by these high-stakes assessments as the results of the tests will be tied to teacher's names as data, and this data may determine teacher pay and tenure in the future. As noted in the introduction, as of the 2014-2015 school year, student scores on the Common Core State Standards assessments through PARCC and Smarter Balanced will provide a percentage of the teacher's overall yearly evaluation.

These newly created assessments will require students to write constructed responses on a computer. The type of writing required on the PARCC and Smarter Balanced assessments are new to the students as well as to the teachers due to the timing of the implementation of the new writing standards, as noted,

> The Race to the Top Assessment Program consortia plan to require writing from sources. They seek to measure a rich writing construct that includes strength of argumentation and clarity and accuracy of explanation. At the same time, their plans indicate the intention to rely heavily on automated scoring technologies. (Deane et al 2)

Along with the new assessments and the new types of writing being assessed on the high-stakes level, AES is also now being introduced to the K-12 assessment. So how will this cross between new high stakes assessments that utilize AES and new evaluation methods that tie test scores to teacher performance grades affect the way teachers teach literacy in the K-12 classroom?

Les Perelman warns that the way these high-stakes assessments are graded will change U.S. K-12 education because of this convergence of key issues, especially the usage of AES, stating, "The machines' huge bias toward word count may encourage teachers to emphasize bloated and vapid prose. They may focus instruction on daily on-demand writing exercises to increase student output and fluency at the expense of critical thinking and frequent and extensive revision of writing" ("When 'state of the art'" 110). Writing pedagogy will change at the classroom level because the stakes of these tests are higher on teachers than they have ever been before, which "may lead to anxiety or other negative effects in the classroom. Tests which are helpful to decision makers (admissions

officers, educational administrators) are not necessarily helpful to teachers and students" (Wall 500). One main consideration that is that of the washback effect, which is an issue that has cut across each incarnation of high stakes assessment and is still a relevant issue today.

Washback, sometimes referred to as backwash, can be generally understood as the effect of an examination on teaching and learning. Not all researchers, however, have agreed to its definition, and washback can be considered positive in some research. Alderson and Wall restricted the use of the term "washback" to "classroom behaviors of teachers and learners rather than the nature of printed and other pedagogic material" (118). They would also consider washback to be what teachers and learners do that "they would not necessarily otherwise do" (117). For the purposes of this paper, washback will be defined by Alderson and Wall's connotation of negative washback.

Ramineni and Williamson also discuss the washback effect in which "an assessment represents an expected level of performance as an outcome of education and educators are held accountable for student performance" (36). Ramineni and Williamson here are discussing washback as it is tied to high-stakes. As high stakes assessment scores are largely becoming tied to teacher efficacy and evaluations, the washback effect becomes important as teachers may "teach to the test" in order to maximize student outcomes because, "when stakes are attached to the scores, teachers will feel pressure to focus narrowly on improving performance on specific tasks" (Darling-Hammond, Part One). In the classroom, washback may take the form of allocating a substantial amount of instructional towards tested subjects versus untested subjects, drilling through practice tests, or devoting more resources towards students who score at the cutoff point for

proficiency on the tests. In discussing high-stakes testing, Brian Jacob, who studied the effects of accountability on student achievement through an economic lens, asserts,

> The notion behind test-based accountability is that it will provide students, teachers and administrators an incentive to work harder as well as help identify struggling students and schools. Advocates claim that accountability will improve student performance by raising motivation, increasing parent involvement and improving curriculum and pedagogy. Economic theory, however, suggests that high-powered incentives may lead to unwanted distortions...Based on similar logic, critics have argued that such policies will cause teachers to shift resources away from low-stakes subjects, neglect infra-marginal students and ignore critical aspects of learning that are not explicitly tested. (Jacob 762)

Jacob here compares educational accountability measures, in particular the tying of test scores to teachers, to incentives within economics. He agrees that accountability can have positive effects such as making teachers and administrators works harder at the early identification of struggling students and other things, but also notes that as occurs in the use of incentives at the economic level, some unwanted effects can also occur in effort to make the incentive, one of which is negative washback.

Interestingly enough, there is little research that has been done on washback effect in terms of high-stakes assessments; most of the research on washback revolves around washback on curriculum rather than on pedagogy, and that in countries other than the United States. This is most likely due to the fact that it is only right now, in this moment within United States education, that the stakes are so high for teachers in the classroom with regards to student standardized test scores. Brian Jacob's 2004 article on the impact

of an accountability policy implemented in the Chicago Public Schools system in the years 1997-1999, is the only research I have found that comes from the United States that touches upon washback effects in the realm of high-stakes testing, and even then it is important to remember that Jacobs looks at the entire study through the economic lens regarding incentives. Jacobs set out to answer three specific questions about accountability in the district: "(1) Does high-stakes testing increase student achievement? (2) If so, what factors are driving the improvements in performance?...(3) Do teachers and administrators respond strategically to high- stakes testing?" (762). Through an analysis of both student-level and administrative data, Jacobs found that the answer to his first questions was that yes, student achievement on the high-stakes exam increased in the fields of math and reading directly following the introduction of the accountability measure (763). This seems to be a resounding plus in the column for high-stakes assessment causing positive washback--student scores definitively increased. However, further research by Jacobs uncovered that on a low-stakes, state test, the accountability policy did not increase student performance. What does this mean? Jacobs states:

> Achievement gains may have been driven by an increase in skill emphasized predominantly on the high-stakes exam. An item-level analysis provides additional evidence that achievement gains were driven in large part by increases in test-specific skills and student effort. Finally, the results suggest that teachers responded strategically to the incentives along a variety of dimensions—by increasing special education placements, preemptively retaining students and substituting away from low-stakes subjects like science and social studies. (763)

What Jacobs notes is negative washback on pedagogy within the classroom due to high-stakes assessment. The teachers strategically responded to the high-stakes assessment policy through a number of undesired methods, "including a narrowing of teaching to focus on the set of skills emphasized on the high-stakes test," which for this study seemed to be the use of drilling and practice tests (Jacobs 763).

