ELECTRONIC ESSAY EVALUATION (EEE)

Herman Brooks, TSgt, USAF
AETC Technology Requirements Branch
San Antonio, Texas

Ken Levi, Ph.D.
AETC Studies and Analysis Squadron
San Antonio, Texas

ABSTRACT

Air Education and Training Command Studies and Analysis Squadron (AETC SAS) in conjunction with the Air Command and Staff College (ACSC) at Maxwell AFB, Alabama is assessing innovative technology for computer assisted evaluation and feedback of essays. The focus of study is the first examination given to all students enrolled in the non-resident program at ACSC. The course subject matter is National and International Security Studies. Educational research has long recognized that essays are more effective for measuring depth of comprehension than multiple-choice examinations. Unfortunately, the process of evaluating, grading, and providing feedback of essay based material is resource and labor intensive. Successful results would allow the non-resident program at ACSC to use a more probing test instrument, essay evaluation, to measure student comprehension through essays vs. multiple choice questions, without an increased demand on resources or labor. Essay examinations completed by resident ACSC students were used to calibrate and validate enhanced computer assisted essay grading software, which will then be used to score essay tests obtained from a sample of the non-resident program students in the fall term 2002. Results from pre-implementation baseline and post-implementation will be analyzed to determine effectiveness.

Biographical Sketch:

TSgt Herman Brooks Jr. is stationed at the Air Education and Training Command Studies and Analysis Squadron, Randolph AFB, TX. He is currently a Test Director in the Technology Innovation Flight, and has served as test team lead for Electronic Classroom and Schematic Power Browser.

Dr. Ken Levi is stationed at the Air Education and Training Command Studies and Analysis Squadron, Randolph AFB, TX. He is currently a Test Director in the Technology Innovation Flight, and has served as team lead for artificial intelligence and small computer development.

ELECTRONIC ESSAY EVALUATION (EEE)

Herman Brooks, TSgt, USAF
AETC Technology Requirements Branch
San Antonio, Texas

Ken Levi, Ph.D.
AETC Studies and Analysis Squadron
San Antonio, Texas

INTRODUCTION

Air Education and Training Command (AETC) Studies and Analysis Squadron (SAS) in conjunction with Air Command and Staff College (ACSC) at Maxwell AFB AL is investigating recently developed technology that provides computer assisted evaluation and feedback of essay materials. The testbed for this project is the resident and non-resident ACSC lesson in National and International Security Studies. AETC SAS will determine whether essay evaluation technology provides meaningful feedback for student depth of subject matter comprehension without increasing manpower requirements.

Previous research

Essay vs. multiple choice tests

Research has found essays superior to multiple choice tests in evaluating higher levels of student learning. Bloom s taxonomy (1956), guideline for all Air Force training and education, outlines six levels of abstraction, starting from basic knowledge level at the bottom up to evaluation level at the top. Multiple choice tests have proven adequate for the knowledge or comprehension levels. But essays excel at measuring higher cognitive levels of learning (Opitz, 2001). In classrooms using multiple choice tests, instruction tends to emphasize drill and practice on decontextualized skills (Wiggins & McTighe, 1998). Essays promote teaching for higher conceptual understanding.

Daraghmeh (1997) conducted a study at An-Najah National University in Nablus, Palestine. He compared 60 students taking essays to 60 students taking multiple choice. He then retested both groups with a combination of essay and multiple choice questions. The essay students learned more about general knowledge than their multiple choice counterparts. Daraghmeh concludes that essays are best to increase [student] academic achievement on all levels of learning.

Despite the proven superiority of essays over multiple choice exams for higher learning, many educators have abandoned written tests because grading them is so time consuming and labor intensive (Thompson, 1999). This dilemma is particularly acute in the area of military education. In contrast to military training, the aim of military education is to instill higher levels of learning, but the number of students involved is often so large as to make essay exams impractical. Electronic essay evaluation (EEE) provides a possible solution to this dilemma.

Potential benefits of EEE

If EEE worked, it could save considerable manpower and funding, while simultaneously elevating the standards of education. Already, the use of *e-rater* to grade a half million Graduate Management Admissions Tests is estimated to save $1.7M annually (Kukich, 2000; Kladko, 2001). Now, each essay is graded by one human and the machine, instead of two humans as before. Another benefit is immediate feedback. Instead of waiting days or weeks for an instructor to grade an essay, students can receive a computerized response almost immediately. Finally, computers are more objective than humans. Instructors are subject to grader drift, a propensity to raise or lower grading standards unconsciously over time (Thompson, 1999). Computers don t have mood swings.

