Running Head: FORMATIVE FEEDBACK VIA AN AUTOMATED ESSAY SCORER

Formative Feedback Via An Automated Essay Scorer: Its Impact On Learners

Cassie Scharber
Center for Applied Research & Educational Improvement
University of Minnesota
United States
scha0809@umn.edu

Sara Dexter, Ed.D.
University of Virginia
United States
sdexter@virginia.edu

Eric Riedel, Ph.D.
Center for Applied Research & Educational Improvement
University of Minnesota
United States
riedel@umn.edu

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

Abstract

The purpose of this research is to analyze preservice teachers' use of and reactions to an automated essay scorer as a means of formative feedback in an online, case-based learning environment. Data analyzed include pre- and post-semester surveys, a user log of students' actions within the cases, instructor-assigned scores on final essays, and in-depth interviews with four selected students. Analysis illustrates that extensive use of the scorer was associated with improved essays as measured by their final human-assigned scores. Study findings also suggest factors that must be considered in using technology to provide formative feedback including the design, nature, accuracy, and specificity of feedback, all of which were found to impact student performances and experiences. These outcomes boost confidence in the potential utility of instructional technology to aid in formative assessment within the ETIPS cases learning environment.

Formative Feedback Via An Automated Essay Scorer: Its Impact On Learners

Objectives

Technological advances promise to contribute to the improved assessment of student learning. Although technology has been primarily been used for summative assessment purposes, it also holds much promise in aiding formative assessment. As an example, the ETIPS cases (http://www.etips.info), which are online, case-based learning environments for preservice teachers to practice instructional decision making, offer formative feedback to students on case challenges via an automated essay scorer. This paper presents the results of a study that details students' use of the ETIPS automated essay scorer and their reactions to the formative feedback provided by it. Preliminary conclusions are drawn to aid in continued development of computer-aided formative assessment within the ETIPS learning environment.

Perspectives

Learning environments need to be, among other things, assessment-centered in that they provide opportunities for learners to revise, improve, and see progress in their thinking as well make learners' thinking visible (Bransford, Brown, & Cocking, 1999). Additionally, formative assessment can be aided by technology by providing not only means for students to represent their understandings, but also an automated source of feedback that can support students' revisions and improvements.

Most of the research on the use of technology-based formative assessment focuses on the use of computer-based assessments to provide students with practice testing opportunities and immediate feedback. For example, Peat and Franklin (2002) used online quizzes, mock exams, and self-assessment modules in a large, freshman biology class. Sambell, Sambell, and Sexton (1999) and Charman and Elmes (1998) used computer-based multiple choice assessments as

practice tests in an undergraduate engineering course and statistics course in geography, respectively. All three studies found that students reported tools used in formative fashion provided them with immediate feedback on their learning progress as well as directed their attention to areas needing further study.

Another application of technology for assessment purposes has been the use of automated essay scoring software, which is usually employed for summative purposes. Ellis Page and others developed the first generation of Project Essay Grader in the mid 1960s (Yang, Buckendahl, Juszkiewicz, and Bhola, 2002; Page, 1966, 1994). Since then, numerous technology-based software and approaches have been developed to measure writing quality. However, the research on these software applications tends to focus on their analysis approach and its accuracy as compared to human scorers (Valenti, Neri, & Cucchiarelli, 2003) rather than on students' responses to and experiences with the scorers.

The ETIPS automated essay scorer, developed in 2003, predicts students' scores on essays they complete in response to online cases that ask them to make an instructional decision regarding technology integration and implementation in a K-12 classroom. Unlike other automated scorers, the purpose of this scorer is to provide formative feedback to students on their essays so they may improve their responses before submitting them to their instructors for final evaluation.