While these issues hold true for any form of high-stakes assessment, the use of AES in the K-12 realm is a new development, making the implications of washback on curriculum and pedagogy significant, especially in terms of teaching composition. Paul Deane states:

> If a student's first writing-experience at an institution is writing to a machine, for instance, this sends a message: writing at this institution is not valued as human communication—and this in turn reduces the validity of the assessment…And finally, if high schools see themselves as preparing students for college writing, and if college writing becomes to any degree machine-scored, high schools will begin to prepare their students to write for machines. (8)

It should be noted here that Paul Deane writes for the ETS, an organization that, as mentioned earlier, produces AES systems. It seems here that Deane seems to be against the usage of AES on high-stakes assessments. Indeed, Deane is even more explicit in this point of view, stating, "We understand that machine-scoring programs are under consideration not just for the scoring of placement tests, but for responding to student writing in writing centers and as exit tests. We oppose the use of machine-scored writing in the assessment of writing" (Deane 8). However, Deane goes on in this article to articulate the many critiques relating to AES to lead to an argument for another, very

particular AES system. He ultimately introduces the features of the *e-rater* AES system to showcase that not all AES systems deserve the criticisms they receive. Deane seems to be taking Mark Shermis's route by essentially stating that while currently AES systems have a host of issues that makes them unusable for high-stakes assessments right now, they do have benefits that should be explored, as he illustrates by discussing the virtues of the *e-rater* system. The *e-rater* system is sold by the ETS, which illustrates again the lack of unbiased research being done on the issue, as Deane undercuts his own points in order to showcase the selling points of his company's system.

However, Deane does express an important point in his article--the unintended effects on the individual student as a learner, and on the teacher's pedagogy, is a major unknown factor in implementing AES into high-stakes writing assessment. Julie Cheville writes about specific concerns in how AES influences writing pedagogy, and the dangers that come with allowing private industry to create a space within curriculum. She admonishes AES for reducing what students believe is good writing into something that is formulaic, stating:

> When the "reader" is a machine calibrated not to meaning but to static compositional features, formulaic considerations subordinate meaning, rendering it subservient to structure. Whether the technology replaces the teacher or supplements instruction, its effect is the same. Formula filters meaning. (Cheville 50)

This is crux of the main disagreement between those who support AES and those who do not-- the formulaic nature of writing. This issue also illustrates a main difference between those who teach writing and those who simply look to assess it. As an English or

composition teacher knows, many times it is when students experiment in writing that their true skills are developed and showcased. This type of experimentation would be marked low by an AES system as it challenges the algorithm, or formula, on which the AES system is based.

In discussing the washback effect, Dianne Wall states,

> The possible negative effects included encouraging teachers to watch the examiner's foibles and to note his idiosyncrasies in order to prepare pupils for questions that were likely to appear, limiting the teachers' freedom to teach subjects in their own way, encouraging them to do the work that the pupils should be doing, tempting them to overvalue the type of skills that led to successful examination performance, and convincing them to pay attention to the purely examinable side of their professional work and to neglect the side which would not be tested. (Wall 500)

This negative effect on teachers that Wall discusses in terms of washback directly translates into a danger within AES. In teaching students writing that is meant to be scored by a machine, there is a loss in the conveyance of effective meaning. Julie Cheville suggests that the errors that are privileged in AES systems are arbitrary, and revolve around what machines can do with language as opposed to what people can do with language. If teachers are teaching writing to these tests, then only certain types of errors, such as spelling or grammar over style, in writing will be privileged at the classroom level, in an effort to maintain stasis with AES.

But there are larger implications to this for students who are learning writing for life, not just for these assessments. Cheville states, "The lack of agreement on what

constitutes an error, as well as conflicting notions of correctness, is a vital dimension of students' language experiences, constituting not a problem but a necessary and important instructional opportunity that automatic scoring technologies override" (50). Because writing is a social construct, what constitutes a major error will differ from type of one type of writing to another (i.e.; argumentative to explanatory) as well the audience for the piece of writing. Cheville discusses a study of three thousand college essays conducted by Connors and Lunsford, in which the top ten most frequently marked errors by college professors included comma placement, sentence fragments, wrong or missing prepositions, etc. (50). However, in a separate study discussed by Cheville, 101 professionals in management positions defined serious errors in writing as run-on sentences, non-capitalization of proper nouns, and lack of subject-verb agreement, among others (Cheville 50). These two studies illustrate how different major errors in writing are, dependent upon the social construct surrounding the type of writing, in this case academia versus non-academia. AES systems do not consider the differences in audience and purpose in writing.

Michael B. Neal echoes many of Cheville's sentiments in his book *Writing Assessment and the Revolution in Digital Texts and Technologies*, focusing on the negative pedagogical impact of AES. He writes of the loss of teacher-student dialog in the classroom: "Much as mechanized writing assessments steal development opportunities from faculty in large-scale writing assessments at institutions, they also threaten opportunities at the classroom level to use writing assessment as a place for dialogue and teaching, which is fundamental to the way many teachers understand and value the teaching of writing" (69). The un-ebbing influx of technology into writing

assessment seems to be reducing literacy into an isolated set of skills rather than looking at writing as an important, holistic, rhetorical act. This skewed view is made clear to the students who must sit for the high stakes assessments that utilize AES, and therefore are being taught, albeit it unconsciously, that only certain skills of writing are required, and these do not include understanding of context or the social nature of rhetorical interaction. In most cases AES systems seem to privilege academic-based writing skills over those that may relate more to the social construct of writing.

