THE CREDIBLE GRADING MACHINE:
AUTOMATED ESSAY SCORING IN THE DOD

Lynn Streeter
Knowledge
Analysis
Technologies
Boulder, CO

Joseph Psotka
Army Research
Institute
Alexandria, VA

Darrell Laham
Knowledge
Analysis
Technologies
Boulder, CO

Don MacCuish
Air Command and
Staff College
Maxwell AFB, AL

The Intelligent Essay Assessor is commercial software that grades essays as accurately as skilled human graders. It was used to critique senior officers' papers in both the Army's and Air Force's Command and General Staff Colleges. The Army Research Institute has supported development of this software, which understands the meaning of written essays. Automatic essay scoring is well-suited to distance learning environments, and for faculty training and calibration. Because the essay feedback is returned in seconds and can indicate which sections should be rewritten, students can make significant revisions before submitting their final product. This tutorial facility could be exploited in many military courses.

In the Army's Combined Arms and Services Staff School (CAS3), Military Writing assignment memos were graded by both the instructors and a subset by recently retired instructors. The results showed that human-to-human reliabilities (Leavenworth graders-to-retired instructors) were identical to the computer-to-Leavenworth graders reliabilities for the overall grade. In addition, the essay grading software was enhanced to supply written tutorial feedback similar to comments given by instructors, including (1) format checking, (2) section critiquing (e.g. Background, Purpose, etc.), returning recommendations of sections needing revision, and (3) plagiarism detection.

The Air Command and Staff College project is exploring the effectiveness of automating the grading of the written examination used for the "National & International Security Studies" course for both residents and distance learners. In this trial, the Intelligent Essay Assessor was used to assess longer papers, averaging over 2000 words, and grades were compared to two faculty members' grades. Again, the automated method was as reliable as human graders. Plans are underway to use the automated facility for formative evaluation, which means that students, not faculty, will review the assessment provided by the software, and use that feedback to formulate a better response prior to final submission in a portfolio writing exercise.

Biographies

Lynn Streeter has worked on officer training and leadership applications, field trials of essay scoring, and other measurement applications based on Latent Semantic Analysis. She holds a Ph.D. from Columbia University, and has advanced Computer Science training and experience, and formal executive and business training. She was a Research Director at Bell Labs and Bellcore where she directed human computer interaction work, AI and linguistics research, and software engineering research. She participated in the invention and early applications of LSA. Thereafter she was a General Manager of a R&D division of U S West Advanced Technologies.

Joseph Psotka is a research psychologist at the Army Research Institute where he has been a Senior Scientist in the Training Laboratory, in Basic Research, and in Organization and Personnel Research. He holds a Ph.D. from Yale University. He was also a Program Director for the Applications of Advanced Technologies program at the NSF. His research currently focuses on the application of cognitive science and automated technologies to learning and instruction, with a special emphasis on their role in developing leadership knowledge, skills, values, and attitudes.

Darrell Laham is Chief Technology Officer at KAT and is an originator and patent holder of the Intelligent Essay Assessor technology. He holds a Ph.D. in Psychology and Cognitive Science from the University of Colorado at Boulder. Dr. Laham has been the chief architect and project leader on several previous and ongoing projects in education, training, and workforce management for the Army, Navy, Air Force, DARPA, NASA, and NIST, as well as commercial applications of KAT's educational technologies.

Don MacCuish, Ed.D. is Associate Dean and Professor of Distance Learning at the Air Command and Staff College. He has over twenty years experience in the design, development, implementation and evaluation of training programs on the secondary, undergraduate and graduate level.

THE CREDIBLE GRADING MACHINE:
AUTOMATED ESSAY SCORING IN THE DOD

Lynn Streeter
Knowledge
Analysis
Technologies
Boulder, CO

Joseph Psotka
Army Research
Institute
Alexandria, VA

Darrell Laham
Knowledge
Analysis
Technologies
Boulder, CO

Don MacCuish
Air Command and
Staff College
Montgomery, AL

INTRODUCTION

Many Department of Defense, government and civilian organizations are attempting to reduce training costs while improving overall job performance. The major cost components are the instructor's student contact hours and the time students spend away from their jobs. For the branches of the Armed Forces, the problems are exacerbated by downsizing and recruitment and retention difficulties. For the world's largest trainer, training as usual is not an option. The desired end state is to cut real costs, decrease the number of instructors and training time, and simultaneously improve student learning and performance.