Participants/Setting

Students enrolled in a post-baccalaureate program at the University of Minnesota - Twin Cities who were taking the required course EDHD 5007: Technology for Teaching and Learning during fall semester of 2003 were asked to participate in this study; all students in this course were seeking initial teaching licensure in secondary English. The instructor, who had previously

used the ETIPS cases in similar courses, assigned all students to complete two ETIPS cases as part of the course requirements. Students were not required to use the automated scorer, however. Thirty-four students were enrolled in the course; twenty-five of these students completed both pre- and post-assignment surveys and the two cases as assigned.

Within each ETIPS case, students take on the role of a first year teacher with a challenge of whether and, if so, how to integrate instructional technology in a simulated school setting. After accessing different information about the school environment, students provide answers to the challenges in the form of short essays for which the automated scorer can then provide feedback.

Data Sources

Data was gathered using both qualitative and quantitative methods including pre- and post-semester surveys, a user log of students' actions within the cases, and in-depth interviews with four selected students. The first survey (Appendix A) was given to students before the ETIPS assignment began. This survey contained items which asked students to self-rate their computer skills and comfort with computers, assess the value of computers, and indicate their philosophies of teaching using validated measurements located in the text *Instruments for Assessing Educator Progress in Technology Integration* (2000) by Knezek, Christensen, Miyashita, and Ropp as well as *Teaching, Learning, and Computing: A National Survey of Schools and Teachers Describing Their Best Practices, Teaching Philosophies, and Uses of Technology* (1998) funded by the National Science Foundation and the U.S. Department of Education. This pre-assignment survey also gathered background information about students' progress in their teacher licensure program.

The second survey (Appendix B) was given to students after they completed the two-case assignment. The post-assignment survey consisted of both open-ended and Likert-scaled questions developed by the researcher and the ETIPS cases project staff for the purposes of learning about students' experiences with the ETIPS assignment and assessment features, primarily the automated scorer. The pre- and post-assignment surveys each took approximately 5-10 minutes to complete during class time.

Data from both surveys was joined with information collected on each student's actions within the case including case information accessed, use of the automatic essay scorer, and final scores assigned by the instructor. Thirteen of the twenty-five students completing the surveys and case assignments also volunteered to take part in in-depth interviews (Appendix C) about their experiences with the automatic essay scorer.

The thirteen possible interviewees were sorted into different types of users according to a) their opinions of the automated scorer (taken from post-assignment survey), b) actual essay scores assigned by the instructor, and c) average number of drafts they completed per response section for both cases. The thirteen volunteers were distributed across five of these typologies (see Table 1) and four students, representing types A, C, F, and G, participated in interviews. These interviews were individually conducted, tape-recorded, and then transcribed and analyzed. In addition, these four students' case responses were analyzed for the number of drafts submitted for feedback, changes in draft length and substance, the ratings assigned by the automated essay scorer, as well as how the automated scores compared to the instructor's scores.

Table 1

Student Typologies

Typology A	**Typology B**	**Typology C**	**Typology D**
Positive opinion of scorer Received scores of 2 2+ drafts per response	Positive opinion of scorer Received scores of 2 0-1 drafts per response	Positive opinion of scorer Received scores of 0-1 2+ drafts per response	Positive opinion of scorer Received scores of 0-1 0-1 drafts per response
5 students	*1 student**	*2 students*	*(no volunteers)*
Typology E	**Typology F**	**Typology G**	**Typology H**
Negative opinion of scorer Received scores of 2 2+ drafts per response	Negative opinion of scorer Received scores of 0-1 2+ drafts per response	Negative opinion of scorer Received scores of 2 0-1 drafts per response	Negative opinion of scorer Received scores of 0-1 0-1 drafts per response
(no volunteers)	*3 students*	*2 students*	*(no volunteers)*

*Student declined to be interviewed.

Results

All twenty-five students engaged the automatic essay scorer and used it as a formative assessment tool during the first case but not the second case. In general, students used the essay scorer more with the first than second case – submitting an average of 3.61 drafts to the scorer on the first case and an average of 1.89 drafts on the second. Nearly all students used the essay scorer at least once before submitting their final answers. Only two students out of twenty-five did not submit a draft to the scorer before submitting their final answers and both did so only on the second case.