The purpose of assessing writing is to know that the student is able to write well in a range of situations and for a range of purposes, not just well enough to fool a machine into giving the writing a passing score, which is one of the possibly unintended effects of using AES. In his 2013 article, "Large-Scale Assessment, Locally-Developed Measure, and Automated Scoring of Essays: Fishing for Red Herrings?" Robert Condon speaks out about this very same problem and argues:

> In short, the ability of AES to evaluate samples that reveal writing ability—not just fluency, accuracy of text production, and sophistication of vocabulary—is scant, if not non-existent, and for that reason assessments that can be scored via AES are poor predictors of students' success in courses that require them to think, to write with an awareness of purpose and audience, and to control the writing process. (103)

Condon suggests here that students are taught how to write responses for writing assessments in a way that has little to do with actual skill and ability and everything to do with gaining a higher score on the assessment. If this is the case, the validity of the assessment is challenged, as the instrument for the assessment itself is flawed and does

not accurately test what it is supposed to test. Therefore, Condon suggests that it is the assessments themselves that are flawed.

Condon's argument—that it is the type of writing assessment that is problematic and that the fight between AES and human scoring is simply a ruse to keep the attention away from the main issue of broken assessments—is eye opening. Condon concedes many of the points made about both AES and human raters brought up in this paper; he agrees that human raters are more on point when it comes to scoring writing quality, while machines can score the more superficial aspects of texts effectively (Condon 101). However, Condon also points out that:

> Ultimately, the focus of large-scale, nationally normed tests to judge a sample of writing is itself too constraining to be useful within an educational context, where the focus is on improving the competencies of a writer. Thus, the type of test AES can score is in conflict with the needs of a student to learn how to improve as a writer and of a teacher, who needs to know how to facilitate that improvement. (101)

Writing assessments on these tests are essentially ineffective at testing the competencies required of a student who is truly adept at writing, because the timing is too constrained to really showcase true strengths in student writing. Condon posits that the information gleaned from the type of exams that use AES is reduced to a simple number, and that this number, "represents a quantification of a severely reduced set of characteristics of good writing—so severely reduced as almost to be irrelevant" (Condon 104). The writing construct itself becomes undervalued in the face of large-scale writing assessments that strive only for validity and reliability. Adding into this mix the threat to teachers from

high-stakes assessment accountability measures, and there is a perfect storm of issues that threatens to warp writing pedagogy within the classroom.

So where do we go from here? As our current society evolves further into a more technology-based culture, the use of artificial intelligence to do jobs that were once traditionally reserved for humans is also becoming more and more prevalent. As a result, I believe that Automated Essay Scoring is here to stay and will only grow in popularity, regardless of what its critics say. However, we should consider doing what Condon suggests: look into the assessments themselves and debate whether or not they truly and validly assess writing as is required and necessary. We should be questioning why we let such forms of assessment, the kinds that *can* be scored by a machine, continue to be used in high-stakes scenarios and determine the academic fates of students. Further unbiased research into the effects of AES on writing pedagogy should be done to truly assess whether it has a measurable positive or negative impact on student writing before it is used in high-stakes assessments that impact the futures of both students and teachers. Research has been done on the use of AES being used effectively in the secondary literacy classroom as a form of formative assessment. This is where writing pedagogy should be focused, on teaching writing effectively in the classroom instead of a means to an end within high-stakes assessment.

## 5. AES for Formative Assessment

Formative assessment within the classroom is a means by which teachers can gather diagnostic data on the strengths and weaknesses of their students' skills, and tailor

instruction to their students' levels. On the other end of the spectrum is the summative assessment, which is used to determine and judge what students have already learned. Unlike high-stakes assessments, which fall under the label of summative assessments, and which occur one to two times a year, usually mandated on a state-level, formative assessments can be done weekly, or even daily, by the teacher. Formative assessments can take the shape not only of traditional assessments such as tests and quizzes, but also homework, classwork, writing assignments, exit slips, or polls, all of which provide a wide range of data that creates a more rounded and ongoing picture of student progress than what is gained from a few high-stakes assessments. When it comes to writing pedagogy, essay assignments are obviously the most prevalent form of writing assessment, as they provide a concrete sample of student skills. While AES is used in high-stakes assessment because it can provide scores within a short amount of time, at the local, classroom level, essay scores are given by the teacher and these take a long time to come by. This is due to the sheer amount of time and effort that goes into effectively and thoughtfully responding to and assessing student writing. Personally, as a working English teacher, with an average of one hundred and fifty students each year, essays take over a week for be graded and commented on because I take my time looking not only at content and context, but also the grammar and other writing issues that abound in student writing. By the time I give back the essays, many of the students have lost interest in the comments and feedback of their essay, but fixate only on the grade itself.

