Running Head: AUTOMATED ESSAY SCORING

Experimental Evidence on the Effectiveness of Automated Essay Scoring in Teacher Education Cases

Eric Riedel
University of Minnesota

Sara L. Dexter
University of Virginia

Cassandra Scharber
University of Minnesota

Aaron Doering
University of Minnesota

Paper prepared for the 86th Annual Meeting of the American Educational Research Association, April 11-15, 2005, Montreal, CA.

Questions or comments about this study can be directed to the first author at the Center for Applied Research and Educational Improvement, University of Minnesota, 275 Peik Hall, 159 Pillsbury Avenue SE, Minneapolis, MN 55455-0208, 612-624-3805, riedel@umn.edu.

Abstract:

This study tests the impact of an automated essay scorer (AES) which provides formative feedback on essay drafts written as part of a series of online teacher education case studies. A total of 70 preservice teachers in four teacher education courses were assigned to complete two cases. Each student was randomly assigned to either a condition where the scorer was available (experimental condition) or a condition where the scorer was unavailable (control condition). Those students in the experimental condition submitted higher-quality final essays and conducted more efficient searches of the case than students in the control condition. Essay scores were positively associated with the number of drafts submitted to the scorer for formative feedback.

Experimental Evidence on the Effectiveness of Automated Essay Scoring in Teacher Education Cases

This study examines the effect of students having available and using an automated essay scorer as a formative feedback tool on their short essay responses to teacher education cases that asked them to detail an instructional decision about technology integration in a classroom. The implication of a reliable automated essay scoring tool in use within a network-based online learning environment is that it makes possible formative feedback, and presumably improved performances, on complex sets of skills. Here, we present findings from a controlled experiment on how preservice teachers' use of an automated essay scoring tool within the ETIPS learning environment impacted the quality of their essay responses as measured by human scorers against the established assignment rubric.

Literature Review

Computer-Based Writing Evaluation

In the field of education, computer-based writing evaluation system development has been going on since the mid 1960s when Ellis Page and others developed the first generation of Project Essay Grader (Yang, Buckendahl, Juszkiewicz, & Bhola, 2002; Page, 1966, 1994). Since then, nearly a dozen technology-based systems and approaches have been developed to measure writing quality.

Valenti, Neri, and Cucchiarelli (2003), identified ten computer-based systems used for evaluating "free text answers" or essay writing and grouped the systems based upon how their methodology for evaluating essays. Three evaluate essays primarily for content: Intelligent Essay Assessor (IEA), Educational Testing Service I (ETS I), Concept Rater (C-Rater). Project Essay Grade (PEG) evaluates essays primarily for style. The remainder evaluate both content and style: Electronic Essay Rater (E-Rater), BETSY, Intelligent Essay Marketing System, SEAR, Paperless School free text Marking Engine (PS-ME), and Automark.

Much of the research on these automated essay scorers is focused on their accuracy, testing how similarly the system scores an essay as compared to a human scorer. The e-Rater system, developed by the Educational Testing Service (ETS), currently leads the field in terms of effective and accurate automated essay scoring (Burstein, Kukich, Wolff, Chi, & Chodorow, 1998; Burstein, Leacock, & Swartz, 2001). Since 1999, E-Rater has been used as one of two scorers (the other scorer is a human) for the essay section of the General Management Aptitude Test (GMAT). The agreement rates between humans and this scoring system on the over 750,000 GMAT essays is 97%. However, there is increasing interest and recognition in how computer-based systems could be used to support formative assessment of students' work (Charman & Elmes, 1998; Sambell, Sambell, & Sexton,1999; Shermis & Burnstein, 2003), and in particular using automated essay scoring software to help students improve their writing (Myers, 2003).

Response to Student Writing

Whether revision takes place and, if it does, fostering substantial revisions in student writing has long been a subject of study. Research on the revision process in the 1970s and 1980s indicated that students' frequent focus on surface matters resulted in little substantive revision (Beach, 1976; Emig, 1971; Sommers, 1982). Furthermore, there is little evidence that several instructional models lead to substantive revision (National Writing Project, 2003).

The nature of the teacher's feedback during the composition process is also critical to whether students revise (Yagelski, 1995). For example, a study of the influence of 11 teachers' feedback on 64 middle-school students revision of drafts illustrates the impact of teacher influence on revision (Matsumura, Patthey-Chavez, & Valdes, 2002; Patthey-Chavez, Matsumura, & Valdes, 2004). Most of the teachers' feedback focused on editing and the students' revisions reflected this focus: 58% of students' revisions involved surface-level changes and 34% involved content-level changes. Furthermore, the content-based revisions did not improve the quality of most drafts – the changes simply added information to the writing.

A meta-analysis of all of the data-based instructional research on writing (Hillocks, 1996) indicates that students focus primarily on form over development and expression of content over the use of writing to discover and express ideas (Beach & Friedrich, 2005). Students assume that they first need to define their organization of content before/prior to writing as opposed to using writing to discover content/ideas. They often focused their attention on creating a single draft aligned with the format, resulting in a draft that was already relatively completed, so that they were only focusing on matters of editing. The primary focus is often simply on editing—correcting errors versus also developing ideas. There is a lack of prewriting, revision, rethinking, or focus on the process of writing.

