Summer 7-2013

Ethical Implementation of an Automated Essay Scoring (AES) System: A Case Study of Student and Instructor Use, Satisfaction, and Perceptions of AES in a Business Law Course

John K. Lewis
Salve Regina University, lewisj@salve.edu

Ethical Implementation of an Automated Essay Scoring (AES) System: A Case Study of Student and Instructor Use, Satisfaction, and Perceptions of AES in a Business Law Course.

J. K. Lewis

Salve Regina University
Phone: 401 835-1921
jklnewport@yahoo.com

Abstract

A pilot study of a vendor provided automated grading system was conducted in a Business Law class of 27 students. Students answered a business law fact pattern question which was reviewed and graded by the textbook vendor utilizing artificial intelligence software. Students were surveyed on their use, satisfaction, perceptions and technical issues utilizing the Write Experience automated essay scoring (AES) software. The instructor also chronicles the adoption, set up and use of an AES. Also detailed are the advantages and disadvantages of utilizing such software in an undergraduate course environment where some students may not be technologically adept or may lack motivation to experiment with a new testing procedure.

Automated grading of student assignments is part of the next wave of textbook enhancements that vendors will be providing to instructors in the near future. Several vendors are conducting beta testing with instructors in hope of offering automated grading as part of their textbook and course support. Of course, such services will be an additional cost to the text. Exactly what will be charged for such services remains to be seen.

The vast majority of previous research in the area of AES has been limited to its use in grading assignments in the STEM fields. Computer science instructors have been experimenting with self-created automated grading software for several decades. Recently automated grading software has been implemented by vendors to score essay questions in online tests such as the Graduate Management Admission Test (GMAT).

The experience of one business law class in using automated grading software as part of such a beta test can give valuable insight into the needs and expectations of the three stakeholders in this situation – instructor, student and vendor. Implementation of an AES raises numerous issues. Should an instructor be willing to relinquish control of the grading process to an outside entity? Are the student needs for feedback and grading fairness being met? Is the technology advanced enough to replace the human element always present in grading assignments? What will be the economic impact on students if such software is adopted?

The implications for the various stakeholders are discussed and addressed. The author comes to conclusions about the pedagogical usefulness of AES systems and offers suggestions for best practices to be employed by instructors interested in implementing such software in their courses.

Keywords

Automated Essay Scoring, AES, summative assessment, formative assessment, essay questions, essay grading, computer ethics, Write Experience

INTRODUCTION

An Automated Essay Scoring (AES) system is defined as computer technology capable of evaluating and scoring written prose (Dikli, 2006). The father of AES systems is Ellis Page. Page created the first AES, Project Essay Grade (PEG) back in the 1960's. He continued his research with PEG and other systems into the new millennium until his death in 2005. Early incarnations of PEG required essays be entered into a mainframe computer on punch cards (Page, 1994). As Page continued to improve PEG other AES systems were created to compete with it. Additionally, many computer science instructors created homegrown AES systems to automate grading of programming assignments. Many commercial tests now use AES systems in tandem with a human grader to score essays including the Graduate Management Admissions Test (GMAT), the Test of English as a Foreign Language (TOEFL) and the Graduate Record Examination (GRE) (Ade-Ibijola, Wakama & Amadi, 2012). Many universities also use AES to score admission essays.

Most AES systems work in a similar manner and have similar requirements. All AES systems need to be trained through the creation of a knowledge base. This is usually done by entering human scored essays into the database. Some systems use texts to create their knowledge base. Most systems grade essays based on style and content. The vast majority use Natural Language Processing (NLP) tools to score submissions.

Assessment

Although many instructors enjoy teaching many do not enjoy the effort required in grading student work. Assessment of student learning is an integral part of teaching and also the most time consuming. As Dreher, Reiners and Dreher (2011) state, "assessment guides the teaching and learning process by providing reciprocal feedback to both educators and students so that they may improve in their respective tasks" (p. 162). Assessment can serve two purposes either formative or summative. Summative assessment measures a students learning up to that point in time in a course. Multiple choice tests are often utilized for summative assessment because they can measure a student's knowledge of facts and the course content. Formative assessment is used diagnostically to both assist the student and the teacher. Formative assessment provides feedback to the student on their progress and helps the teacher to refine teaching and learning methods to maximize student progress. There are various methods of formative assessment but essay writing is one of the most common.