It is here, in this space at the local level, that I believe Automated Essay Scorers have the ability to be truly beneficial in terms of writing pedagogy, rather than as a form of high-stakes assessment. Automated Essay Scorers can potentially provide

instantaneous feedback to students in terms of grammar, spelling, and punctuation-- all key components to effective writing, but one that is formulaic as these are all issues in writing that follow explicit rules, which can be termed as lower-level concerns in writing. There has been considerable research in utilizing AES in this way, referred to as Automated Writing Evaluation (AWE) or Automated Essay Evaluation (AEE). AWE and AEE differ from AES in that AES is used primarily as a means of summative assessment, usually within high-stakes assessments. AES and AEE refer to the automated systems that provide feedback on student writing and function as means of formative assessment, not simply providing a score as a method of summative assessment. While most beneficial would be the feedback AWE systems can provide in terms of grammar and structure, some systems also provide stylistic feedback which may be useful on a smaller level, at the teacher or instructor's discretion.

Proponents of AWE claim that its usage facilitates more writing practice, and increases students' motivation to write and revise, as students no longer have to wait until teachers get through the mass grading and evaluating-- the feedback is instantaneous. In order to test this, Professors Douglas Grimes and Mark Warschauer, neither of whom is associated with any corporation that creates AWE system, evaluated the effects of the AWE system *My Access* in eight middles schools in Southern California over three years. Between years one and three, the researchers found that the number of essays revised by students rose from 12% to 53%, which backs up what many proponents of AWE suggest-- students are more likely to revise their writing if it is returned in a timely manner (Grimes and Warschauer 15). The nature of the improvements made during revisions is not addressed in the study, which focused on teacher and student attitudes towards

utilizing AWE in the classroom. Grimes and Warschauer conclude that, "mindful use of AWE can help motivate students to write and revise, increase writing practice, and allow teachers to focus on higher- level concerns instead of writing mechanics" (34).

Interestingly enough, Grimes and Warschauer make sure to note that, " However, those benefits require sensible teachers who integrate AWE into a broader writing program emphasizing authentic communication, and who can help students recognize and compensate for the limitations of software that appears more intelligent at first than on deeper inspection" (34). This last point speaks to the case that these automated systems, be they AES or AWE, are limited in scope, and cannot be trusted completely to assess writing as they are, in the end, simply machines. A human element is needed to counterbalance the weaknesses of the machine, and provide the feedback on high-level concerns, such as complexity of ideas and depth of analysis, within writing that go beyond the scope of any machine.

The importance of feedback on writing is a critical aspect of AWE. AES systems used for summative assessment do not provide students with feedback, as they are used to simply grade a student response without the student ever understanding the reason they received the score they did. Research suggests that feedback in educational environments is most effective when " it informs the learner how to do the task better as opposed to providing praise for correct performance or punishment for mistakes" (Kellogg 174). AES systems do exactly this: they provide a high score for writing that follows its algorithmic design for what the system deems to be "good writing," and punishes writing that deviates from the same design, but there is no other specific feedback received by the student other than the score.

AWE systems can fill this need in the classroom by giving feedback in a timely manner to benefit the individual student as he or she works through the writing process. Kellogg, Whiteford and Quinlan looked into the effects of AWE feedback on writing at the college level, specifically studying a freshman composition course that used *Criterion* for essay feedback. *Criterion* is an AWE system created by the ETS that not only provides a holistic score to a piece of writing, but also identifies errors in grammar, mechanics, usage, and provides general comments on style (Kellogg 177). In the study, students in a freshman composition course received no feedback or varying amounts of formative feedback on their first drafts of three practice essays with the level of feedback received labeled as "none, intermittent, or continuous;" which level of feedback received was chosen by the researchers. Students in one section of the course received no feedback throughout all three essays, another section received feedback only on the second essay (intermittent feedback), and in a third section of the course the students received feedback on all three essays (continuous feedback). *Criterion's* feedback focuses on errors in regards to grammar, usage, mechanics, and general stylistic comments; *Criterion* also assigns the essay a holistic score from 1-6 generated by the *e-rater* AES system. Feedback is generated as a report with suggestions for corrections, such as "'You have used passive voice in this sentence. Depending upon what you wish to emphasize in the sentence, you may want to revise it using the active voice'" (Kellogg 179).

The students wrote a first and then a revised final draft of an essay in a writing lab held once per week. After a retention interval of two weeks, a test essay was written without the use of feedback to assess transfer of learning. The results of the study were interesting; in terms of holistic, rubric scores, no reliable gains were seen. However, the

transfer essay written by the students who received continuous feedback "showed that students learned to reduce errors of mechanics, usage, grammar, and style," which means students retained this feedback from the AWE system and utilized it in later writing (Kellogg 173). The authors do note that all the students, regardless of the level of feedback given, were able to revise their essays beneficially due to the nature of writing in drafts. However, in terms of the errors highlighted by *Criterion*, only those students who received continuous feedback were able to truly reduce the number of those specific errors, showing that the feedback was most effective when received every time (Kellogg 187).

The conclusion to this study by Kellogg, Whiteford and Quinlan also notes, similar to the Grimes and Warschauer study, that AWE is at its heart only a machine and that, "Although automated feedback has these advantages, human judgment remains the 'gold standard' for evaluating an essay. Writing is a mode of communication, and ultimately only a human reader can evaluate the relative effectiveness of that communication" (Kellogg 190). Kellogg, Whiteford and Quinlan explain that though ideally each student would receive individually tailored feedback from an instructor on how to write effectively, it is simply not a practical expectation. AWE then provides a practical solution to this issue, but should not be expected to be the solution to summatively assessing writing. Formative feedback should also include human feedback on higher-level concerns in writing. AWE seems to be a helpful system to provide the feedback on lower-level concerns, allowing for teachers or instructors to focus on the feedback needed to address higher-level concerns. It is of note that Quinlan works for the

ETS, which is one entity that supports AES and creates AES systems to use in high-stakes assessment, yet Quinlan does not seem to support AES for high-stakes assessment.