The relevance of new learning technologies--particularly effective ones--to the DoD is clear. In order to increase the number of personnel who can be trained and keep costs down, the DoD is using more distance learning in enlisted and officer education. Some examples include:

- The U. S. Army awarded a five-year contract to PricewaterhouseCoopers to develop the Army University Access Online Program, an educational Web portal designed to provide distance learning to U.S. soldiers worldwide:

  The Army University Access Online Program will enable U.S. enlisted soldiers to use an education Web portal to take distance learning courses and earn certificates, associate's degrees, bachelor's degrees, and master's degrees while they continue to serve. The Army estimates that 12,000 to 15,000 soldiers will participate in the program in 2001 - the program's first year of operation - and that as many as 80,000 soldiers will participate in the program by its fifth year. (http://www.pwcglobal.com/extweb/ncpressrelease.nsf)

- The Army and Air Force are replacing residential officer study with distance education modules. All branches of service have large nonresidential training programs.

For example, The U. S. Army War College as well as the Command and General Staff College offer their degree programs through nonresident distance learning programs.

To deliver high quality instruction and assessment, distance learning training must find novel ways to incorporate the coaching and feedback that make human instructors so special. In addition, there is a need for deeper assessment to assure that job relevant knowledge and skills have been acquired. Constructed response questions (as opposed to multiple-choice questions) have greater face validity for this purpose (Birenbaum, & Tatsuoka, 1987) and, as a result, are used extensively in military education and training. Expressing what has been learned in an essay promotes learning and makes for better writers and communicators (Chi, 2000; Greenwald et al., 1998; Kintsch et al. 2000; Steinhart, 2000). However, to promote learning itself, we need assessments that are returned in seconds, not days and months, contain useful information about what the student has done well and not, and offer good advice about what to do next. Currently, essay tests that return substantive commentary and guidance, such as pointers to sources of missing information or identification of conceptually wrong, empty, or redundant sentences are far too labor intensive to be used often. Comments from instructors are rarely returned in time to be of optimum pedagogical value (Warfield, Johnstone, and Ashbaugh, 2002). Indeed, the difficulty of providing substantive feedback encourages the use of essays primarily to teach content-independent writing skills such as spelling, punctuation, and grammar, rather than the more important matter of effectively conveying a message.

There are some notable advantages to machine scoring of essays:

- A computer can be given the time and resources to examine an almost unlimited amount of relevant source material before being used to score essays on a particular prompt.
- A computer can examine and analyze essays in much more detail than a human.

- A computer can compare every essay in a set of virtually any size with every other—something that would be impossible for a human to do for a mere 300 essays, in a year of work, at three minutes per comparison.
- A computer can be completely consistent in its evaluations, from essay to essay, from time to time. It will not get tired, bored, irritated, or inattentive, nor will its standards drift.
- A computer can be entirely objective and without bias based on acquaintance with students or extraneous knowledge of their characteristics.
- A computer can perform many complex and sophisticated analyses that humans are incapable of computing without assistance.
- A computer can be free of the reasoning and judgment errors, myths, false beliefs, and value biases that plague all human judges.

This is not to deny that there are many things that only humans can do with an essay (Palincsar, Brown, & Campione, 1994). There is still much to be learned from excellent human graders and editors. The state-of-the-art with respect to computer generated commenting and critiquing falls far short of a skilled person.

Currently, there are three companies that offer automated essay scoring services: ETS Technologies with its e-rater system (http://www.ets.org/), Vantage Learning (http://www.vantagelearning.com/) with Intellimetric, and Knowledge Analysis Technologies with its Intelligence Essay Assessor (IEA) (http://www.knowledge-technologies.com/). All three service providers produce holistic grades that are comparable to human grades, although the methods for calculating those grades differ among the three providers. ETS Technologies emphasizes natural language processing for scoring grammar and organization, while Knowledge Analysis Technologies uses Latent Semantic Analysis and other machine learning and statistical methods for scoring. The IEA emphasizes content coverage as the primary determinant of score, but includes scoring of grammar, mechanics, style, and organization when appropriate. Vantage Learning does not disclose its methods beyond a general acknowledgement of the use of Artificial Intelligence technologies.