Summative assessment

Economic considerations present in education today often dictate large class sizes. Time and effort limitations often necessitate the use of multiple choice exams by instructors. Students experience a great deal of summative assessment but less formative assessment. Blayney and Freeman (2008) point out that multiple choice questions are more efficient, especially with vendor provided pre-existing questions, but they "do not test higher order application or provide extensive feedback that students can use to identify their own misunderstandings" (p. 156). Technology has often been used to assist instructors with grading multiple choice exams, for example the ubiquitous Scantron, which is still in use at many universities. "Such automated assessment can provide a quick, reliable, cost-effective means of assessing large numbers of students" but does not provide formative assessment (Dreher, Reiners & Dreher, 2011, p. 165). Multiple choice testing while useful for summative assessment mostly addresses surface learning and cannot adequately assess application of knowledge to real life situations. Other limitations of the multiple choice format include lower reliability due to student guessing, lower validity due to inadvertent hints provided by response format, and lower validity due to inability to measure complex constructs (Wang, Chang & Li, 2008). Blayney and Freeman (2008) found that students retained less when multiple choice questions were graded by a scanner and returned at the next class than when tested on an answer until correct basis.

Formative assessment

The formative assessment provided by open-ended essay questions is the most productive method of assessing student learning. Formative assessment can collect "detailed information about students' learning status for planning instructional feedback" as well as knowledge and application of concepts (Wang, Chang & Li, 2008, p. 1463). In particular, essay questions are considered by many educators to be the most useful tool for assessing learning outcomes. Open-ended questions require the ability to recall, organize and integrate ideas and the ability to express

oneself in writing (Ade-Ibijola, Wakama & Amadi, 2012). Discussion questions "often display wider aspects of students' individuality, personal perspective, and creativity" (Carpenter, Delugach, Etzkorn, Farrington, Fortune, Utley & Virani, 2007, p. 227).

Essay and discussion questions can also be used to improve students' abilities to solve real world business problems. In today's globalized business environment students must be creative thinkers, problem solvers, planners, decision-makers and able to participate in team activities. Such skills can only be improved through application of the higher levels of Bloom's taxonomy. Bloom's taxonomy has six levels – knowledge, understanding, application, analysis, synthesis and evaluation (Bloom, 1956). Multiple choice questions can assess knowledge and understanding, but the higher levels of application, analysis, synthesis and evaluation require essay and discussion questions. These higher level functions require transfer of theory to practical situations (application), identification of relevant components and logic in the learning material (analysis), combining information to produce new products (synthesis) and making decisions that create an impact on a given application (evaluation) (Dreher, Reiners & Dreher, 2011). Student success is increased when they are "given challenging, real-world practice assignments with rapid, meaningful feedback" (Matthews, Janicki & Patterson, 2012, p. 73).

Disadvantages of Essay Questions

While essay questions are advantageous to student learning and assessment there are obvious disadvantages for the instructor. Grading of essay and discussion questions is both laborious and time consuming even with the help of teaching assistants. "When there are many students and time is short, feedback detail is reduced, assessment quality may be compromised, and in extreme cases a 'tick and flick' approach to grading may seem a tantalizing option" (Dreher, 2006, p. 189).

Another issue with essays is grading inconsistency. Grading essays fairly is much more difficult than grading multiple choice questions. There is no right or wrong answer and determining a grade is often a subjective decision by the grader. There is nothing more aggravating to students than grading inconsistency. Cheang, Kurnia, Lim and Oon (2003) found that grades may vary according to the instructor's mood, alertness and other factors. Evaluating answers in random order can lead to assigning different grades to answers of similar quality. According to Escudeiro, Escudeiro and Cruz (2011), "evaluating an answer of average quality after evaluating a set of answers of much lower quality may lead the evaluator to inflate the grade assigned to the average answer" (p. 15). The opposite effect may lead to a lower grade than warranted for the average answer. Use of a rubric may address this inconsistency. Rubrics lead to more consistency when assessing students' work but add to the laboriousness of the grading process (Carpenter, Delugach, Etzkorn, Farrington, Fortune, Utley & Virani, 2007). As Byrne, Tang, Tranduc and Tang (2010) ask the question arises, "how can a human grader score essays adequately when the number of essays is large and the time to evaluate them short" (p. 1). Increasingly the answer is to turn to automated essay scoring systems.

Advantages of Automated Essay Scoring (AES) Systems

One of the biggest advantages of an AES is that a student receives quick feedback. An AES provides feedback to a student almost instantaneously with submission in most cases. Especially in large classes the speed of feedback is far superior to that of human graders. This is one reason why so many computer science instructors have created homegrown AES systems to grade introductory computer programming courses. Grading of programming is tedious and can be handled much more quickly and efficiently by an AES. Quick feedback is vital and often listed as a best practice for teaching and often recognized as being essential in student motivation to learn (Barker, 2011). In distance education where interaction with an instructor may be minimal or sporadic the use of an AES can not only score essays but tutor the student as well (Duwairi, 2006).