The four interviewed students' essay drafts were analyzed in greater detail and their revision and score records provide possible reasons why the first case tended to produce a higher number of drafts than the second case. In both assigned cases all four students attempted at least once to improve their essay responses before submitting them to their instructor. Most of the changes made were substantial in terms of writing content and quality. Despite these edits, the software log shows that the automated scorer did not reward the students' continued efforts with higher scores. Also, when an automated score was available for comparison on a final draft of a

student's writing, (meaning that a student submitted a draft for feedback, received automated feedback, made no changes/edits, and submitted the response to the instructor) the instructor's score was typically one point higher than score predicted by the software. When automated scores continually showed no improvements despite students' revisions, it appears these students lost confidence in the scorer's ability; several students directly stated that their experiences led them to outright disregard its feedback.

While the scorer may have caused frustration during the first case, that students did use the scorer more often during the first case is notable since the average number of drafts is positively related to the average final human score assigned (Spearman's rho=.42, $p < .04$). The same relationship is not present in the second case when students appeared not to rely on the scorer in forming their final essays (see Table 2). The number of drafts submitted during the first case was also positively related to the how much impact students reported the scorer having on their final essays (Spearman's rho=.46, $p < .05$). One open-ended question on the survey asked students to discuss the degree to which scores from their drafted responses impacted the final essays they submitted. Twenty-four of twenty-five students responded to this question and up to two responses were coded. The most frequent type of coded response, mentioned by two-thirds (n=16) of the respondents, was that they tried to please and then beat the scorer. The second most common type of coded response (n=9) was that they used the scorer then gave up.

Table 2

Mean Number of Drafts Submitted to Essay Scorer by Final Human Score

		Final Human Score		
		0	*1*	*2*
Case 1	Part 1	-	3.2	6.0
	Part 2	-	2.1	2.4
	Part 3	1.0	2.4	5.7
Case 2	Part 1	2.0	1.2	1.5
	Part 2	1.7	2.7	2.3
	Part 3	1.0	2.0	1.9

The four interviewed students suggested various reasons why they might have initially tried to respond to the scorer's feedback but then gave up. During the interviews, students became rather animated in describing their experiences with the scorer and its feedback, and, despite never being asked, all volunteered information about how the scorer and its feedback made them feel. Whether competition, guilt, desire, anger, or disdain, the ETIPS automated scorer and its feedback provoked emotional responses from these students.

Students' overall attitudes towards the automatic essay scorer were analyzed with two measures. The first measure was a summary scale that was constructed by combining four close-ended questions asking about the usefulness of the automated essay scores in composing essays and the user's confidence in the accuracy of the scorer (alpha=.83). The second measure was a single, related question asking students to rate on a five-point response set how much the automated scores impacted their final essays. Respondents' scores on the summary scale were unrelated to their skill or comfort with computers or the degree to which they thought technology was important in education, as were their responses to how much impact they felt the scorer had on their final essays. Rather, it was the total number of writing courses the respondent had taken in or outside their teacher education program that was positively related both to the summary

scale of the scorer's quality (Spearman's rho=.46, $p < .05$) and students' reports of how impactful the scorer was on their final essays (Spearman's rho correlation=.50, $p < .05$) (see Figure 1).