In discussing teacher response to writing, it is also important to consider whether students will understand a teacher's response and then actually make use of that response in revising their writing based on the demands of a rhetorical context. This presupposes that both the student and the teacher share a common understanding of the teacher's expectations for the writing assignment, the criteria for effectively completing that assignment, and the purpose for their responses to the writing (Kim, 2004; Wallace & Hayes, 1992). One primary issue is the fact that regardless of the nature and quality of feedback, students will simply comply with what they perceive their teacher wants them to do in order to obtain a good grade, even though their suggestions may not help improve their writing (Sperling & Freedman, 1987; Straub, 1996).

According to the reviewed studies, students seem to agree about two types of comments they find helpful. First, they favor comments that suggest ways of making improvements (Reed & Burton, 1985; Ferris, 2003). Second, students favor comments that explain why something is good or bad about their writing (Land & Evans, 1987). Students' ability to use these comments to make revisions depends on the rhetorical strategy they were addressing. Students are more successful in making changes in response to comments having to do with providing information, requests for specific changes, or comments on grammar and mechanics, than with questions dealing with challenges to the students' ideas or argument (Conrad & Goldstein, 1999; Ferris, 2001). All of this suggests that written comments may be particularly effective in fostering certain kinds of revisions—adding details/examples, improving coherence, or dealing with editing matters because the written comments can focus on specific aspects of a draft (Conrad & Goldstein, 1999).This research points to the need for teachers to both provide content-based feedback and to show or direct students on how to use that feedback to improve the quality of their drafts based on their ability to address the rhetorical aspects of writing within specific social contexts.

In summary, the research suggests that in order for an automated essay scoring system to provide formative feedback that supports students in improving the quality of their writing, its feedback must provide recommendations for improvement and explain why the essay needs improvement. In the next section we give background information on the automated essay scoring (AES) tool itself, research on students' writing revisions from an earlier version (v.1) of

the AES, and the nature of the feedback in the current version (v. 2) of the AES, which was revised with guidance from findings from the earlier version well as the literature on writing response reviewed here. The findings section then reports what we learned about the effectiveness of the current version (v. 2) of the AES from this experiment.

Background

ETIPS cases provide a simulated school environment for preservice teachers to practice making decisions regarding technology integration in education (see http://www.etips.info to view sample cases). Each case poses a case challenge to make a decision about how to integrate technology as a beginning teacher in a particular setting. The challenges vary by the setting and which of six educational technology integration and implementation principles (eTIPs) they emphasize (Dexter, 2002). Users are asked to plan their search using an embedded tool called the PlanMap which they can reference later to reflect on their search. Users then search among 68 pieces of information in seven categories (e.g. students, staff, technology infrastructure) for information relevant to the question. After their search, users respond to the challenge by writing a short three-part essay answering the case question. An automated essay scorer provides predicted scores on drafts of the case essay and suggests additional information (based in part on student's prior search) that might help the student in improving his/her essay.

ETIPS Automated Essay Scoring Tool

The ETIPS software offers another automated essay scorer (AES) to the field of computer-based writing assessment. The software was developed in 2003 as a component of online cases for teacher education. Unlike in other automated essay scoring systems, the feedback given by the automated scorer in the ETIPS cases is designed to be used in a formative fashion. That is, students have the option to obtain predicted scores on their essay responses before they submit final responses to their instructor for grading.

The ETIPS automated essay scorer uses a Bayesian model to score essays both for content and style. The current AES examines various features of essay responses including vocabulary, word usage, specific phrases, and grammatical structure. It then compares these features in students' essays to those same features in training essays that have already been scored by human experts. By examining the correlations between students' essays and the trained essays, the AES can predict how likely students are to receive a score of 0, 1, or 2 against the rubric (see Appendix A).

On the webpage where students submit their answer, students can click different buttons and submit their responses either for automated feedback or to their instructor as their final answer. If they chose to get automated feedback, afterwards they can chose to go back into the case context to search for additional information, re-draft their responses, or submit their responses as final answers to their instructors. There is no limit as to the number of times students can submit responses to the ETIPS automated essay scorer for feedback. Additionally, when instructors score student essays, they are able to view the number of drafts a student submitted for automated feedback as well as the estimated score the scorer gave them; instructors are not able to view the actual drafts.

In the earlier of the AES, the feedback from the software was a bar graph showing the percent likelihood that they would receive the score of 0, 1, or 2 on the essay and a short

explanation of a "good answer" (see Figure 1). For an explanation for their score, students could view the rubric online.