One concern of AES is how opposed students seem to be when it is used to score their writing on high-stakes assessment. Would students have the same adverse reaction to using AWE in the classroom for the purposes of learning and formative assessment? The student response to AWE systems used to better student writing is showcased in a 2013 study by Rod Roscoe and Danielle McNamara, who studied the effects of an intelligent tutoring system (ITS) called The Writing Pal (W-Pal) at the high school level. W-Pal offers "*writing strategy instruction* along with game-based practice, essay writing practice, and formative feedback to high school students," which differs slightly from AWE that only provides feedback (1010). W-Pal attempted to guide students in terms of both lower-level concerns such as grammar and usage, but also encouraged students to consider higher-level concerns such as organization and style (Roscoe 1014). Because W-Pal is still experimental, this article focuses not on the instructional efficacy of the product, but rather on how students in two high schools interacted with the system, and the student perception of the design of the system. The conclusions drawn by the study showed that students were open to the idea of ITS, and indeed viewed it as a valuable addition to the classroom. Along with several presentational issues in the software, one of the main concerns was in the feedback that the students received on their writing through the ITS. Students wanted more specific and individualized feedback from the ITS, which the authors determined could be made better by tweaking the algorithms in the system. So it seems that students, while horrified by AES used in high-stakes assessments, are open to AWE systems providing them with knowledge and feedback on writing in a

classroom setting. Indeed, the students in the study wanted more specific feedback from the AWE system to help guide their writing.

It does seem that AWE has a lot of the answers to the issues that make AES problematic- it provides beneficial feedback that can be employed by the students as they work through the writing process, and students seem to be more open to utilizing it in the classroom. However, when discussing W-Pal as a machine, Roscoe and McNamara also argue that, "a certain level of permanent ambiguity may have to be embraced, and the focus must be on guiding students toward progress and independence, rather than delivering, correcting, or testing a well-defined body of knowledge" (Roscoe 1023). This is the same issue that both the Grimes and Warschauer, and Kellogg, Whiteford and Quinlan touch upon in their studies-- AWE systems, like AES systems, are mechanized systems that can only go so far in their efforts to critique and provide meaningful feedback to student writing. The missing ingredient is the human aspect, the teacher. Roscoe and McNamara note that, "The fundamental purpose of AWE systems is the facilitation of writing assessment rather than teaching students about writing principles, goals, and strategies. Without such instruction, students may not be prepared to utilize the detailed writing feedback these tools offer" (1012). Without the teacher guiding instruction, the feedback generated by the AWE system loses relevance and meaning to the student. An instructor has the power the take the AWE feedback in a piece of student writing and use it as formative assessment, thereby creating instruction tailored to the critiques put forth by the AWE to support student writing. Therefore, the ideal situation would showcase the strengths of the AWE system in terms of lower-level feedback, while simultaneously incorporating teacher/instructor feedback on higher-level concerns.

Assessing the student knowledge gained by both the AWE system and through teacher feedback may be gauged through the usage of portfolios, specifically ePortfolios.

Portfolios are widely seen as a best practice in the field of writing pedagogy and assessment. Michael Neal defines portfolios as multiple student created texts, formal to informal, which showcase the writing process from drafting to revision, which incorporates student reflection on their writing process (81). The Common Core Anchor Standards for Writing include two standards that align with the usage of portfolios for writing. Writing Anchor Standard Five states that students need to "Develop and strengthen writing as needed by planning, revising, editing, rewriting, or trying a new approach," while Writing Anchor Standard Ten states that students should, "Write routinely over extended time frames (time for research, reflection, and revision) and shorter time frames (a single sitting or a day or two) for a range of tasks, purposes, and audiences" (http://www.corestandards.org/ELA-Literacy/CCRA/W/). Neal believes that portfolios have the ability to "demonstrate more about students' knowledge and skill than a single piece of writing or a final product could communicate" (81). The Common Core Standards seem to support Neal's argument that portfolios are the most beneficial for truly valid writing pedagogy and assessment. And yet the PARCC assessments, like all high-stakes assessments, ask students to create a piece of writing in a timed frame on a subject they have little true knowledge of prior to the exam, and use the scores to determine student skill level.

Portfolios, however, allow students to have ample time to write and go through the entire process of writing, as the deadlines are up to the individual instructor or teacher. Grimes and Warschauer, in a 2008 study on using AWE in the classroom, note

that, "Currently, it [AWE] seems most helpful for motivating and assisting young writers to reduce the number of mechanical errors, thereby freeing the teacher to focus on content and style" (Warschauer 34). AWE used in conjunction with portfolios would be beneficial in helping provide teachers or instructors with the time to focus on feedback pertaining to higher-level concerns, as it could take care of providing students with the feedback on lower-level concerns.

The necessity for a human to act as a secondary reader of student writing is well documented in much research associated with AWE. As illustrated, most researchers seems to believe that AWE is can only provide so much useable feedback when it comes to student writing, and that the teacher/instructor component is most essential if AWE is truly supposed to function as a means of formative assessment for student writing. Research also shows that students are more open to taking feedback on their writing for lower-level concerns from AWE system rather than AES systems. Therefore, the ideal method for utilizing AWE as formative assessment seems to be for using AWE in electronic portfolios (ePortfolios). Les Perelman notes that for a construct as multi-faceted as writing that "Portfolio evaluations clearly offer the most promising platform for assessing this complex construct" ("Construct Validity" 129). One of the Common Core Anchor Standards for Writing explicitly states, "Use technology, including the Internet, to produce and publish writing and to interact and collaborate with others" (http://www.corestandards.org/ELA-Literacy/CCRA/W/). ePortfolios provide a means by which students can practice writing in a digital context, a skill that the Common Core Standards deem necessary to success in the 21st Century. Utilizing AWE in conjunction with ePortfolios may provide a much-needed crossover of the strengths of AWE and

ePortfolios with the added benefit of teacher feedback and guidance in order to truly showcase student growth in writing.