Automated essay grading is now moving into the assessment mainstream. For example, as of 2001 the second grader for the GMAT (Graduate Management Aptitude Test) essays is *e-rater* rather than a second human scorer. Both ETS Technologies

(http://www.ets.org/research/erater.html) and KAT (http://www.knowledge-technologies.com/ifr-iea.html) have working demonstrations of their product on their web sites allowing interested parties a chance to test drive the technology. With most of the vendor applications, a holistic score is returned and sometimes analytic scores, such as scores for content, mechanics, organization, focus, etc. All three providers of automatic essay scoring have tested their product on tens of thousands of essays from students of many different ages and have products used both for summative assessment and for practice testing.

While all have products directed to language arts and standardized writing tests such as K-12 state writing assessments and placement tests for college writing courses, only the IEA has been tested in topics ranging from biology, history, psychology, and military leadership.

The research reviewed herein was conducted using the Intelligent Essay Assessor technologies under Army Research Institute (ARI) and Air Education and Training Command (AETC) ETTAP contracts awarded to Knowledge Analysis Technologies.

THE INTELLIGENT ESSAY ASSESSOR

The Intelligent Essay Assessor (IEA) judges the quality of the overall content of an essay as reliably as skilled human graders over a large range of topics. Figure 1 shows the reliability between human graders and IEA for different types of essays—standardized essays, such as the College Board's Graduate Management Aptitude Test (GMAT) essays and classroom essays primarily in college courses across topics ranging from history to psychology to physiology.

Inter-rater reliability for standardized and classroom tests
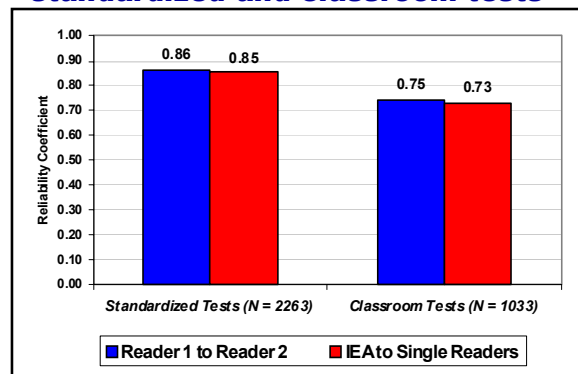
Figure 1. Comparison of human-to-human reliabilities to human-to-Intelligent Essay Assessor reliabilities across standardized and classroom tests.

Inter-rater Reliability as Criterion

The inter-rater reliability level is the standard criterion used for the Intelligent Essay Assessor. (Other vendors often use percentage of agreement on the score points, which is a somewhat less sensitive measure.) The scores given by the IEA are compared to the scores given by a human reader for a particular set of essays. If the reliability level between the IEA and a human reader is comparable to the reliability level between two human readers, the model is considered a success. The reliability is evaluated through a correlation coefficient based on the score pairs for the set. A reliability of 1.0 indicates absolutely perfect agreement between readers. In a test of over 3,000 essay grades IEA had human vs. system reliabilities between 0.70 and 0.80. The Educational Testing Service found that grader-to-grader reliabilities in their exams ranged from 0.50 to 0.85 (Braun, 1988). This wide range of reliabilities is important when evaluating the Air Force and Army results, as the reliabilities observed were lower than those previously encountered by us.

How the IEA Works

The IEA measures abstract factual knowledge based on extensive background readings, texts, and news sources, not just superficial factors such as word counts, word length, keywords, or punctuation. The IEA's assessment focuses on the understanding of the subject matter that goes into the creation of an essay. For example, before scoring essays on military leadership, IEA reads all of FM 22-100, Military Leadership, other relevant essays that have been graded by senior officers, and books such as Woodward's "The Commanders". From its reading IEA constructs a very large semantic network of all the words in all the contexts found in the background texts. It can also supplement this text, if it is too specialized, with a representative sample of the kinds of materials an average American reader encounters. The semantic space that is created from all these materials permits IEA to read any essay and understand the many synonyms and alternate ways of stating the same important ideas.

IEA is based primarily on meaning and content, but it also can measure mechanics and style and other analytic traits of the writing (See Figure 2). While the exact measures and how these are combined are proprietary information, how they are computed in general is as follows.