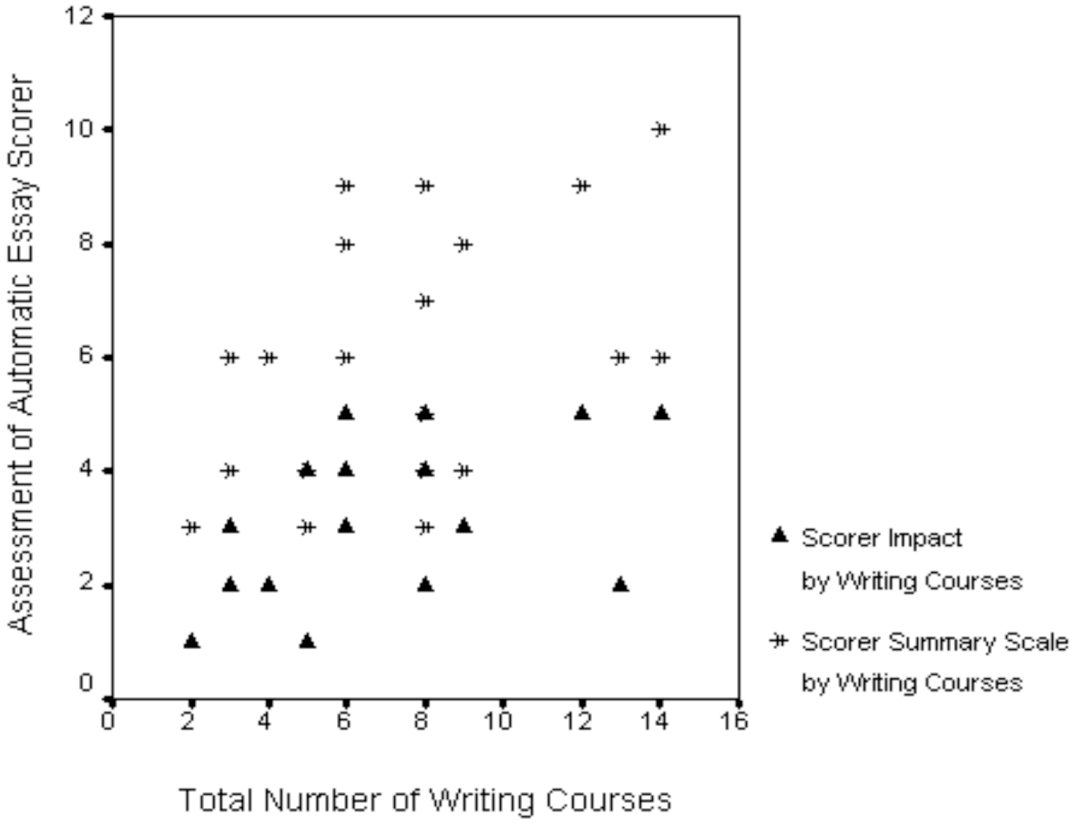

Figure 1. Assessment of Automatic Essay Scorer by Total Number of Writing Courses.

Implications and Conclusion

Although the ETIPS automated essay scorer represents a relatively new type of software application in instructional technology, 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. Extensive use of the ETIPS scorer was associated with improved essays as measured by their final human score. This outcome boosts confidence in the potential utility of instructional technology to aid in formative assessment. The results of this study also point to the importance

of the students' subjective experience with the formative assessment tool. While students used the scorer extensively in the first case, they believed that the scorer was not accurately measuring their improvements. In addition, the strong relationship between students' prior writing background and their opinion of the scorer suggests students' beliefs about the writing process impacted their use of the scorer.

Technology has the potential to foster more effective learning environments through providing formative assessment to learners. The in-depth data about students' reactions to and opinions of the current ETIPS automated essay scorer and its formative feedback presented in this paper provide insight as to the factors that must be considered in using technology to provide formative feedback; design, nature, accuracy, and specificity of the feedback are very important in aiding students' better performances on and positive experiences with learning environments.

References

Bransford, J. D., Brown, A. L., & Cocking, R. R. (1999). *How people learn: Brain, mind, experience, and school.* Committee on Learning Research and Educational Practice, National Research Council. Washington D.C.: National Academy Press.

Charman, D. & Elmes, A. (1998, November). A computer-based formative assessment stategy for a basic statistics module in geography. *Journal of Geography in Higher Education, 22*(3), 381-385.

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.

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.