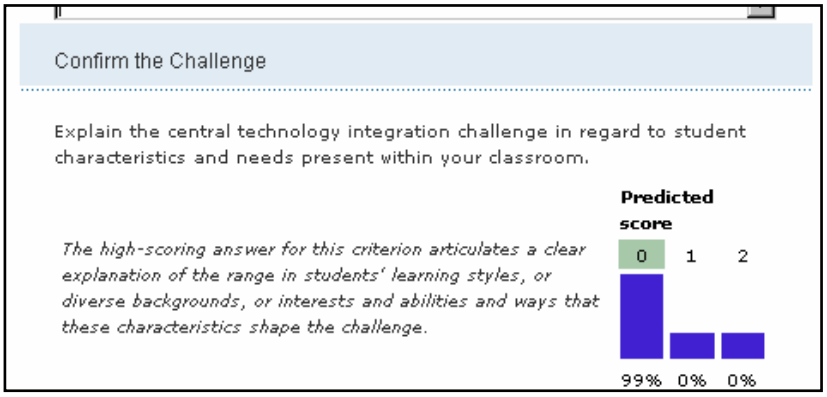

Figure 1. Example of feedback generated by the first version of the ETIPS automated essay scorer, including the bar graph which illustrated the percent likelihood that the student would receive a score of 0, 1, 2 from his/her instructor.

In a study of students' reaction to and use of version one of the ETIPS AES (Scharber & Dexter, 2004), students did not hesitate to use and experiment with the scorer as a means of improving their work---even though they were not required to use it. The changes students made to their essays were positive ones in terms of writing structure: they provided additional examples or details, or referenced either the key question from the case introduction or the technology integration principle about which the case was designed to provide practice thinking about. While extensive use of the ETIPS scorer was associated with improved essays as measured by their final human score, some students became frustrated when they revised their work without the scorer's responding with a prediction of an improved score. Some students gave up on using the scorer as a formative feedback tool. In reviewing the essay drafts of students who did submit multiple version of their responses to the AES, using the assignment rubric to score and compare first drafts to successive drafts we found that their scores did not improve, even though in the majority of cases there was room for them to do so (i.e. they were not yet at the highest level on the rubric, a score of 2).

We concluded that the nature of the feedback given to students needs to be more detailed. A simple prediction of a 0, 1, or 2 score evidently does not, in combination with the rubric to which these cores refer, provide enough guidance to students as to how to improve their essays.

In keeping with literature on writing response we sought to summarize what characterized an essay receiving a score of 0, 1, or 2, as well as make specific suggestions for improvement. In the current version (v. 2) of the AES the score prediction data is combined with the ability of the software to track which case information students accessed (see Figure 2). Based on information collected about what information they did not search as well as their predicted score, the AES recommends that the user access specific information items that may improve their predicted score which can in turn be used to improve their essay.

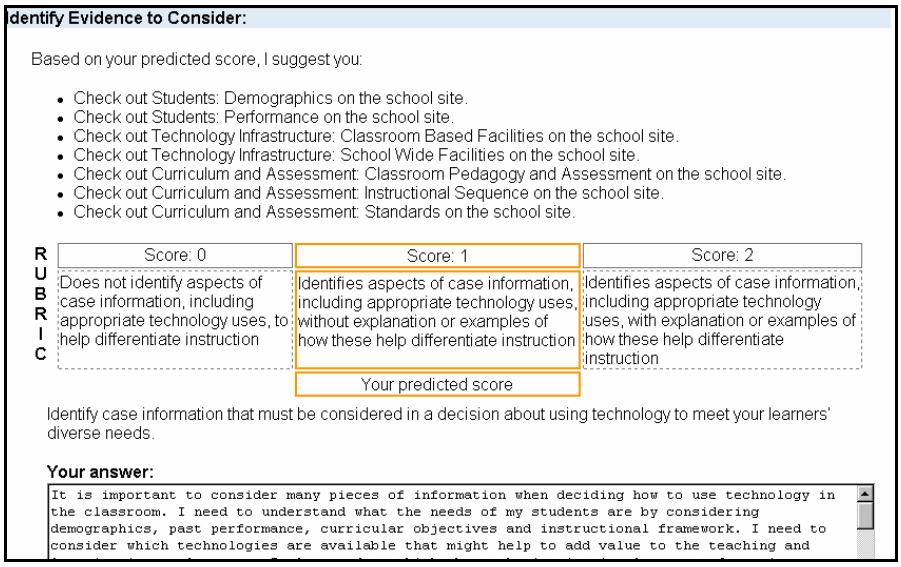

Figure 2. Example of feedback generated by the second version of the ETIPS automated essay scorer.

Findings

Data and Methods

Prior to the fall 2004 semester, teacher education faculty who previously had participated in testing the ETIP cases were invited to participate in experiments involving the automated assessment features of the cases. Three faculty, teaching four courses, agreed to participate. The first taught two sections of an introductory educational technology course required of all preservice teachers enrolled in a post-baccalaureate licensure program at a major research university. One section was for students in the English education cohort while the second section contained Science education students. The second instructor also taught an introductory educational technology course required of all preservice teachers. The licensure program included undergraduates, however, and the course was not specific to a particular licensure program. The third instructor taught a math education methods course. Each faculty member agreed to assign students to complete at two ETIP cases addressing eTIP 2 (added value) outside of class after an introduction to the cases in a class session. The cases were limited to those case scenarios involving middle or high schools. A single exception occurred in that the math methods courses assigned one of the cases to involve an elementary school. Note that technical difficulties with the first case in the English educational technology course led to the data being unreliable. These difficulties were not apparent to students and did not affect their work in the first case.