History of EEE

Attempts to automate the grading of essays began in the 1960 s with Ellis Page s Project Essay Grader (PEG). The program was based on surface features, such as word length, essay length, or the number of commas and prepositions. The technology leaped forward in the 1990 s with the evolution of natural language processing (NLP). *E-rater*, from the Educational Testing Service, based it s evaluations on a form of word matching, consisting of the weighted frequencies of vocabulary terms. In 1998, Knowledge Analysis Technologies (KAT) introduced a new methodology Latent Semantic Analysis (LSA) that enables computers to go beyond mechanics and syntax and attend to what

students are saying, rather than on how exactly they are saying it (Kukich, 2000).

Latent semantic analysis

Unlike earlier EEE technologies, LSA focuses on content over form. Instead of specific words and phrases, it identifies the underlying meanings behind them (Landauer, 2000). LSA can recognize that the phrase the doctor operated on the patient is equivalent to the surgeon wielded the scalpel (Kladko, 2001).

The technology involves building a semantic web, in which words are weighted in terms of their co-occurrence with each other and located within a vast concept space. Relative position within the space denotes the similarity of meaning between words and passages (Landauer, 2000).

LSA learns not only the synonyms for words but also their proper contexts. For example, the word fly is related to zipper. But when used in the phrase high fly, LSA gave it a .31 similarity rating to baseball and a .37 rating to ball. The similarity to zipper, however, was assessed at only .03 (Thompson, 1999).

Typically, semantic networks are built from training sets of 100-300 essays (Landauer, 2000). These essays have already been graded by humans, and LSA learns from them how to distinguish between good and bad answers. An alternative training method for the software is to use a gold standard, a model answer on which all other essays are judged (McCollum, 1998).

LSA strengths and weakneses

The main strength of LSA is that it focuses on content and judges essays by substance. It can also provide immediate feedback, letting the writer know content areas covered poorly or well.

On the other hand, Tom Landauer, KAT founder and CEO, freely admits that LSA is not meant to assess creativity. Nor does it judge the logic or coherence of a student s argument (Landauer, 2000). No particular regard is paid to the quality of the student s vocabulary, either. As Calfee (2000) observes, LSA only reinforces the tendency among today s high-schoolers to eschew a risky lexicon. You wouldn t use LSA in a course on creative writing. For fact-based subjects, like science or history, however, LSA should be ideal.

LSA reliability

Like most other EEE technologies, LSA boasts a high success rate. The correlation of grades among instructors ranges from 0.75 to 0.85 (Kukich, 2000). If computer grading can meet or exceed this range, it s considered as good as human. According to KAT, LSA scores typically correlate with human scores at the 0.85-0.91 level. This is similar to the inter-rater reliability for *e-rater* of 0.87-0.94 or for PEG of 0.78 (Kukich, 2000; Daniel & Cox, 2001).

LSA applications

To date, LSA has been applied at the civilian high school and college levels, and, to a limited extent, in the military. LSA gives Boulder CO sixth graders much more feedback than a teacher could normally provide to a class of 20-30 students (Kladko, 2001). The University of Colorado, Florida State University and New Mexico State University also have LSA applications (McCollum, 1998; Thompson, 1999).

Within the military, a preliminary trial of LSA is currently ongoing at the Army s Combined Arms and Services Staff School (CAS3) at Ft. Leavenworth, KS (Crowson, 2002).

The jury is still out on the success of the CAS3 venture. According to Lieutenant Colonel Mark Crowson (2002), results so far are not encouraging. The match between LSA scores and instructor scores on the same essays is low. Current inter-rater reliability is a disappointing 0.50. The computer system does a fair job of detecting grammar and spelling errors, but can t grade content.

To date, LSA has not been applied within the Air Force. Implementation at Air University provides an AF case study to parallel the application of LSA in the Army. Both studies provide a foundation for assessing the effectiveness of EEE in military education.

.

STUDY APPROACH

Calibration and implementation

Essay examinations completed by AY 02 resident ACSC students, and manually graded by their instructors, were used to calibrate and validate enhanced computer assisted essay grading software. The software then scored essay tests from a separate sample of fall 2001 resident program students. Computer and instructor scores were compared.

After calibrating and testing, the LSA model is ready for implementation. Plans are to deploy the system on a subset of the 7000 non-resident students in the fall term 2002. Essay tests will be substituted for the multiple choice tests normally given to distance learning students. Grades assigned to these essay tests will be compared to a subset of tests scored by an instructor.

National and International Security Studies (NISS)

The subjects of this study are students in NISS at the Air Command and Staff College (ACSC), Maxwell AFB AL. This is one of two lessons in the 57 hour International Relations course. NISS students learn the purpose and objectives of National Security Strategy (NSS), the role of military force in NSS, and strategic concepts in national military strategy. Resident students attend lectures and break off into smaller sections of 13-14 members each for discussion and assignment. Non-residents receive a compact disk containing all course materials for self study. Whereas residents currently take essay tests, all non-resident tests are multiple choice. By converting those multiple choice exams into essays, this project aims to make the educational experience of the two student groups more comparable.