Besides speed of feedback an AES system can also provide consistency in grading, cost-savings for the educational institution, time-savings for the instructor, and reduced error in scoring. Time and date of submission is recorded automatically, and since computers can only be objective no personal bias in grading is possible (Olivier, Herson & Sosabowski, 2001). Another AES benefit is its ability to act as a plagiarism detector. In one study an AES discovered an extremely high rate of plagiarism finding that 98 out of 712 assignments were copied from another student's work (Cheang, Kurnia, Lim & Oon, 2003).

We have all encountered students who become discouraged or belligerent when subjected to criticism even when it is constructive. Students may be more open to such criticism if it is delivered impersonally through an AES. An AES also encourages students to revise their work before submitting it for final grading. In one study students who used an AES wrote three times as many words on an essay as students who did not use an AES (Giles, 2011). Finally, many Net Generation students enjoy the gamification aspect of utilizing an AES. Students treat the AES system as a video game "in which doing well involves redrafting work to get a higher score" (Giles, 2011, p. 22).

Disadvantages of Automated Essay Scoring (AES) Systems

While there are many advantages to employing an AES there can be some challenging problems involved in their use. Some students do not like the idea of a computer scoring their work. Lai (2010) found that students preferred their writing be evaluated by a peer rather than by an AES. Students often worry that an AES cannot understand novel ideas and concepts, or properly grade answers that were not part of its training (Landauer, Laham & Foltz, 2000). Barker (2011) found that as AES feedback increased so did the amount of challenges by students about their grade. Another issue, especially with homegrown AES systems is limited feedback. For instance, when used for programming grading the student may only be told that their code was incorrect. Blayney and Freeman (2008) used a homegrown AES to score basic accounting problems with mixed results. Some students were able "to self-diagnose specific mistakes, but many "novice learners were overwhelmed by having the onus to identify their current error and the improvement advice" offered by the AES (Blayney & Freeman, 2008, p. 165).

Early AES systems such as PEG were susceptible to cheating as writing long essays tended to result in a higher score. Critics often claim that AES systems cannot evaluate the subtleties of good writing. A computer program was incapable of measuring the most important qualities of fine writing such as content, organization and style (Hearst, 2000). In one study, instructors abandoned use of an AES system for this very reason. When 10 of 33 students requested a human re-grading the result was an average increase of 24% in their scores. The instructor found that "machine readers could not detect subtleties of writing such as irony, metaphor, puns, connotation and other rhetorical devices" (Byrne, Tang, Tranduc & Tang, 2010, p. 35). Even AES developers admit that their systems are incapable of scoring English composition assignments or creative writing; the systems work best when scoring factual knowledge (McCollum, 1998).

There can also be technology issues with use of an AES. Because the student is submitting their essay online the speed and reliability of the Internet connection is an important consideration. There is also the very real problem of computer anxiety. Lai (2010) found that over half of the students expressed confusion with the complex AES interface and that the "slow speed of the Internet connection frequently made them impatient and uncomfortable (p. 443). While most Net Generation students are familiar with and enjoy using computers there is still a substantial minority who fear and loathe technology. Some students find the frequent computer interaction required by an AES to be dehumanizing (Lai, 2010). Lastly, most AES systems require a long period of training before they can be employed. "In order to train a machine learning system, a corpus of holistically scored essays needs to be collected, so that it can be used as training data for the system" (Santos, Verspoor & Nerbonne, 2012, p. 5).

Ethical Considerations in Using an AES

Surprisingly very little has been discussed about the ethical considerations of replacing a human grader with a computer system. One obvious ethical advantage is that the AES system is completely unbiased. No human grader can be completely objective even if the author of the essay is unknown. Certain writing styles and choices of topic or language can affect a human grader if only on a subconscious level. For a professor who interacts with students on a regular basis the possibility of bias entering into the grading process is a very real possibility. Favored students are more likely to be graded leniently while out-of-favor students may be held to a stricter standard. A computer is not affected by such considerations. The essay is graded based on content and style in accordance with the training materials that have been inputted.

However, use of a computer for grading may be dehumanizing. This is especially true if an AES is the only grader or in an online course where the student has little interaction with the professor. Also use of an AES may be unfair to those students who are technology challenged or who suffer from computer anxiety. Use of an AES could be

unfair on an economic basis as well. Those who are less affluent may not have Internet access or a computer in their home. They may have to take the test in a less than optimal environment such as a noisy public library. Use of AES systems in a K through 12 environment for standardized testing purposes could also favor more affluent school systems over those with less resources.

AES systems may still be susceptible to cheating. Williamson, Xi and Breyer (2012) point out that "the potential for vulnerability in scoring unusual or bad-faith responses inappropriately" is still a possibility (p. 3). Recently Les Perelman, a professor of Comparative Media Studies at the Massachusetts Institute of Technology, proved that AES systems can be tricked by savvy test takers. AES systems can be "dazzled by big but meaningless words" and if the correct algorithms are used "a bad essay full of botched facts can still get a good score" (Berrett, 2013, p. 4). Perelman also recently attacked the validity of studies that claimed AES systems scored similarly to human graders in head-to-head tests. "The standard method for comparing the reliability of machine scores to human scores is to compare the reliability of the machine scores to each of two human scores and then, compare those scores to reliability of the human scorers to each other, in those studies, as in many others, humans clearly outperformed machines" (Perelman, 2013, p. 8).