In a joint venture between the faculty of New Jersey Institute of Technology's (NJIT) first-year writing program (including Norbert Elliot) and the ETS (that included researchers Paul Deane and Chaitanya Ramineni) a study was done on the implications of using ETS's *Criterion* AWE system throughout the ePortfolios required of freshmen students at NJIT. By utilizing AWE in ePortfolios, the distrust of technology as an assessor became a non-issue, as the writing itself was being done digitally. As noted by the authors of the study, "students compose in an interactive medium in which an AES system such as Criterion becomes part of a fluid environment where a machine score is viewed as an invitation to revise instead of a judgment to be suffered" (Klobucar et al 106). The end results of the ePortfolios were still read and graded by two instructors with rubrics; *Criterion* here was only used as formative assessment on the part of the student, who had the power to make revisions as suggested by the AWE system. The end results of the study are mixed; because this study was undertaken at the college level, students simply stopped using *Criterion* throughout the semester because they could not be compelled to use it by the instructors. The authors note that this study may be too specific to yield significant results, and therefore their results may be too limited to be useful.

What the authors do bring to light are the methods for utilizing AWE systems in a meaningful way for writing pedagogy. As mentioned earlier, the world is becoming a more digitalized place; allowing students to practice writing digitally can only be beneficial to their growth in writing for the real world. The collaboration of ePortfolios and AWE in this study opens up the arena of writing pedagogy research into creating, "an

environment designed to encourage student writing, with automated feedback driven by an analysis of student responses, such features may have additional value as cues for feedback that is fully integrated with the writing process" (Klobucar et al 114). AWE and AEE systems like *Criterion* provide new possibilities for writing pedagogy and assessment, because as the world changes, so must writing pedagogy to fit the needs of this "brave new world." However, just as other researchers have noted, this study also posits that,

> The roles that writing assessment systems play depend on how they are integrated into the practices of teachers and students. If automated scoring is informed by enlightened classroom practice—and if automated features are integrated into effective practice in a thoughtful way—we will obtain new, digital forms of writing in which automated analysis encourages the instructional values favored by the writing community. (Klobucar et al 115)

The researchers here, including those who work for ETS, do insist that in order for AWE systems to be truly beneficial to students, a teacher or instructor must be part of the writing pedagogy process. Effective writing pedagogy cannot hinge on technology alone.

While certainly helpful, AWE systems, like their AES counterparts, are by no means infallible, and are susceptible to "gaming." Tim McGee, a writing program administrator, became intrigued by Pearson's AWE system, *Intelligent Essay Assessor* (IEA), which promised to measure factual knowledge unlike most AES systems that are used strictly for scoring. McGee signed up for a trial version of the software, and ran a host of experiments in gaming. He chose a sample essay provided by the IEA on the circulatory system, which was deemed a 4 on a scale of 1-5. McGee then rewrote the

essay into IEA in reverse, in which the first sentence became the last sentence and so on. As McGee notes, "The meaning of the individual sentences is unchanged, but the assembled whole has suffered a substantial reduction in both cohesion and coherence, not to mention factual accuracy" (87). The IEA gave the reversed essay a score of 4-- the score it received when it was written in order as a sample essay. McGee outlines other attempts at gaming in his article. McGee entered an essay about the Great Depression, and received a high score of 5. He then reversed the facts in the essay; for example, reversing the dates of the stock market crash and the end of the New Deal. The machine scored this new essay with a 5, not changing its original high score. (McGee 80-89)

What is clear after reading McGee's article is that the *Intelligent Essay Assessor* could not read for meaning as a human could. Despite what marketers and advertisements suggest, machines are unable to read for meaning, nor can they provide accurate human responses, in terms of high-level concerns, to human writing. This is why all of the research in this section makes note of the need for a human eye to double check the AWE system's feedback. However, this fact does not mean that AWE systems are abominable and therefore should not be used in writing pedagogy. As showcased above, AWE systems can provide supplementary support to a writing teacher or instructor in terms of the parts of writing that *are* formulaic, much like Microsoft Word's Spelling and Grammar Check does to a lesser extent already. What these systems cannot do is take the place of the teacher or instructor in terms of providing guidance on higher-level concerns in writing, and the research seems to agree with this point. Carl Whithaus states, "When software works well for a particular task, writing researchers should build pedagogies that incorporate these features. When the use of software produces decontextualized,

invalid writing assessments, writing researchers need to point out the faults of these systems" (176). Therefore, we as teachers of writing in the K-12 context should not simply dismiss all automated writing evaluation or automated essay evaluators, but should utilize them for their strengths, mainly as another type of formative assessment in our classrooms.

## 6. Conclusion

As I write this paper in December 2014, more and more states are pulling out of Race to the Top and repealing the use of the Common Core State Standards (Turner). The reasons for this are multifaceted, and the fight has involved stakeholders across education and politics at all levels. However, just because the Common Core (and therefore PARCC and Smarter Balanced) is losing some steam as it struggles to change the US education system, it does not mean that AES will share the same losses. It is hard to imagine a future in writing assessment without AES. The technological gains that have been, and continue to be, made are just too great to be ignored, not to mention the monetary gains AES brings as well. Through PARCC and Smarter Balanced, AES has found a foothold in the K-12 education market, and will most likely not disappear anytime soon.