IEA's content measures are based on Latent Semantic Analysis (LSA). IEA computes the overall content similarity in LSA space between a new essay and essays on the same topic that have been graded by

humans, and next determines the nearness of the new essays to human graded essays. IEA then predicts based on the proximity in semantic space to the human graded essays what grade a human would have given to the new essay.
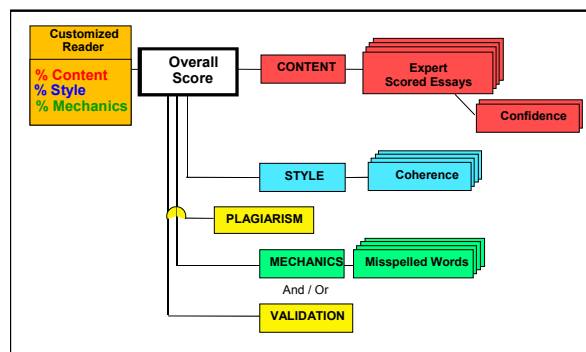
Essay Analysis System

Figure 2. IEA System Architecture.

IEA's style measures look at coherence of one sentence to the next as well as the overall coherence of the essay (i.e. how well all the sentences in the essay relate to the essay as a whole). IEA's determination of style is also based on the readability of the essay using various published readability scores. IEA's mechanics measures weight possible misspellings and compute word choice variables (level of diction). Although the original analysis of these huge amounts of text require algorithms that are compute-intensive, the output of this analysis can be used for text analysis in very short processing times less than two seconds on the average.

In the studies reported here we used the IEA technology in the Army and Air Force Command and General Staff Colleges, constituting two separate pilot tests of the technology. These two military applications differed in some important respects from other uses of automatic essay scoring technology. In the Army application, there was a strong requirement for "written feedback" rather than simply holistic or analytic scores. Thus, a central part of that study tried to mimic the feedback a skilled human grader gives to staff memo assignments. In the Air Force application, the papers to-be-graded were substantially longer (averaging 2,000 words) than the standard essay types that have graded in the past (100 to 500 words).

THE AUTOMATED SCORING STUDIES

Army CAS3 Study

A CAS3 Military Writing course was selected as the pilot. The contract was awarded by TRADOC with

Communications Technologies as the overseer and Knowledge Analysis Technologies as the subcontractor. Several hundred Captains attend CAS3 for six-week sessions. The Military Writing course teaches officers how to write staff memos. In this study, we collected between 300 and 500 memos (two to three pages) for each of four writing assignments—two informal and two formal staff memos. The two formal memo assignments were radio fielding and force modernization, and the two informal memo assignments were dental readiness and environmental compliance. Memos were graded by the small group instructors using the following criteria: substance, organization, style, and correctness. The overall grade structure was Outstanding (O) (the best), Excellent (E), Satisfactory (S), Needs Improvement (NI), and Needs Major Improvement (NMI). Each instructor had about 15 students in his or her section. Approximately, 1500 memos were collected.

The human-to-human reliability is the standard against which automatic methods are judged, requiring two humans grading the same memos. Since the only human grades were from Leavenworth CAS3, with instructors grading only a single memo, a decision was made to hire three retired former CAS3 instructors to independently grade a subset of each of the four memos. The CAS3 Operations office provided ten names of recently retired instructors; from this list we chose three. Each independent, former instructor grader was given up to 30 memos to grade for each assignment, and was paid for each memo graded. The grades they gave were compared to those of the CAS3 instructors. If the grades given by the Leavenworth graders and the independent graders were significantly different, it would indicate the graders applied different criteria. Having independent human graders was critical to establish how well IEA was working.

For the first three writing assignments, 30 memos were selected from six instructors having the requisite spread in their grades. From each qualifying section four memos were randomly selected: One memo with an Excellent grade, two with Satisfactory grades, and two with grades of Needs Improvement. Each retired CAS3 instructor was sent 20 out of the 30 memos to grade. In so doing, ten of the memos were shared in common with another retired instructor. For the fourth memo, one retired instructor graded 30 memos and the other graded 60 (30 in common and 30 new). An independent model was built for instructor R1 for this combat decision memo to determine how well IEA would do with one highly skilled, consistent grader, and this resulted in a reliability of 0.60 on the 60 memos.

IEA to Instructor Comparisons

Informal Dental Memo 1	Informal EPA Memo 1	Radio Decision Memo 2	Combat Decision Memo 2	Average
IEA-to-L .48	IEA-to-L .55	IEA-to-L .47	IEA-to-L .50	.50
R₁*-to-L .58	R₁-to-L .60	R₂-to-L .34	R₂-to-L .47	.50

NOTE: R_1, R_2, = Retired grader 1, 2

Table 1. Reliability scores between the IEA and the Leavenworth instructors.