Even the most ardent advocates of AES systems admit that they are limited to identifying words and phrases that are characteristic of a strong answer, no AES can understand context or the deeper meaning of language typical of quality writing. Use of an AES may also send the wrong message to students. "It tells students that writing is so unimportant that we're not willing to take the time to read it" (Berrett, 2013, p. 2).

The National Council of Teachers of English (NCTE) recently released a position paper denouncing the use of AES systems in high stakes testing. The organization is particularly concerned with plans by many states to use AES systems to grade tests used to measure Common Core State Standards (CCSS). The NCTE believes "computers are unable to recognize or judge those elements that we most associate with good writing, and that computers use different, cruder methods than human readers to judge students' writing" (NCTE Position Statement on Machine Scoring, 2013, p. 1). In addition, the NCTE believes that "computer scoring favors the most objective, surface features of writing, and removes the purpose of written communication – to create human interactions through a complex, socially consequential system of meaning making" (NCTE Position Statement on Machine Scoring, 2013, p. 2). The NCTE also believes that AES are unfair to non-native speakers of English.

AES Evaluation Criteria and Performance Measurement

According to Kaplan, Wolff, Burstein, Li, Rock and Kaplan (1998) the four most important characteristics used to measure the effectiveness of an AES system are accuracy, defensibility, coachability and cost-effectiveness. An AES must be accurate when compared to a human grader. An assigned grade must be defensible through explanation of grading criteria and comparison to a rubric. Coachability is not a desired characteristic in AES. A coachable AES is one that is "based on simple, surface based methods that ignore content, students could train themselves to circumvent the system and so obtain higher grades than they deserve" (Kakkonen, Myller, Sutinen & Timonon, 2008, p. 277). One of the criticisms of PEG was that it was coachable simply be writing long essays filled with facts. Cost-effectiveness is self-explanatory and is mostly measured through the savings of time and labor for the instructor.

The vast majority of AES studies have measured AES performance through correlation with a human or multiple human graders. The other two methods used are multiple regression correlation and accuracy of results (error rate).

AES Studies

Since most AES scoring is based on training through essays graded by human scorers it is worth noting that in general human graders tend to give roughly equivalent scores to student essays when using the same grading rubric. This has been shown through research by commercial testing services such as ETS and through AES studies that employed multiple human graders. Carpenter, Delugach, Etzkorn, Farrington, Fortune, Utley and Virani (2007) found that there was a high correlation between human graders of the same student essay response in an engineering course. Despite the growth and improvement in AES technology acceptance in academia has been slow. Dreher, Reiners and Dreher (2011) found that while most instructors can define an AES, most have never used one for

scoring essays and have no interest in using one. The vast majority of instructors who have used an AES have done so for correcting multiple choice exams through the use of optical character scanning recognition (Dreher, Reiners & Dreher, 2011).

Homegrown Systems

Some interesting results have been achieved by researchers who have created their own custom designed AES systems. Most of these systems were created by STEM instructors particularly to assist with scoring of introductory programming exercises. In an early study, Johnson (1996) processed student essays through a commercially available grammar checker and compared the scores to those of six human judges. In correlating the human and AES scores he found that the AES outperformed three of the human judges and was nearly as reliable as the other three (Johnson, 1996). Olivier, Herson & Sosabowski (2001) found that pharmacology students preferred AES scoring over human scoring due to increased speed of feedback and grading, as well as the flexibility of working on and submitting their assignment from any location that had Internet access.

Cheang, Kurnia, Lim and Oon (2003) created an AES they called the Online Judge to score programming assignments in three courses at the National University of Singapore. They found an astonishing rate of plagiarism among their students (98 out of 712 assignments were plagiarized) but otherwise the system worked satisfactorily (Cheang, Kurnia, Lim & Oon, 2003). Unfortunately no other hard evidence was presented in the article. Duwairi (2006) used a self-created AES to score 200 essays in a database management course and correlated the scores with those of two human graders. Although human grader scores correlated more closely with each other than with the AES the results were close enough to justify continued use of the AES (Duwairi, 2006).