As students registered at the ETIP website as assigned they were automatically directed to an online consent form. Those who consented to participate in the experiment then were given a short, online, pre-case survey. Each student was then randomly assigned by the software to one of two conditions: cases having the AES available (experimental condition) or cases not having the AES available (control condition). All other aspects of the cases, including having the PlanMap available, were the same for both conditions. The conditions remained the same for

each student across the two cases. Following completion of the two cases, students were automatically forwarded to a short, online, post-case survey. After the fall 2004 semester, all essays were blindly scored by a member of the research team using a three-score rubric (see Appendix A).

Results

Figures 3 and 4 illustrate the range of scores for each experimental condition for each class on the first and second cases respectively. The three individual scores are added together for each case to form a summary measure with a range from 0 to 6. The experimental condition includes those students who had access to the AES while the control condition includes those students who did not. With a single exception of the scores for the first case in the introductory educational technology course, those students in the experimental condition on average scored higher than those in the control condition.

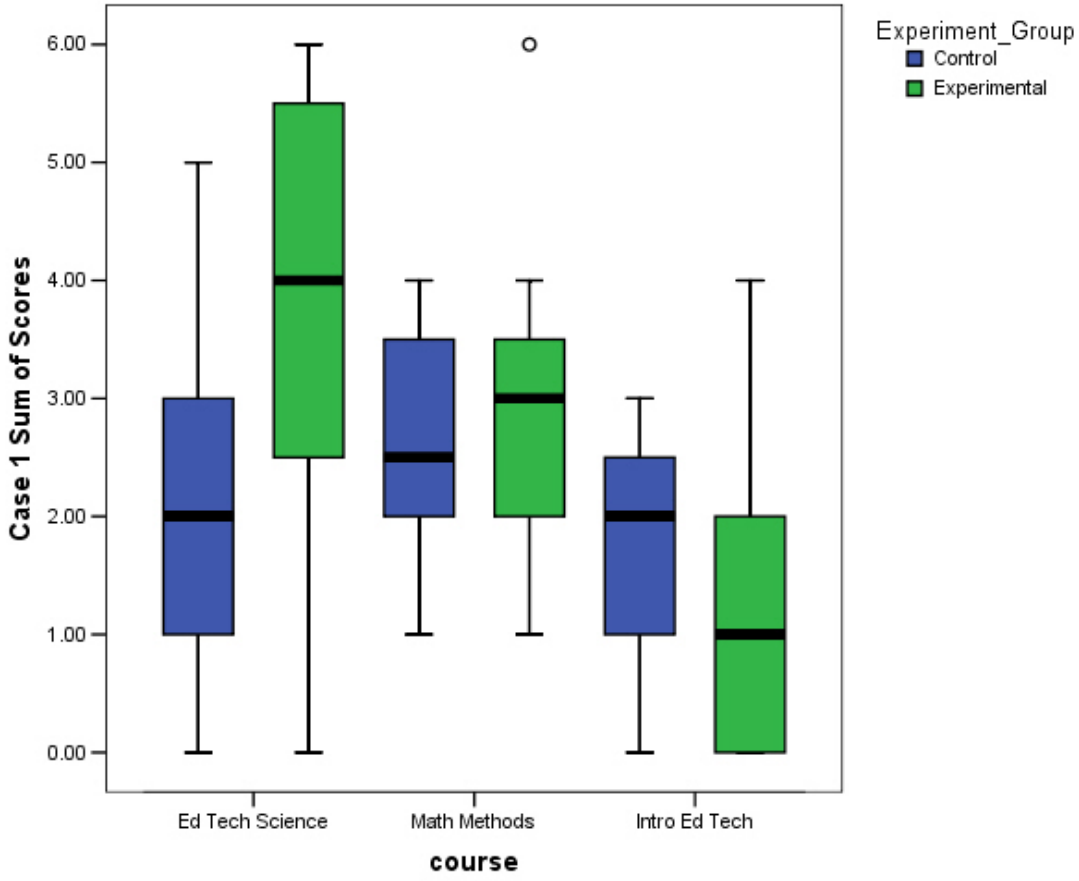

Figure 3. Box plot of first case sum of essay scores by class and experimental condition. Solid horizontal lines indicate the median score. High and low boundaries of box indicate 75th and 25th percentiles respectively. High and low lines outside of box indicate maximum and minimum range of continuous scores.