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

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
Pre-Assignment Survey

Students' Experience with the Automated Scoring Feature of the ETIPS Cases

Introduction: This survey asks questions about your experiences with and use of computer technology, you as a teacher in training, and your beliefs about the place of technology in education. This survey helps inform a project doing research on technology that assists preservice teachers in practicing decision making about technology integration. There are no right or wrong answers. You may skip any question you do not feel comfortable answering.

1. Please list the email address* you will use to access the ETIPS Cases assignment: _____

**This information will only be used to link you with this survey, the post-survey; the online survey; and data generated from completing the assignment given to you by your instructor.*

Your Computer Skills

2. **Instructions:** The statements below refer to how confident you feel in your ability to do a technology-related task. Check the box that indicates your level of agreement or disagreement with each statement.
SD = Strongly Disagree**, D** = Disagree, **U** = Undecided, **A** = Agree, **SA** = Strongly Agree

I feel confident that I could...	***SD***	***D***	***U***	A	***SA***
a. Send e-mail to a friend.					
b. Send a document as an attachment to an e-mail message.					
c. Use an Internet search engine (e.g., Google or Alta Vista) to find web pages related to my subject matter interests.					
d. Search for and find the Smithsonian Institution website.					
e. Create my own website					
f. Use a spreadsheet to create a pie chart of the proportions of the different colors of M&Ms in a bag.					
g. Create a newsletter with graphics and text in 3 columns.					
h. Save documents in formats so that others can read them if they have different word processing programs (e.g., saving Word, ClarisWorks, RTF, or text).					

Comfort with Computers

3. **Instructions**: The statements below describe how different people feel about using computers. Check the box that best describes your agreement with each statement.
SD = **Strongly Disagree,** D = **Disagree,** U = **Undecided,** A = **Agree,** SA = **Strongly Agree**

	SD	***D***	***U***	***A***	***SA***
a. I get a sinking feeling when I think of trying to use a computer.					
b. Working with a computer makes me feel tense and uncomfortable.					
c. Working with a computer makes me nervous.					
d. Computers intimidate me.					
e. Using a computer is very frustrating.					
f. I feel comfortable working with a computer.					
g. Computers are difficult to use.					
h. I think that computers are very easy to use.					
i. I have a lot of self-confidence when in comes to working with computers.					
j. Computers are hard to figure out how to use.					

Assessing the Value of Computers

4. **Instructions**: The statements below describe different opinions about the value of computers or using and learning about computers. Check the box that shows your level of agreement or disagreement with each statement.
SD = **Strongly Disagree,** D = **Disagree,** U = **Undecided,** A = **Agree,** SA = **Strongly Agree**

	SD	***D***	***U***	***A***	***SA***
a. Students should understand the role computers play in society.					
b. All students should have some understanding about computers.					
c. All students should have an opportunity to learn about computers at school.					
d. Computers can stimulate creativity in students.					
e. Computers can help students improve their writing.					
f. Computers can aid in assessment of students.					

(Please turn to the next page.)

You as a Teacher

5. Content area(s) seeking licensure in: _____

6. What is your average grade in the teacher education courses you have taken so far? (check one)

____ A+	____ A	____ A-
____ B+	____ B	____ B-
____ C+	____ C	____ C-
____ D +	____ D	____ D-

7. How many college-level courses **outside of your teacher licensure program** have you taken that dealt specifically with the *mechanics of writing* and/or *writing assessment*? Circle your answer.

0 1 2 3 4 5 6 7 8 9 10

8. How many **teacher education courses** have you taken that dealt specifically with the *mechanics of writing* and/or *writing assessment*? Circle your answer.

0 1 2 3 4 5 6 7 8 9 10

9. Have you taken the course, *CI 5155: Contemporary Approaches to Curriculum: Instruction and Assessment* at the University of Minnesota?

YES NO

(Please turn to the next page.)