Blayney and Freeman (2008) used an Excel based AES to score accounting essays within a learning management system (LMS). Most of the students (82%) liked the instantaneous feedback of the AES and 76% believed the AES motivated them to learn and to keep trying until the correct answer was obtained (Blayney & Freeman, 2008). In another study, researchers created their own AES called Automatic Essay Assessor (AEA) based on other systems like PEG and IEA (Kakkonen, Myller, Sutinen & Timonon, 2008). AEA was able to score essays as accurately as two human graders and its grades achieved a correlation of .90 with the course instructor (Kakkonen, Myller, Sutinen & Timonon, 2008). Wang, Chang and Li (2008) scored 226 Taiwanese high school student responses to an essay question with three home grown AES systems. The first used a pure heuristics based grading (PHBG) method, the second was a data driven classification and the third a regression based grading (RBG) method (Wang, Chang & Li, 2008). The results of the three systems were compared to the scoring of two independent human graders. Although all three systems performed satisfactorily the best results were achieved by the data driven classification which achieved a correlation of .92 with the human graders (Wang, Chang & Li, 2008).

Some more recent studies have found issues with AES systems. One study which experienced negative results was conducted by Byrne, Tang, Tranduc and Tang (2010) using a self-created AES called eGrader to score 33 student essays in comparison with three human graders. Although correlation of .85, .75 and .74 was achieved by eGrader the researchers decided to discontinue its classroom use (Byrne, Tang, Tranduc & Tang, 2010). Student complaints led to the discovery that while the AES correctly scored objective elements the subjective elements of the essays were too complex for the AES to measure (Byrne, Tang, Tranduc & Tang, 2010). Barker (2011) created three prototype AES systems to score computer science assignments. Feedback was cut from four to six weeks to next day results and most of the students felt the scoring was fair and useful (Barker, 2011). However, many of the instructors considered the feedback provided too inflexible and as feedback was improved through each prototype more students began to challenge their grade (Barker, 2011). Escudeiro, Escudeiro and Cruz (2011) created an AES called d-Confidence which was used to score 31 answers to an essay question in a software engineering course. Correlation with a human grader (course instructor) was low as scores only correlated 68% of the time.

Ade-Ibijola, Wakama and Amadi (2012) scored 50 software engineering students essays using a home grown AES and compared the scores to those of a subject matter expert (SME). The correlation coefficient between the SME and the AES was only .71 (Ade-Ibijola, Wakama & Amadi, 2012). Matthews, Janicki, He and Patterson (2012) created a system called Adaptive Grading/Learning System (AGLS) to score assignments in a management information systems course which were compared to the grades assigned by teaching assistants (TA). AGLS grades were generally lower than TA grades because it found more errors (Matthews, Janicki, He & Patterson, 2012). In addition, the researchers found that feedback was greater with the AES, and response time was significantly

improved; however, there was no improvement in the quality of feedback provided to the students by AGLS (Matthews, Janicki, He & Patterson, 2012). In a study of Dutch students English proficiency a home grown AES achieved correlation coefficients of .87 with trained human graders (Santos, Verspoor & Nerbonne, 2012).

Commercial Systems

Surprisingly there has been little independent testing of commercial systems. Most of the data about correlation with human scorers is provided by the software developers themselves.

PEG

Page (1994) conducted a large scale study of PEG effectiveness using senior essays from the National Assessment for Educational Progress (NAEP). NAEP essays were scored by two human judges and Page recruited six more human judges to score each essay on a six point scale. The human judges achieved a multiple regression correlation of .877 with each other, in comparison PEG achieved a correlation of .869 (Page, 1994). Although these results were similar there were outliers which Page did not explain but which were probably caused by PEG's predilection to overrate long essays of dubious quality. The PEG system was sold by Dr. Ellis Batten Page to Measurement Incorporated in 2002. In January 2012, the Hewlett Foundation invited nine major vendors of artificial intelligence (AI) scoring of student essays to participate in the Automated Scoring Assessment Prize (ASAP) competition. AES system scores were correlated with the scores of two professionally trained readers. In January, the Hewlett Foundation invited Measurement Incorporated (MI) and eight other major vendors of artificial intelligence (AI) scoring of student essays to participate in the Automated Scoring Assessment Prize (ASAP) competition. The competition included essays written to eight different prompts by students in various grade levels. Each essay had been scored by two professionally trained readers. The human readers had agreement indices of .75. PEG achieved the highest agreement index with the human readers at 0.79 (PEG Software Leads Automated Essay Scoring Competition, 2012).

IEA

Landauer, Latham and Foltz (2000) used linear regression to compare IEA scoring with human graders on 3,926 essays on 15 diverse topics with a resulting correlation of .85. Even better results were achieved with 900 creative narrative essays from the GMAT with a correlation coefficient of .90 which was identical to that of two human graders (Landauer, Latham and Foltz, 2000). IEA is currently owned by Pearson Knowledge Technologies a subsidiary of Pearson Education. A recent Pearson white paper claims that its Oral Reading Fluency testing system (IEA based) achieved scores that correlate with human scores at 0.98, while the correlation between pairs of human raters was 0.99 (Streeter, Berstein, Foltz & DeLand, 2011).