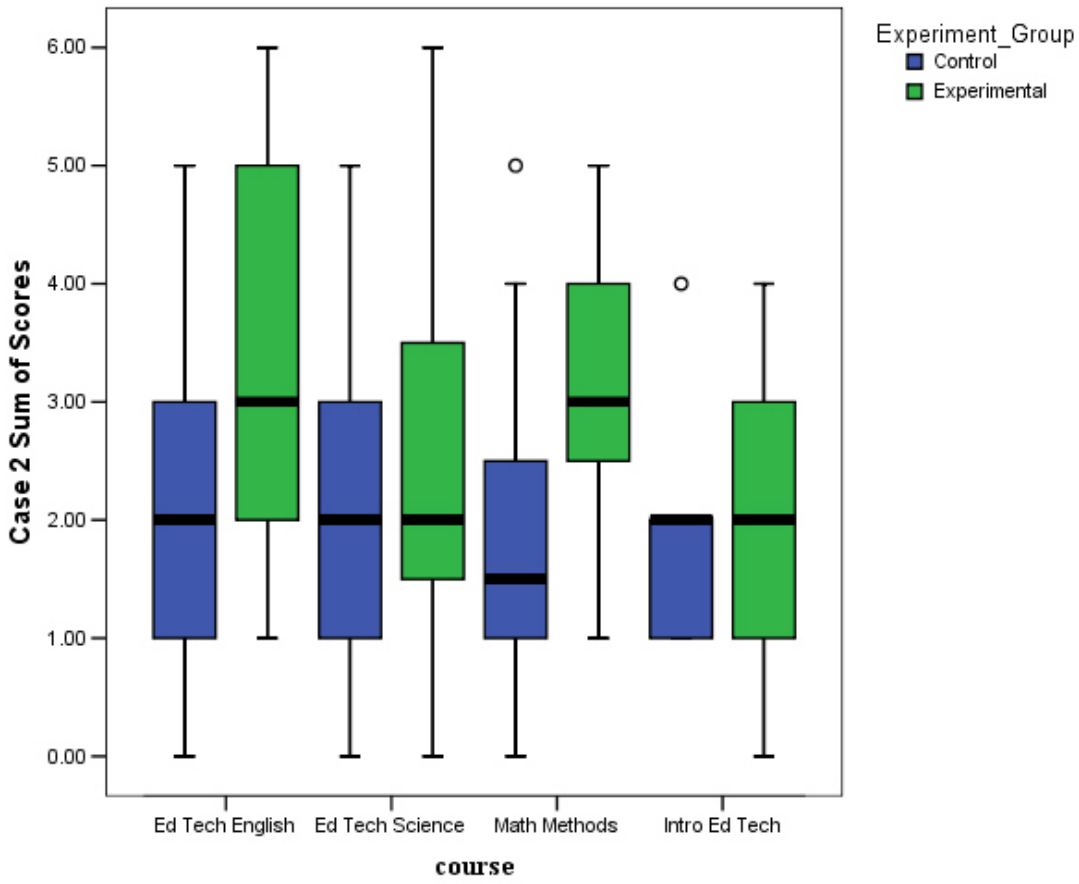

Figure 4. Box plot of second case sum of essay scores by class and experimental condition. Solid horizontal lines indicate the median score. High and low boundaries of box indicate 75th and 25th percentiles respectively. High and low lines outside of box indicate maximum and minimum range of continuous scores.

These apparent differences can be tested for statistical significance. Three outliers (one in an experimental condition, two in control conditions) are removed to meet statistical assumptions of a univariate analysis of variance. Two ANOVAs were run using the essay sum of scores for each case as the dependent variables. Experimental group and class are included as the fixed factors. The results are presented below in Tables 1 and 2. For the first case, there are statistically significant differences between courses in the essay scores but no effect for experimental group. In the second cases, there is a statistically significant main effect for experimental group but not for course. In neither analysis is there an interaction effect between course and experimental group. If the outliers are included in the analysis, the main effect for experimental group has only marginal statistical significance in the second case ($p=.075$). A Mann-Whitney test, a non-parametric test which is less affected by outliers, does show a statistically significant effect for experimental condition ($z=-1.9868$, $p < .05$) when including outliers however. Thus the decision to exclude the outliers in order to use the more powerful analysis of variance is justified.

Table 1.
Analysis of Variance of Sum of Essay Scores in First Case

Source	Sum of Squares	df	Mean Square	F	p-value
Course	18.522	2	9.261	4.919	.012
Experimental Group	3.150	1	3.150	1.673	.203
Course x Experimental Group	5.794	2	2.897	1.539	.226
Error	82.842	44	1.883		
Total	113.520	49			

Table 2.
Analysis of Variance of Sum of Essay Scores in Second Case

	Sum of Squares	df	Mean Square	F	p-value
Course	10.274	3	3.425	1.584	.202
Experimental Group	11.672	1	11.672	5.399	.023
Course x Experimental Group	5.251	3	1.750	.810	.493
Error	134.045	62	2.162		
Total	164.343	69			

The experimental condition is actually the availability of the AES as a formative assessment device – not its actual use. We do, however, infer that it is the use of this feature that is the cause of the differences between the experimental and control groups. Due to the nature of random assignment, no other systematic differences should exist. There remains the possibility that students in the experimental condition perform better simply because they are in an experiment examining their performance in relation to an AES, a kind of Hawthorne effect. To address this possibility, we look at the relationship between the essay scores and the number of drafts submitted to the scorer. If the AES is to prove helpful in improving the quality of essays, there should be some positive relationship between scorer use and essay quality. Figure 5 below is a scatter plot of number of drafts submitted to the AES for the first case and the sum of essay scores for the first case. Figure 6 is the same scatter plot except for the second case. While the relationship is difficult to discern in the first case it is positive in the second case. Those who submit five or more drafts tend not to score below a 2.0 in the summary score.