Table 1 shows the reliability scores between the IEA and the Leavenworth instructors, and highest retired CAS3 grader and Leavenworth instructors for each of the four assignments. The average reliability across the four assignments is identical for the best-retired graders and for IEA. Also, of the retired CAS3 graders only R1 and R2 achieved good reliabilities with the Leavenworth graders. Over the three assignments that R3 graded, the reliability with the Leavenworth instructors was essentially zero. Table 1 shows the reliability for the best retired instructor-Leavenworth instructor for each of the four assignments. The major result to emerge from this study was that the automated method equaled the best performance of instructors overall—both had overall reliabilities of 0.50. The automated method was more reliable than the average retired instructor.

One of the goals of the study was to provide instructor-like feedback and comments. Since many of the errors were formatting errors, we built format checkers to look at various parts of the memo, such as distribution list, signature line, etc. If these were not correctly formatted, a message to that effect was returned. For assessing the content of various sections of the memo, we built software that parsed the student memos into sections and then compared each student's section to one or more ideal essays. The measure of goodness was the similarity in content between the student memo and the ideal. If the similarity was below a certain threshold, the student was prompted to consider rewriting that particular section. Figure 3 shows example feedback for one student memo. An overall score is returned along with separate scores for content, style and mechanics as well as for readability.
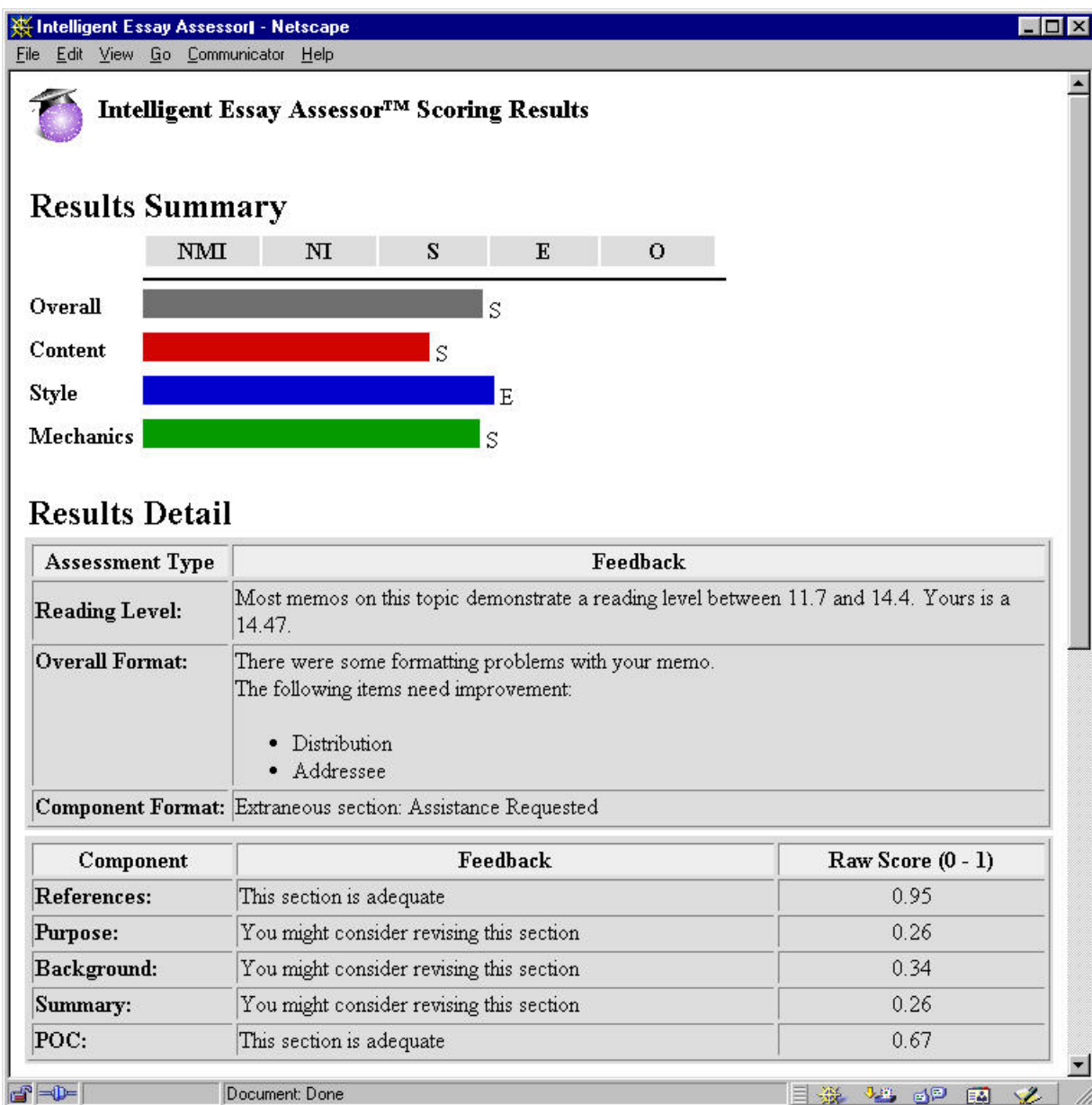
Figure 3. **Example feedback for ARMY staff memo.**