The most vital issue, that which will determine the direction of the future of writing assessment, is how to define the writing construct. For those who are pro-AES, writing is easily defined, broken down into key components that can be judged as "good" or "bad," which can then in turn be coded into a machine that will also be able to assess

"good" or "bad" writing. Those who are critical of AES believe that writing can rarely be broken into arbitrary categories like "good" or "bad;" that the nature of writing, from a rhetorical perspective, is determined by a whole host of factors-- audience, topic, style, purpose-- the list goes on. This view generally comes from those who have not had the opportunity to think deeply about writing and its complexity. Purposeful and successful writing, at its core, is that which showcases meaningful communication and connection from human to human, in a complex and socially consequential way. A machine simply cannot duplicate that; there is no algorithm that can be written that can judge that the way a human can, regardless of the leaps in technology that have been made. Asking students to write to a machine that will judge and score their writing, in a high-stakes assessment, negates the true purpose of writing and trivializes its importance.

However, there are some surface elements of writing, such as grammar and punctuation, which can indeed be categorized and checked. While these factors are on the lower end of the scale as to what makes a piece of writing truly successful, they are an important factor of writing done well. It is with these elements of writing that AES should be utilized, not in summative, high-stakes assessment, but as a tool for formative assessment in the guise of AWE. Xioming Xi, a researcher for the ETS, acknowledges the shortcomings that AES systems have had thus far in his 2014 article, but claims that strides are being made presently as the pool of researchers who are working on AES systems have been broadened to include both computational and applied linguists. In the end however, the author admits:

> However, it is also important to realize that the accuracy of feedback given by computers, although acceptable in low-stakes practice environments with

> instructor support, leaves considerable room for improvement to emulate the judgment of trained linguists. (298)

Xi, notably writing as an employee for the ETS, echoes my own sentiments here that AES is still too underdeveloped to be used in the high-stakes testing realm, but it does have enough utility to be employed in lower-stakes, formative assessment in a classroom setting where a teacher is present. The teacher, instructor, or professor represents that key component missing from AES that renders them inadequate-- the human element that is necessary for true writing assessment.

My contribution is that using AES in high-stakes assessments should be banished altogether, due to issues on multiple levels, from current politics in education to weaknesses in the AES systems that can be manipulated. However, instead of focusing on the negative aspects of AES as a whole, research should be done into using the strengths of AES in the writing classroom through its incarnation as AWE. AWE systems can provide formative feedback and support on the individual student's writing on low-level issues such as grammar, punctuation, spelling and usage. This can leave the teacher with enough time to effectively provide student feedback and guidance on higher-level issues in regards to issues such as style, tone, purpose, audience, etc. By utilizing the strengths of both the AWE system and the teacher's knowledge in a digital context, a true assessment of student writing can be rendered through the ePortfolio system, thereby negating the major issues of using AES systems in high stakes assessments.

Works Cited

Bejar, Isaac I., Michael Flor, Yoko Futagi, and Chaitanya Ramineni. "On the Vulnerability of Automated Scoring to Construct-irrelevant Response Strategies (CIRS): An Illustration." *Assessing Writing* 22 (2014): 48-59. *JSTOR*. Web. 10 Dec. 2014.

Cheville, Julie. "Automated Scoring Technologies and the Rising Influence of Error." *The English Journal* 93.4 (2004): 47-52. Online.

Clarke, Marguerite M., George F. Madaus, Catherine L. Horn, and Miguel A. Ramos. "Retrospective on Educational Testing and Assessment in the 20th Century." *Journal of Curriculum Studies* 32.2 (2000): 159-81. *JSTOR*. Web. 10 Dec. 2014.

Condon, William. "Large-Scale Assessment, Locally-Developed Measures, and Automated Scoring of Essays: Fishing for Red Herrings?" *Assessing Writing* 18 (2013): 100-108. *ERIC*. Web. 10 Nov 2013.

Darling-Hammond, Linda, and Frank Adamson. *Beyond the Bubble Test: How Performance Assessments Support 21st Century Learning*. San Francisco: Jossey-Bass, 2014. Kindle file.

Deane, Paul. "On the Relation Between Automated Essay Scoring and Modern Views of the Writing Construct." *Assessing Writing* 18 (2013): 7-24. *ERIC*. Web. 10 May 2014.

Deane, Paul, Frank Williams, Vincent Weng, and Catherine S. Trapani. "Automated Essay Scoring in Innovative Assessments of Writing from Sources." *Journal of Writing Assessment* (2013): 1-17. *JSTOR*. Web. 10 Dec. 2014.

"English Language Arts Standards » Anchor Standards » College and Career Readiness Anchor Standards for Writing." *Common Core State Standards Initiative*. Web. 1 Jan. 2015. <http://www.corestandards.org/ELA-Literacy/CCRA/W/>.

Grimes, Douglas, and Mark Warschauer. "Utility in a Fallible Tool: A Multi-Site Case Study of Automated Writing Evaluation." *The Journal of Technology, Learning, and Assessment* 8.6 (2010): 1-44. Web. 1 Jan. 2014.

Herrington, Anne, and Charles Moran. "Writing to a Machine Is Not Writing At All." *Writing Assessment in the 21st Century: Essays in Honor of Edward M. White*. Eds. Norbert Elliot and Les Perelman. New York: Hampton Press, 2012. 219-232. Print.