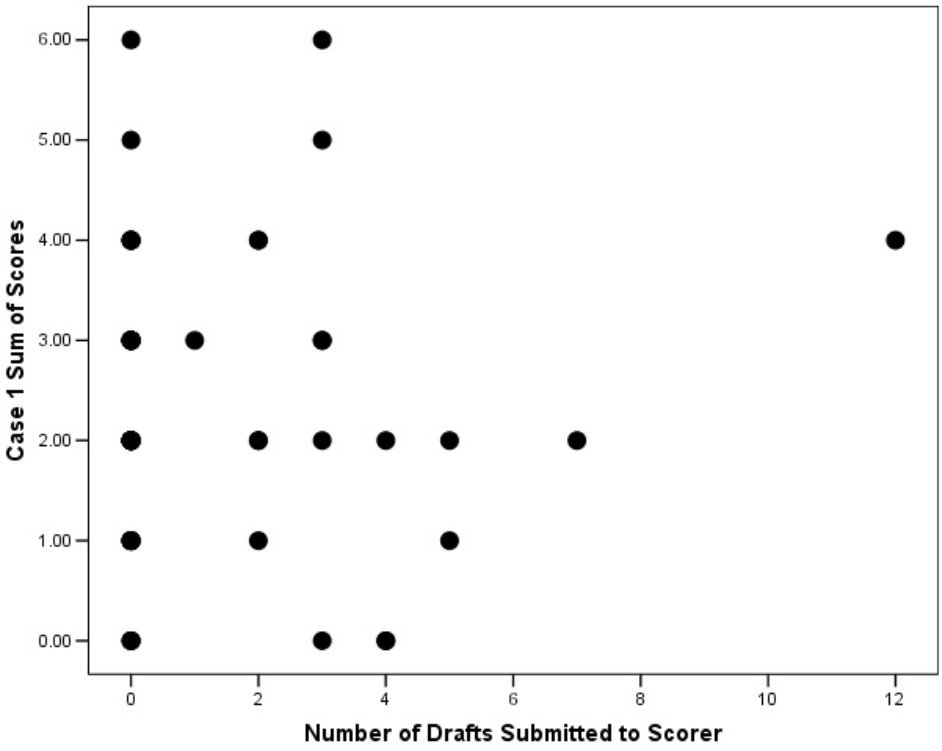

Figure 5. Scatter Plot of Drafts Submitted to AES by Essay Scores (First Case)

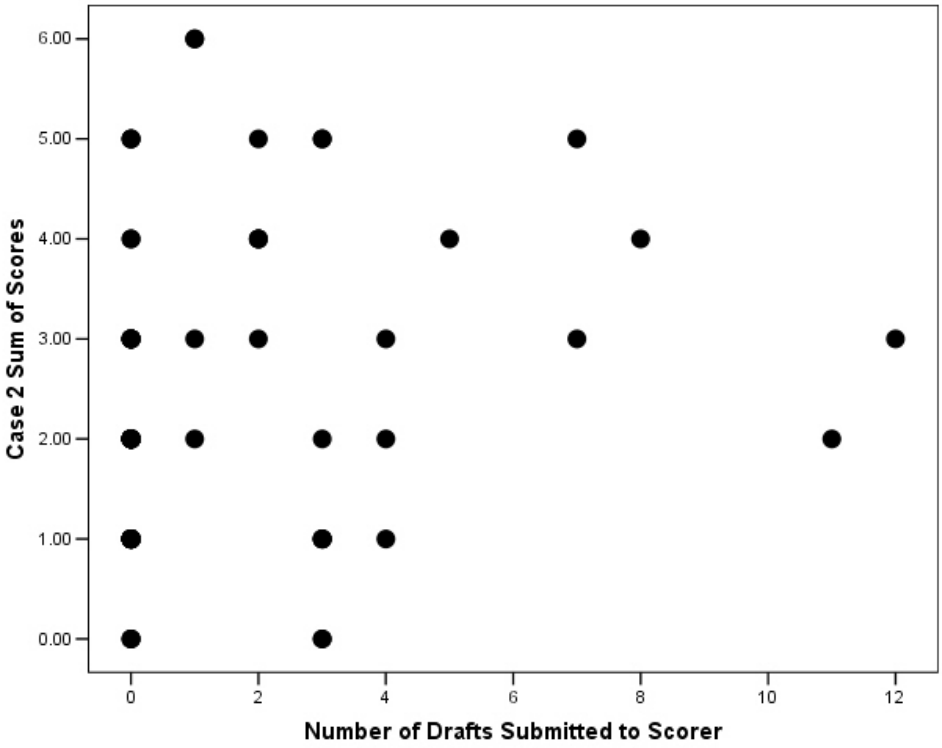

Figure 6. Scatter Plot of Drafts Submitted to AES by Essay Scores (Second Case)

Since the AES can play an important role in the search of the case through suggesting the user access additional information relevant to the question, it is also worthwhile to investigate whether the characteristics of the users' case searches differed by experimental group. Two outcome variables are examined in Table 3 below in addition to the sum of essay scores for each case. The mean number of steps is the average number of times the user seeks to access a separate piece of information (including returns to the same information) during a case. The mean number of relevant items is the average number of individual items judged relevant to answering the case question by designers (not counting returns to the same item). These mean scores on each of these are provided in Table 3 below. There appear to be only slight differences between the two conditions for each of the additional outcomes. Students in the control condition appear to make a marginally more extensive search, as indicated by the number of steps taken in the case. Students in the control condition appear to have equally or marginally less relevant search as indicated by the number of relevant items accessed.