STUDY RESULTS

There are two key aspects of this project. First is establishing confidence in newly-developed software tools that evaluate essay materials as effectively as a human might. Second is demonstrating that such software can perform effectively when scaled up from the research environment, providing accurate and meaningful assessments on thousands of essays in less time than humanly possible. Both of these aspects are addressed in the following phases:

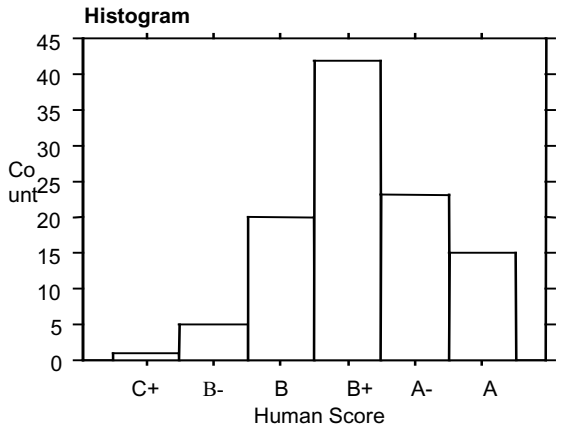
Calibration

The first calibration

The contractor Knowledge Analysis Technologies (KAT) trained the LSA software with essays. These essays, in electronic format, had been taken by NISS students in the fall 2001 and manually graded by instructors. Of the total 302 essays sent to KAT, 106 were used in the software calibration.

Essays were graded on a six point scale (2.3 = C+, 2.7 = B-, 3.0 = B, 3.3 = B+, 3.7 = A-, 4.0 = A). The human scores (Figure 1) distributed around a mean of 3.4 with a modal response of B+. Only 6 of the 106 essays earned scores below a B (3.0).

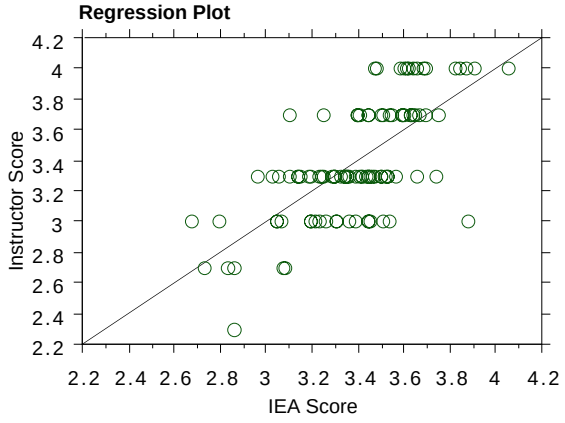
Figure 1. Distribution of Human Grades



Inter-rater reliability is calibrated by comparing instructor grades to computer grades. Figure 2 shows a scatterplot of the scores.

Results from the first calibration were reported in December 2001. The inter-rater reliability correlation between the human instructor and the computer was 0.70.

Figure 2. Inter-rater Scatterplot



Ordinarily KAT would compare the 0.70 obtained for computer-human reliability to the human-human reliability number. The NISS essays, however, were only graded by one instructor apiece. So, the extent to which two or more humans would grade any given essay the same could not be determined.

Agreement

Agreement statistics are calculated by combining all scores falling within one point of each other. Ninety-two percent of the computer graded scores fell within one point of the human scores. As figure 3 shows, half of the computer graded essays were the same as scored by instructors. Another 45 were scored within 1 point. This result is somewhat mitigated by the fact there are only six points to begin with.

Figure 3. Inter-rater Agreement

	2.3	2.7	3	3.3	3.7	4	
2.3	**0**	1	0	0	0	0	**1**
2.7	0	**3**	2	0	0	0	**5**
3	0	2	**7**	8	2	1	**20**
3.3	0	0	10	**24**	6	2	**42**
3.7	0	0	1	7	**11**	4	**23**
4	0	0	0	2	5	**8**	**15**
	0	**6**	**20**	**41**	**24**	**15**	**106**

	Instructor-IEA (%)
Exact Agreement	**50.00**
Adjacent Agreement	**92.45**
Exact Agreement at each score (%)	
2.3	**0.00**
2.7	**60.00**
3	**35.00**
3.3	**57.14**
3.7	**47.83**
4	**53.33**
Adjacent Agreement at each score (%)	
2.3	**100.00**
2.7	**100.00**
3	**85.00**
3.3	**95.24**
3.7	**95.65**
4	**86.67**

Exact
Adjacent

The second calibration

The results of the first calibration were somewhat disappointing. The KAT standard is 0.75 (MacCuish, 2002), although they will accept scores as low as 0.60 (Laham, 2002). Because of some irregularities in the data, KAT decided to clean up the data and try again. Unfortunately, the results of the second calibration were equivalent to the first.