Air Command and Staff College (ACSC) Study

The motivation for the study was to test the applicability of automatic grading to distance learning as well as residential classes. As the Air Force increases distance learning to reduce travel costs, decrease time away from assignment, and decrease home-life disruptions, there will need to be effective and cost effective ways to assess students' progress. Today most assessment relies on multiple-choice exams, which requires the student to merely

recognize the correct answer. The essay is a more authentic form of assessment, since it is based on actively producing knowledge. If automatic essay assessment could be used in distance learning, it could enhance instructional efficiency and effectiveness.

In this trial 300 student take-home essay final exams from the ACSC resident International Relations Course were analyzed using the IEA technology.

The average length of these essays was 2050 words, longer than other applications of the technology.

To determine the best way to score a set of essays, the IEA computes a number of statistics on a training set of essays. These statistics are used to construct a regression model that best predicts the human grade. To this end, a training set of 106 essays was fitted by a regression model. This training model was then applied to the set of 194 test essays. Models with a modest number of parameters are preferred to avoid overfitting the training set and thereby limiting generalization to the test set. With a six parameter model, the reliability score for the training set was **0.54**; with eight parameters **0.63**. The generalization to the remaining set of essays used as a test set was **0.34** for the six parameter model and **0.43** for the eight parameter model. When the six parameter model was applied to the entire set of 300 essays, the reliability was **0.57**, in the same range as with the CAS3 results.

To obtain human-to-human reliabilities, a second reader, who is a professor at the college independently scored the 106 training essays and 103 of the 196 test essays. For the training sample, human-to-human reliability was **0.33**, for the test sample, **0.31**, both of which are relatively low. There were instances of the second grader disagreeing substantially with some of the small group instructors. Thus, there appear to be different grading criteria being applied to these longer papers by the readers.

With these inter-rater reliabilities serving as criteria, we can determine how well the IEA performed by looking at the reliabilities achieved when applying the training model to the double-scored test set of 103 essays. The IEA to instructor reliability for the test set is **0.36**; IEA to second reader is **0.35**. Thus, the IEA performed at least as well, if not slightly better, than the humans did in an equivalent test.

Because these reliabilities were lower than normally encountered, analyses were conducted to examine possible sources of variation. To this end, bibliographies were removed, international students' papers were removed, the amount of text was reduced to the first 600 words and then to first 300 words.

In this assignment, practically all students included a significant amount of quoted material in their essays. It was assumed that the students may be quoting much of the same material. Such sharing of text (and its meanings) would make it more difficult for LSA measures to reliably predict an essay's grade.
None of these manipulations altered the results previously cited in any substantial way. For instance,

removing international students decreased the predictive power, since removing them reduced the range of the distribution. The results are shown in Table 2.

Essay Types	Number Trained	Para-meter	Reliability Score
Training Essays	106	6	.54
Training Essays *(1)	106	8	.63
Training Essays *(3)	106	5	.51
Training Essays *(4)	106	8	.62
Training & Test Essays *(2)	265	8	.48
Training & Test Essays *(1,2)	265	8	.48
Training & Test Essays *(2,3)	265	5	.33
Training & Test Essays *(2,4)	265	7	.42
Training & Test Essays	300	8	.57
Training & Test Essays*(1)	300	8	.55
Training & Test Essays *(3)	300	8	.41
Training & Test Essays *(4)	300	7	.51
Training Essays *(2)	93	8	.56
Training Essays *(1,2)	93	6	.50
Training Essays *(2,3)	93	4	.45
Training Essays *(2,4)	93	3	.46

* 1. (without Bibliographies)
  2. (without Internationals)
  3. (First 300 Words)
  4. (First 600 Words)

Table 2. Overview of the grading models created from the ACSC IR602 Essays.