Table 3
Comparison of Experimental and Control Groups on Outcome Measures

			Outcome Measure					
Case	Number of Users		***Sum of Scores for Case (Range=0-6)***		***Mean Number of Steps Taken in Case***		***Mean Number of Relevant Items Accessed***	
	Control	Experim.	Control	Experim.	Control	Experim.	Control	Experim.
1	23	27	2.13	2.56	28	25	6	6
2	34	36	1.91	2.81	24	23	6	7

Tables 4 through 7 show the results of analyses of variance testing whether these differences are statistically significant. Similar to the method used to test for differences in essay scores between the two conditions, an analysis of variance is used to predict the number of steps and number of relevant items for each case. Each of the tables shows the main effects for class and experimental group as well as whether there is an interaction effect between the two. There is no main or interaction effect for either class or experimental condition for the first case but differences do appear for the second case. In the second case, there are main effects for experimental condition on the number of steps and number of relevant items accessed. There is also a main effect for class on the number of relevant items accessed. There are no interaction effects however.

Table 4
ANOVA for Total Number of Steps Taken in First Case

Source	Sum of Squares	df	Mean Square	F	p-value
Course	282.576	1	282.576	.692	.408
Experimental Group	3235.587	3	1078.529	2.640	.055
Course x Experimental Group	187.055	3	62.352	.153	.928
Error	31052.015	76	408.579		
Total	34757.560	83			

Table 5
ANOVA for Total Number of Steps Taken in Second Case

Source	Sum of Squares	df	Mean Square	F	p-value
Course	35.438	1	35.438	.131	.718
Experimental Group	2834.122	3	944.707	3.501	.020
Course x Experimental Group	159.305	3	53.102	.197	.898
Error	18347.179	68	269.811		
Total	21378.039	75			

Table 6
ANOVA for Number of Relevant Items Accessed in First Case

Source	Sum of Squares	df	Mean Square	F	p-value
Course	11.547	1	11.547	3.111	.082
Experimental Group	9.875	3	3.292	.887	.452
Course x Experimental Group	15.132	3	5.044	1.359	.262
Error	282.085	76	3.712		
Total	315.952	83			

Table 7
ANOVA for Number of Relevant Items Accessed in Second Case

Source	Sum of Squares	df	Mean Square	F	p-value
Course	18.635	1	18.635	4.638	.035
Experimental Group	39.852	3	13.284	3.306	.025
Course x Experimental Group	2.233	3	.744	.185	.906
Error	273.241	68	4.018		
Total	330.947	75			

Discussion

The presence of the automatic essay scorer appears to have a moderate but robust impact on the quality of the case search and essay. The difference in outcomes between students who did not have access to the AES and those who did is more pronounced in the second case. In that case, students who had access to the AES had higher quality essays, a shorter search of the case, and a search more similar to that of an expert. The differences were not large but were statistically significant even controlling for differences among classes. The inference that it was the actual use of the AES, rather than just its presence, seems to be correct insofar as the level of

use of the scorer was moderately predictive of increases in essay scores. In general, the scorer appears to function as intended.

Earlier research on the first version on the ETIPS automated essay scorer revealed that the design, nature, accuracy, and specificity of the feedback provided by the AES were all very important in aiding students' better performances on and positive experiences with it (Scharber & Dexter, 2004). The second version of the AES, which was tested in this study, changed the design of the feedback, with the rubric that highlighted their predicted score now appearing where there had been three bar graphs depicting the likelihood of receiving one of the possible scores. We infer that this helped the students direct their attention to the content of their essays.

Other researchers have found that providing revision or other self-regulatory strategies promotes students' writing achievement (Danoff, Harris, & Graham, 1993; Fleming & Alexander, 2001; Graham & Harris, 1989; Stoddard & MacArthur, 1993), and that such strategies can be paired with rubrics for positive benefit (Duke & Miller, 2004). We infer that the changes to the nature and specificity of the feedback in the second version of the AES were a first step toward a cognitive strategy that promoted self-regulatory behavior in writing revisions. That is, by analyzing how many of the expert recommended pages students visited and then directing them to pages they had not yet visited we directed their attention to the most pertinent information on which they should base their answer. Further research on this aspect of the AES is needed so we can better design more specific iterations of cognitive strategy interventions as students submit subsequent drafts. For example, the AES could suggest specific topics they address in their answer, after its content analysis, and relate this to where they might find the information to draw upon for that discussion.

Finally, further development work and research is needed regarding the accuracy of the AES. Students quickly become frustrated or lose confidence in the scorer if they feel that it isn't responding at all or very accurately (Scharber & Dexter, 2004). However, we do not know how much a more accurate scorer will improve the quality of their writing.