E-rater

E-rater was used from 1999 through 2006 to score the GMAT. According to Valenti, Neri and Cucchiarelli (2003) the agreement rate between E-rater and human scorers of the GMAT on over 750,000 essays was over 97%. In 2006 the GMAT switched to IntelliMetric scoring which is based on the BETSY AES. E-rater is owned by ETS and currently is the software that runs the Criterion Online Writing Evaluation service. Although the ETS website provides a number of white papers and testimonials about Criterion none provide any hard data as to effectiveness.

BETSY

BETSY is now owned by Vantage Learning and is the software behind the IntelliMetric AES system; it also powers the My Access writing assessment tool. Since 2007 IntelliMetric has been used to score the GMAT. According to Valenti, Neri and Cucchiarelli (2003) BETSY achieved an accuracy rate over 80% on a test involving 462 essays. Dikli (2006) reports IntelliMetric in a test involving 8th grade student essays achieved an adjacent correlation scoring of .95 with human scorers and .99 with expert human scorers. According to the Vantage Learning IntelliMetric website when using a 6-point scale, two experts will agree with each other within 1 point about 95% of the time; IntelliMetric typically agrees with either expert about 97% to 99% of the time (IntelliMetric FAQs, 2012).

Markit

Markit was created by Robert Williams and Heinz Dreher of the Curtin University of Technology in Australia (Williams & Dreher, 2004). In a study of 20 essays in a business law class Markit scores were compared to the scores assigned by the course instructor (Williams & Dreher, 2004). The average human score was 61.75 while the Markit average was 62.35, the correlation between Markit and the human grader was .79 (Williams & Dreher, 2004). Williams (2006) in a study using Markit to score 290 high school student essays found a correlation of .79 with three human graders. Markit is freely available software available at http://www.essaygrading.com/project.jspx. The site does not provide any more recent evaluation information.

CURRENT STUDY

A pilot study using a vendor provided beta AES called Write Experience was used to score a legal fact pattern essay in an undergraduate Business Law course at a small liberal arts university in the Northeast United States. Students had one week to access the Write Experience web site to answer the question. The question was treated as a quiz score grade. Students also took six other in-class multiple choice quizzes during the semester. The lowest quiz grade was dropped and the quizzes accounted for forty percent of a student's final grade.

Participants

The business law class had an enrollment of twenty seven students. Students were mostly upperclassmen. All except one student were business majors. Only 23 students chose to participate in the online essay. Two students experienced technical difficulties and were unable to upload an answer. Twenty-one students provided a score-able response to the online essay question.

Instrument

Student reaction and perception of the Write Experience AES and their satisfaction with the online scoring process was measured through an online survey delivered through Survey Monkey. The survey consisted of 16 questions; the majority of items were close-ended Likert Scale questions. Several open-ended questions were also part of the survey. Because extra credit was given for responding to the survey anonymity was not provided although students were assured their responses would be confidential.

Methodology

A mixed methodology was used to analyze survey results. Likert scale questions were quantitatively analyzed while open ended questions were qualitatively analyzed.

Data Gathering

Students were provided with a link to the instrument posted in Survey Monkey. The survey was open for five days after which it was closed. Twenty three responses were obtained to the survey. Nineteen students who had provided usable essays responded, the two students who had experienced technical problems responded, and two students who had not taken the online essay responded as well. The high response rate was due to an extra credit incentive which provided responding students with 10 extra points on their lowest quiz grade. Respondents were 56.5% female and 43.5% male. 87% were seniors and 13% were juniors.

Quantitative Results

The students' experiences using the Write Experience software were mixed. 56.5% found the software easy/very easy to use, while 44.5% found it difficult/very difficult. The vast majority of the students (87%) did not use the My Tutor feature of the Write Experience software. Two-thirds of those who used My Tutor found it helpful in finishing the essay. My Editor was more heavily used – 52.2% used this feature while 47.8% did not. Most of the students who used My Editor found it helpful/very helpful (75%).

Most of the students found answering the legal fact pattern question to be difficult/very difficult 69.5%, while 30.5% found it easy. None of the students found answering the essay question very easy. Most students preferred the multiple choice in-class quiz format as 73.9% found the online essay more difficult. 26% of the students though the online essay was easier. Most of the students disliked the Write Experience online essay as 78.3% said they were unlikely or very unlikely to recommend using it to other students. Only 21.7% of the students like the software enough to recommend other students use it. The percentages were exactly the same for recommending a class to another student using the Write Experience software.

Qualitative Results

Thematic analysis was used to analyze the qualitative data the researcher collected through four open-ended questions. In the first step the researcher familiarized himself with the data by reading and rereading the responses. Next initial codes were created using the respondents own language. These codes were then interpreted and collapsed into over-arching themes that emerged as the data was analyzed. As a final step the themes were reviewed to discover patterns and to examine the more interesting themes in more detail.