Test

Resident student test

After the second calibration, KAT used the calibrated software to grade the remaining 196 essays which they had received from ACSC. This test set of essays was kept separate from the initial training set to see how well the software could grade essays which it had never seen before. The inter-rater reliability between instructor and computer scores on these 196 exams was 0.50 (Laham phone conversation, May 15, 2002).

Non-resident test

The final stage of the study will be to apply the calibrated software to non-resident student essay exams in the fall 2002. Results should be available in October.

Student comments

After the non-resident test is complete, participating students will be asked to fill out critiques, and to comment about the experience of taking distance learning essays, versus multiple choice exams.

DISCUSSION

This project examined whether EEE would be appropriate for both summative and formative evaluations. In a summative evaluation, test scores are computer graded. Instructors then have the option of reviewing the essays manually or simply submitting the grade provided by the software.

In a formative evaluation, the computer scores don t count. Instead, the student uses these scores and the feedback that the computer provides to improve his/her performance. The EEE becomes a way to prepare for the course exam.

After receiving results of the LSA calibration, ACSC administrators concluded that, at least for resident students, electronic essay evaluation would not be used for summative evaluation. Focus shifted, instead, to formative applications.

The nonresident course has an optional portfolio requirement. Students are asked to respond to six writing prompts. For the 2002 students, this requirement will become mandatory. LSA may be used for grading and feedback of these exercises.

According to KAT (Laham, 2002), however, the portfolio grading application would employ a different kind of EEE technology. Instead of latent semantic analysis, the software would detect mechanical features, such as word and essay length, which are indirectly correlated to essay quality. Student then would receive feedback about what quintile they occupy (top fifth, bottom fifth, etc.) relative to the rest of the class. This would provide students both the opportunity to demonstrate the higher conceptual learning that essays elicit as well as formative feedback to prepare them for their final exams.

SUMMARY

The purpose of this study was to determine whether EEE technology could provide AF instructors with meaningful feedback about the depth of student comprehension, without increasing manpower. Military education provides a useful trial for EEE technology because the use of essay exams to instill higher learning has been impeded by the daunting task of grading large numbers of students, especially the non-residents. The current study at Air University provides an Air Force counterpart to a similar effort at the Army s Combined Arms and Services Staff School.

Latent semantic analysis was chosen as the preferred EEE technology because of its ability to judge student essays by their content, rather than by surface, or formal, features such as number of commas or key words.

Results, so far, are disappointing. The calibration on a training set of 106 NISS students yielded a mediocre .70 inter-rater reliability between humans and the computer. Subsequent application of the calibrated model to a test set of 196 NISS essays produced an inter-rater reliability of .50. This figure is the same as obtained by CAS3.

Based on findings to date, AU administrators shifted plans for the use of LSA from summative to formative evaluation. On the one hand, this would avail a large number of students the opportunity to glean the benefits of essays. On the other hand, even the formative employment of LSA will only succeed as long as students remain unaware that they are being judged not by substance, not by the content of their writing, but rather by their form.

REFERENCES

Bloom, B.S. (Ed.). (1956). *Taxonomy of educational objectives: the classification of educational goals: handbook I, cognitive domain.* New York; Toronto: Longmans, Green.

Calfee, R.. (2000, September/October). To grade or not to grade. In the debate on automated essay grading. *IEEE Intelligent Systems*, 35-37.

Crowson, M. (2002). *Information provided to author*.

Daniel, G. & Cox, K. (2001). Automated essay grading. *Web Tools Newsletter.* Retrieved May 3, 2002, from http://www.tools.cityu.edu.hk/news/newslett/.

Daraghmeh, A. (1997). The effect of question s types

and levels on student s academic achievement. Retrieved, May 3, 2002, from wysiwyg://50/ http://www.geocities.com/Athens/Rhodes/ 2713/Abstract.html.

Kladko, B.. (2001). Computer programs pass judgement on student writing. *The Record, 1-6.*

Kukich, K. (2000, September/October). Beyond automated essay scoring. In the debate on automated essay grading. *IEEE Intelligent Systems*, 22-27.

Laham, D. (2002). *Information provided to author.*

Landauer, T., Laham, D., & Foltz, P. (2000, September/October). The intelligent essay assessor. In the debate on automated essay grading. *IEEE Intelligent Systems*, 27-31.

MacCuish, D. (2002). *Information provided to author*.

Opitz, M. (2002). *Knowledge base for teacher education.* Long Island: College of Saint Benedict, Saint John s University.

Thompson, C. (1999, July/August). New word order: the attack of the incredible grading machine. *LinguaFranca,* 28-37.

Wiggins, G. & McTighe, J. (1998). Depth vs. coverage in teaching and standardized test prep. In G. Wiggins and J. McTighe (Eds.), *Understanding by design* (pp. 131-133).