It is recognized that automated essay scoring systems offer much promise as formative assessment tools, particularly in online or networked learning situations, such as the context for this research. These study results are a positive indication that these hopes are warranted and point out some important features to which developers must attend as well as some further research that is needed.

References

Beach, R. (1976). Self-evaluation strategies of extensive revisers and non-revisers. *College and Communication, 27,* 160-164.

Beach, R., & Friedrich, T. (2005). Response to Writing. IN C.A. MacArthur, S. Graham, & J. Fitzgerald. (Eds.). Handbook of Writing research. Boston: Allyn And Bacon.

Danoff, B., Harris, K. R., & Graham, S. (1993). Incorporating strategy instruction within the writing process in the regular classroom: Effects on the writing of students with and without learning disabilities. *Journal of Reading Behavior, 25*(3), 295-322.

Dexter, S. (2002). eTIPS-Educational technology integration and implementation principles. In P. Rodgers (Ed.), *Designing instruction for technology-enhanced learning* (pp. 56-70). New York: Idea Group Publishing.

Duke, B.L., & Miller, R. (2004) *The influence of using cognitive strategy instruction through writing rubrics on high school students' writing motivation and achievement.* Paper Presentation at the Annual Meeting of the American Educational Research Association, San Diego, CA, April 2004.

Emig, J. (1971). The composing processes of twelfth graders. Urbana, Illinois: National Council of Teachers of English.

Fleming, V. M., & Alexander, J. M. (2001). The benefits of peer collaboration: A replication with a delayed posttest. *Contemporary Educational Psychology, 26,* 588-601.

Graham, S., & Harris, K. R. (1989). Components analysis of cognitive strategy instruction: Effects on learning disabled students' compositions and self-efficacy. *Journal of Educational Psychology, 81*(3), 353-361.

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

Myers, M. (2003). What can computers and AES contribute to a K-12 writing program? In M.D. Shermis & J.C. Burstein (Eds) Automated essay scoring: A cross-disciplinary Perspective. Mahwah, NJ: Lawrence Erlbaum Associates.

National Writing Project. (2003). *Because writing matters: Improving student writing in our schools.* San Francisco, California: Jossey-Bass.

Page, E. B. (1966). The imminence of grading essays by computer. *Phi Delta Kappan, 48,* 238-243.

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

Patthey-Chavez, G. B., Matsumura, L., &Valdez, R. (2004). Investigating the process approach to writing instruction in urban middle schools. *Journal of Adolescent and Adult Literacy, 47*(6), 462-479.

Peat, M. & Franklin, S. (2002). Supporting student learning: The use of computer-based formative assessment modules. *British Journal of Educational Technology, 33*(5), 515-523.

Sambell, K., Sambell, A., & Sexton, G. (1999). Student perceptions of the learning benefits of computer-assisted assessment: A case study in electronic engineering. In S. Brown, P. Race, & J. Bull, *Computer-assisted assessment in higher education* (pp. 179-191). London: Kogan Page Limited.

Scharber, C. & Dexter, S. (2004, March). Automated essay score predictions as a formative assessment tool. Paper presented at the fifteenth international conference of the Society for Information Technology and Teacher Education, Atlanta, GA.

Shermis, M.D. & Burstein, J.C. (2003). Automated essay scoring: A cross-disciplinary Perspective. Mahwah, NJ: Lawrence Erlbaum Associates.

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

Stoddard, B., & MacArthur, C. A. (1993). A peer editor strategy: Guiding learning disabled students in response and revision. *Research in the Teaching of English, 27*(1), 76-102.

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

Yagelski, R. (1995). The role of classroom context in the revision strategies of student writers. *Research in the Teaching of English*, 29, 216-338.

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

APPENDIX A

Rubric for eTIP 2

eTIP 2: Technology provides added value to teaching & learning

Consider how technology can add value to your ability to meet the diverse needs of your learners in the context of both your curriculum and the school's overall improvement efforts.

Criteria	0	1	2
Confirm the Challenge: Explain the central technology integration challenge in regard to student characteristics and needs present within your classroom.	Does not present an explanation of the range in students' learning styles, or diverse backgrounds, or interests and abilities and ways that these characteristics shape the challenge	Presents a limited explanation of the range in students' learning styles, or diverse backgrounds, or interests and abilities and ways that these characteristics shape the challenge	Articulates a clear explanation of the range in students' learning styles, or diverse backgrounds, or interests and abilities and ways that these characteristics shape the challenge
Identify Evidence to Consider: Identify case information that must be considered in a decision about using technology to meet your learners' diverse needs.	Does not identify aspects of case information, including appropriate technology uses, to help differentiate instruction	Identifies aspects of case information, including appropriate technology uses, without explanation or examples of how these help differentiate instruction	Identifies aspects of case information, including appropriate technology uses, with explanation or examples of how these help differentiate instruction
State Your Justified Recommendation: State a justified recommendation for implementing a viable classroom option to address the challenge.	Does not state a recommendation for using, or not using, a particular technology to differentiate instruction to meet the diverse needs of learners	Presents a limited recommendation for using, or not using, a particular technology to differentiate instruction to meet the diverse needs of learners	Presents a well-justified recommendation for using, or not using, a particular technology to differentiate instruction to meet the diverse needs of learners