Since the essay grader is trained to mimic human graders it is helpful to know how well the KAT system performed in this situation. Thus, two equivalent training models were developed. The first was based on instructor scores. The second was based on the second reader scores. The set used in this analysis was the original training set of 106 essays. The inter-rater reliability of instructor score to second reader score was .31. The reliabilities of the

automated system trained on the Instructor score were .48 to Instructor scores and .37 to Second Reader scores. When the system was trained on the Second Reader the reliabilities were .45 to Instructor scores and .35 to Second Reader scores. Table 3 shows these reliabilities.

	Instructor	2nd Reader
Instructor Score	----	0.31
KAT Trained on Instructor	0.48*	0.37
Kat Trained on 2nd Reader	0.45	0.35

 * Reliability for complete test set of 194 essays was .43

Table 3. Human and Inter-rater reliabilities for Test Set (N=103)

The significance of this is that the Essay Grader, as promised, performed at least as well as human to human (instructor to second reader) graders. Although we had hoped for higher reliability coefficients we believe that the reliability coefficients obtained are both explainable and understandable.

CONCLUSIONS

The automated grading software performed as well as the better instructors in both trials, and well enough to be usefully applied to military instruction. The lower reliabilities observed in these essay sets reflect different instructors applying different criteria in grading these assignments. The statistical models that are created to do automated grading are also limited by the variability in the human grades. That is, to the extent that there is noise in the human grade distribution, the IEA model will be less robust.
There are obvious applications of automatic grading to distance learning. The grading and feedback with the automated methods is instantaneous, determined largely by network transmission time. The actual grading takes about two seconds. Thus, students could answer constructed response questions, submit a paper, and receive instant feedback, and then work on revising the quality of the written submission. The automated assessment could also be implemented in more informal environments, such as online interaction and chat discussion groups. The final result could be transmitted to the instructor for further in-depth critiquing. The automated methods could offload more of the mundane commenting, such as spelling and formatting errors. Since the automated grades are based on modeling multiple instructors or could be based on the best, most reliable instructor, there is assurance that the grading is consistent and fair.

In addition, there is room to introduce more writing assignments into the curriculum without requiring extra human grading. We have devised a method for automated scoring that requires no human grades. While it is somewhat less reliable than a human-based method, it can be used for lower stakes assessment. For instance, the Air Force would like to institute "Portfolio Writing Assignments" whereby a student would write an essay to a directed prompt and receive an assessment and some feedback and then iterate to produce a better product. Figure 4 on the next page provides a conceptual map of this application.