"Human Readers." *Human Readers*. 1 Jan. 2013. Web. 6 Jan. 2015. <http://humanreaders.org/petition/index.php>.

Kellogg, Ronald T., Alison P. Whiteford, and Thomas Quinlan. "Does Automated Feedback Help Students Learn to Write?" *Journal of Educational Computing Research* 42.2 (2010): 173-96. *JSTOR*. Web. 6 Jan. 2015.

Klobucar, Andrew, Paul Deane, Norbert Elliot, Chaitanya Ramineni, Perry Deess, and Alex Rudniy. "Automated Essay Scoring and the Search for Valid Writing Assessment *International Advances in Writing Research: Cultures, Places, Measures*. Fort Collins, Colorado: The WAC Clearinghouse and Parlor Press, 2012. 103-117. Print.

McGee, Tim. "Taking a Spin on the Intelligent Essay Assessor." *Machine Scoring of Student Essays: Truth and Consequences*. Logan, UT: Utah State UP, 2006. 79-92. Print.

McNeil, Michele, and Catherine Gewertz. "States Seek Flexibility On Testing." *Education Week* 32.35 (2013): 1. *ERIC*. Web. 13 Sept. 2014.

"NCTE Position Statement on Machine Scoring." *NCTE Comprehensive News*. NCTE Executive Committee, 1 Apr. 2013. Web. 6 Jan. 2015. <http://www.ncte.org/positions/statements/machine_scoring>.

Neal, Michael B. *Writing Assessment and the Revolution of Digital Texts and Technologies*. New York: Teachers College Press, 2011. Print.

Perelman, Les C. "Construct Validity, Length, Score, and Time in Holistically Graded Writing Assessments: The Case against Automated Essay Scoring (AES)." *International Advances in Writing Research: Cultures, Places, Measures*. Fort Collins, Colorado: The WAC Clearinghouse and Parlor Press, 2012. 121-132. Print.

Perelman, Les C. "Critique of Mark D. Shermis & Ben Hammer, "Contrasting State-of-the-Art Automated Scoring of Essays: Analysis". "*Journal of Writing Assessment* 6.1 (2013): 1-11. *The Journal of Writing Assessment*. Web. 1 June 2014. <http://journalofwritingassessment.org/article.php?article=69>.

Perelman, Les C. "When "The State of the Art" is Counting Words." *Assessing Writing*: 104-111. *Science Direct.* Web. 1 July 2014.

Powers, Donald E., Jill C. Brustein, Martin Chodorow, Mary E. Fowles, and Karen Kukich. "Stumping E-rater: Challenging the Validity of Automated Essay Scoring." *Computers in Human Behavior* 18 (2002): 103-134. *ERIC*. Web. 10 Nov 2013.

Ramineni, Chaitanya, and David M. Williamson. "Automated Essay Scoring: Psychometric Guidelines and Practices." *Assessing Writing* 18 (2013): 25-39. *ERIC*. Web. 10 May 2014.

Roscoe, Rod D., and Danielle S. McNamara. "Writing Pal: Feasibility of an Intelligent Writing Strategy Tutor in the High School Classroom." *Journal of Educational Psychology* 105.4 (2013): 1010-025. Print.

Rothermel, Beth Ann. "Automated Writing Instruction: Computer-Assisted or Computer-Driven Pedagogies?" *Machine Scoring of Student Essays: Truth and Consequences*. Logan, UT: Utah State UP, 2006. 199-210. Print.

Shermis, Mark D. "State-of-the-Art Automated Essay Scoring: Competition, Results, and Future Directions from a United States Demonstration." *Assessing Writing* 20 (2014): 53-76. *JSTOR*. Web. 10 Dec. 2014.

Turner, Cory. "Common Core Repeal, The Day After." *NPR*. NPR, 30 Dec. 2014. Web. 6 Jan. 2015.

Wall, Dianne. "The Impact of High-stakes Testing on Teaching and Learning: Can This Be Predicted or Controlled?" *System* 28 (2000): 499-509. *JSTOR*. Web. 10 Dec. 2014.

Wang, Jinhao, and Michelle Stallone Brown. "Automated Essay Scoring Versus Human Scoring: A Comparative Study." *Journal of Technology, Learning, and Assessment* 6.2 (2007): 4-29. *ERIC*. Web. 10 Nov 2013.

Warschauer, Mark, and Douglas Grimes. "Automated Writing Assessment in the Classroom." *Pedagogies: An International Journal* (2008): 22-36. Print.

Whithaus, Carl. "ALWAYS ALREADY: Automated Essay Scoring and Grammar-

Checkers in College Writing Courses." National Writing Project, 1 Jan. 2011. Web. 1 July 2014. <http://digitalis.nwp.org/resource/2090>.

Williamson, David M., Xiaoming Xi, and F. Jay Breyer. "A Framework for Evaluation and Use of Automated Scoring." *Educational Measurement: Issues and Practice* 31.1 (2012): 2-13. *ERIC*. Web. 10 May 2014.

Winerip, Michael. "Facing a Robo-Grader? Just Keep Obfuscating Mellifluously." *The New York Times.* 22 Apr. 2012, sec. Education.Web.

Xi, Xiaoming. "Automated Scoring and Feedback Systems: Where Are We and Where Are We Heading?" *Language Testing* 27: 291-300. Web. 1 July 2014.

Zhang, Mo. "Contrasted Automated and Human Scoring." *ETS: R&D Connections* 21. (2013): 1-11. Web. 10 May 2013.