Several interesting themes arose from the qualitative data. In response to a question on what they liked most about using the Write Experience software student responses centered around five main areas: they liked the user-friendliness of the software, they appreciated the opportunity to answer a question in more-depth and to demonstrate their knowledge and understanding of the material, they liked having unlimited time to answer the question and the opportunity to do it outside the classroom, they appreciated the opportunity to use other sources of information such as the text and their notes, and many were glad not to have to memorize answers or to take yet another multiple choice test.

In response to a question about what they most disliked about the Write Experience software many students complained that the software was not user-friendly, that the essay question was confusing or too difficult, that feedback/grading took too long and that they were unclear about how to properly answer the question or how much to write. The most interesting and unexpected response and one which was a consistent theme was the students discomfort with having an unknown entity grade and critique their work. Although they were told that a computer would grade their work many students did not understand the grading process or experienced anxiety about not having their professor assess their work. One response was typical of the students discomfort: “There was one thing I wasn’t too fond about. It worried me a bit that I had no idea who was grading my paper. I didn’t know their grading style, what they were looking for, etc. I would have felt a bit more comfortable if my professor was to read it and grade it since I would have a better idea of what he would want.”

In response to a question on what could improve the Write Experience software five overarching themes emerged: a better, clearer essay question, more user-friendly features, a rubric that explained the grading process, faster feedback/grading, and the opportunity to practice in class.
In a final comments section many students expressed antipathy toward the software complaining that they could not cut and paste their answers from a Microsoft Word document, and that the My Tutor and My Editor features did not work or were buried too deeply to be found. Many also wished they had the opportunity to practice using the software in the classroom or had the question requirements explained more clearly.

Discussion

For the most part student essay scores using the Write Experience software were low. Several students who wrote long answers and obviously thoroughly researched the topic did quite well. Several students who provided mediocre answers barely passed. The majority of the students failed the online essay. There were several reasons for the low grades. Some students admitted they did not take the assignment seriously and made only a minimal effort to answer the question. Most of the students had no experience answering a legal fact pattern question and did not know the proper method of answering such a question. The one non-business student who was an administration of justice major familiar with this type of question scored quite well on the essay. Another factor is that many business students are simply very poor at answering essay questions or doing any type of writing. Due to course size the majority of their exams are multiple choice and they are assigned few papers to research and write. This pilot project

highlights the need for the business department to assign more essay exam and research papers.

Several students experienced frustration with the Write Experience software. Besides the two students who failed to post an answer, several others mentioned that they were unable to paste their answers from Microsoft Word into the Write Experience interface. Two features of the Write Experience interface – My Tutor and My Editor were heavily underutilized by the students. My Editor was a spell checker and grammar corrector which could be used to correct style and spelling errors before final submission of the essay. Such errors would result in a lower score on the essay. My Tutor provided input into the content of the student's essay before submission. My Tutor would suggest that further content should be provided, or alternate approaches to answering the question should be used. Students who followed the My Tutor advice should provide a more complete essay answer and thus achieve a higher score. Few students used these two features. Some claimed they never saw them, and others said they tried to use them but they did not help or work properly.

Limitations and Future Research

Results of this study are not generalizable as they were achieved through a sample of convenience in one small business law course. Results using other majors or larger class sizes may differ considerably. Because the Write Experience software was still in beta and some students experienced technical problems this could have affected student perceptions. Studies using other AES systems or comparing other systems to Write Experience would make valuable contributions to the literature.

Conclusions

Although AES systems can be helpful with automatically grading essay questions their usage in a small course is probably ill-advised. The technical requirements for adopting a question and assigning a user name/password to each student is considerable. For the instructor familiarizing oneself with the administration side of the software is time consuming. There is also a definite learning curve for the student in learning how to access and properly use the interface. For better results the instructor should use class time to demonstrate how to use the interface to the students. For this particular AES scoring of essays took much too long. The essays could have been graded by hand in several hours yet the students did not receive their scores until six weeks after logging into the software. Student satisfaction with the Write Experience software was low. The vast majority of students did not enjoy the process and would prefer in-class or take-home essay questions. Since the class size was not that large the instructor will incorporate more fact pattern essay questions in future courses but will grade them by hand. This will help with formative assessment while also increasing the research and writing skills of the students.

REFERENCES

Ade-Ibijola, A.O., Wakama, I. & Amadi, J.C. (2012). An expert system for automated essay scoring (AES) in computing using shallow NLP techniques for inferencing. *International Journal of Computer Applications,* **51**(10), 37-45.

Barker, T. (2011). An automated individual feedback and marking system: An empirical study. *The Electronic Journal of e-Learning*, **9**(1), 1-14.