Effective written and oral communications are critical to military mission success. Perhaps the one message to take from this work is that automated grading allows students to improve their written communication skills with realistic assignments without instructor burden.
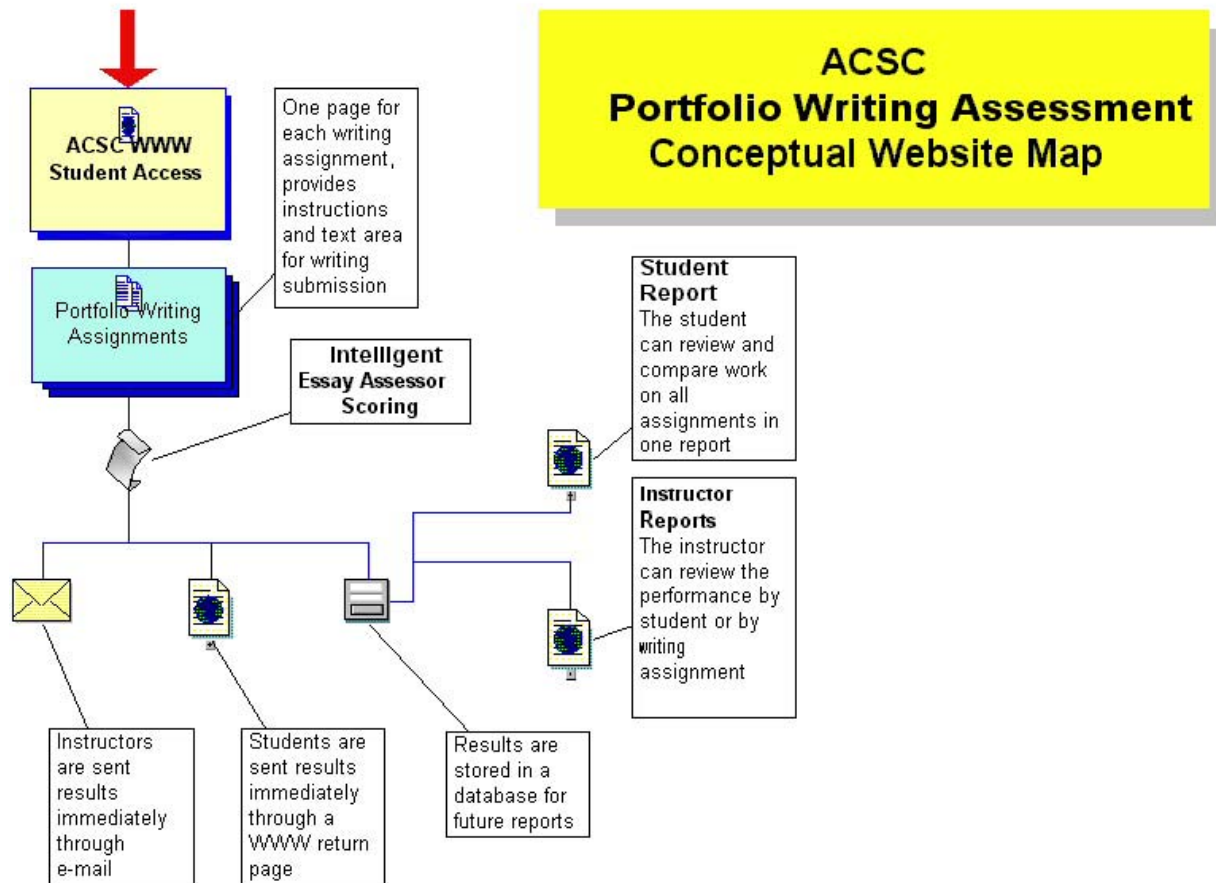
Figure 4. Portfolio Writing Assessment Application

REFERENCES

Birenbaum, M., & Tatsuoka, K. K. (1987). Open-ended versus multiple-choice response formats--It does make a difference for diagnostic purposes. Applied Psychological Measurement, 11, 385-395.

Braun, H. (1988). Understanding scoring reliability: Experiments in calibrating essay readers. Journal of Educational Statistics, Spring, Vol. 13, No. 1. pp. 1-18.

Chi, M. T. (2000). Self-explaining expository texts: the dual processes of generating inferences and repairing mental models. In: R. Glaser (Ed.), Advances in Instructional Psychology (161-237). Mahwah, NJ: Erlbaum.

Greenwald, E. A., Persky, H. R., Campbell, J. R., Mazzeo, J. (1999). The NAEP Writing Report Card for the Nation and the States, National Center for Education Statistics, 1999-462, Washington, DC, http://nces.ed.gov./nationsreportcard/pdf/main1998

Kintsch, E., Steinhart, D., Stahl, G., Matthews, C., Lamb, R., and the LSA Research Group. (2000). Developing summarization skills through the use of LSA-based feedback. *Interactive Learning Environments,* 8(2), 87-109.

Landauer, T. K., Laham, D., & Foltz, P. W. (2001, in press). Automatic essay assessment with Latent Semantic Analysis. In B. Clauser (Ed.) Special issue on computerized scoring of complex item types. Applied Measurement in Education.

Palincsar, A. S., Brown, A. L., & Campione, J. C. (1994). Models and practices of dynamic assessment. In: G. P. Wallach, & K. G. Butler (Eds.), Language learning disabilities in school-aged children and adolescents (132-144). Boston: Allyn & Bacon.

Steinhart, D. (2000) Summary Street: An LSA-Based Intelligent Tutoring System for Writing and Revising Summaries. Unpublished Doctoral Dissertation, University of Colorado.

Warfield, T. D., Johnstone, K. M., and Ashbaugh, H. Effects of repeated practice and contextual-writing experiences on college students' writing skills. *Journal of Educational Psychology*, Vol 94(2), Jun 2002. pp. 305-315.