Berret, D. (2013). English teachers reject use of robots to grade student writing. *Chronicle of Higher Education,* **59**(34), A1-A4.

Blayney, P. & Freeman, M. (2008). Individualised interactive formative assessments to promote independent learning. *Journal of Accounting Education, **26**,* 155-165.

Byrne, R., Tang, M., Tranduc, J. & Tang, M. (2010). *Journal of Systemics, Cybernetics & Informatics,* **8**(6), 30-35.

Carpenter, S.L., Delugach, H.S., Etzkorn, L.H., Farrington, P.A., Fortune, J.L., Utley, D.R. & Virani, S.S. (2007). A knowledge modeling approach to evaluating student essays in engineering courses. *Journal of*

Engineering Education, **96**(3), 227-239.

Cheang, B., Kurnia, A., Lim, A. & Oon, W.C. (2003). On automated grading of programming assignments in an academic institution. *Computers & Education, **41**,* 121-131.

Dikli, S. (2006). An overview of automated scoring of essays. *The Journal of Technology, Learning and Assessment,* **5**(1), 1-35.

Dreher, C., Reiners, T. & Dreher, H. (2011). Investigating factors affecting the uptake of automated assessment technology. *Journal of Information Technology Education,* **10***,* 161-181.

Dreher, H. (2006). Interactive on-line formative evaluation of student assignments. *Issues In Informing Science & Information Technology,* **3**, 189-197.

Duwairi, R.M. (2004). A framework for the computerized assessment of university student essays. *Computers in Human Behavior,* **22***,* 381-388.

Escudeiro, N., Escudeiro, P. & Cruz, A. (2011). Semi-automatic grading of students' answers written in free text. *Electronic Journal of e-Learning,* **9**(1), 15-22.

Giles, J. (2011). AI makes the grade. *New Scientist,* **211**(2828), 22-25.

Hearst, M.A. (2000). The debate on automated essay grading. *IEEE Intelligent Systems & Their Applications,* **15**(5), 22-27.

Johnson, V.E. (1996). On Bayesian analysis of multirater ordinal data: An application to automated essay grading. *Journal of the American Statistical Association,* **91**(433), 42-51.

Kakkonen, T., Myller, N., Sutinen, E. & Timonen, J. (2008). Comparison of dimension reduction methods for automated essay grading. *Educational Technology & Society,* **11**(3), 275-288.

Lai, Y.H. (2010). Which do students prefer to evaluate their essays: Peers or computer program. *British Journal of Educational Technology,* **41**(3), 432-454.

Landauer, T.K., Laham, D. & Foltz, P.W. (2000). The Intelligent Essay Assessor. *IEEE Intelligent Systems & Their Applications,* **15**(5), 27-31.

Matthews, K., Janicki, T., He, L. & Patterson, L. (2012). Implementation of an automated grading system with an adaptive learning component to affect student feedback and response time. *Journal of Information Systems Education,* **23**(1), 71-83.

McCollum, K. (1998). How a computer program learns to grade essays. *The Chronicle of Higher Education,* **45**(2), A37- A40.

NCTE Position Statement on Machine Scoring. (2013). National Council of Teachers of English, 1-12. Retrieved from http://www.ncte.org/positions/statements/machine_scoring

Olivier, G.W.J., Herson, K. & Sosabowski, M.H. (2001). WebMark – A fully automated method of submission, grading and commentary for laboratory practical scripts. *Journal of Chemical Education,* **78**(12), 1699-1703.

Page, E.B. (1994). Computer grading of student prose, using modern concepts and software. *The Journal of Experimental Education,* **62**(2), 127-142.

Perelman, L.C. (2013). Critique of Mark D. Shermis & Ben Hammer, Contrasting state-of the-art automated scoring essays: Analysis. Retrieved from

http://graphics8.nytimes.com/packages/pdf/science/Critique_of_Shermis.pdf

Santos, V.D.O., Verspoor, M. & Nerbonne, J. (2012). Identifying important factors in essay grading using machine learning. Preprint of paper submitted to Tsagari, D. (ed.) for inclusion in *Selected papers in memory of Dr. Paulus Parlou – Language testing and assessment around the globe: Achievements and experiences*. Frankfurt, Germany: Peter Lang.

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

Williams, R. (2006). The power of normalized word vectors for automatically grading essays. *Issues in Informing Science & Information Technology,* ***3*****,** 721-729.

Williams, R. & Dreher, H. (2004). Automatically grading essays with Markit. *Issues in Informing Science & Information Technology,* **1***,* 693-700.

Williamson, D.M., Xi, X. & Breyer, F.J. (2012). A framework for evaluation and use of automated scoring. *Educational Measurement: Issues and Practice,* **31**(1), 2-13.