10. **Instructions**: The statements below describe different philosophies of teaching. Please check the box that indicates your agreement or disagreement with each statement.
SD = **Strongly Disagree,** D = **Disagree,** U = **Undecided,** A = **Agree,** SA = **Strongly Agree**

	SD	***D***	***U***	***A***	***SA***
a. It is better when the teacher – not the students – decides what activities are to be done.					
b. Instruction should be built around problems with clear, correct answers, and around ideas that most students can grasp quickly.					
c. How much students learn depends on how much background knowledge they have – that is why teaching facts is so necessary.					
d. Students should help establish criteria on which their work will be assessed.					
e. Assessment should talk place at the end of a project (summative assessment).					
f. Assessment should take place throughout a project (formative assessment).					

END OF SURVEY. THANK YOU!

APPENDIX B
Post-Assignment Survey

Students' Experience with the Automated Scoring Feature of the ETIPS Cases

Introduction: This survey asks you to reflect on your experiences with the ETIPS Cases and their assessment features. This survey helps inform a project doing research on technology that assists preservice teachers in practicing decision making about technology integration. There are no right or wrong answers. You may skip any question you do not feel comfortable answering.

1. Please list your name and the email address* you used to access the ETIPS Cases assignment for your course:

_____

**This information will only be used to link you with this survey; the pre-assignment survey; the online survey; and data generated from completing the assignment given to you by your instructor.*

Assessment Features of the ETIPS Cases

Automated Essay Scorer

2. To what extent were the automated essay scores useful or not useful in composing your responses to the case challenges? Circle your response.

Not at all useful	**A little useful**	**Somewhat useful**	**Useful**	**Very useful**	***Did not use***

3. How confident were you in the accuracy of the scores that the automated scorer assigned to your responses? Circle your response.

Not at all confident	**A little confident**	**Somewhat confident**	**Confident**	**Very confident**	***Did not use***

4. How accurate do you think the automated scorer was in scoring your responses? Circle your response.

Not at all accurate	**A little accurate**	**Somewhat accurate**	**Accurate**	**Very accurate**	***Did not use***

5a. To what extent did the automated scores you received on your *drafted* responses impact your *final* responses to the case challenges?

No impact	A little impact	Somewhat impacted	Impacted	Impacted a lot	*Did not use*

5b. Please discuss your response to question 5a in further detail.

6. How easy to interpret were the automated scores? Circle your response.

Difficult to interpret	A little difficult to interpret	Somewhat difficult to interpret	Easy to interpret	Very easy to interpret	*Did not use*

7. List any comments and/or suggestions you have for the developers of the ETIPS Cases regarding the automated essay scorer.

Rubrics

8. To what extent were the rubrics useful or not useful in responding to the case challenges? Circle your response.

Not at all useful	A little useful	Somewhat useful	Useful	Very useful	Did not Use

9. To what extent did the rubrics impact your final responses to the case challenges? Circle your response.

No impact	A little impact	Somewhat impacted	Impacted	Impacted a lot	Did not use

10. Please comment below on the quality of the rubric used to evaluate your responses to the case challenges. Did you feel the rubric appropriately captured the different qualities of possible responses to the challenges? Do you have any suggestions on how to improve the rubric? (A copy of the rubric is attached as the last page of survey.)

Combination of Features

11. Please circle below any features you used during the ETIPS Case assignment to help develop your responses to the case challenges.

a. Automated essay scores

b. Rubrics

c. Search-Path Map

d. (other, please list) _____

12. In the lines provided to the left of the features listed below, please rank the order of importance the features played in helping you form your responses to the case challenges. *(1 = most important, 2= next important, 3 = least important, 4= did not use)*

a. _____ Automated essay scores

b. _____ Rubrics

c. _____ Search-Path Map

d. _____ (other, please list) _____

General Questions

13. Explain your rationale for composing or not composing multiple drafts of your responses for your ETIPS Cases assignment.

14. To what extent were the ETIPS Cases useful or not useful in learning about technology use in education? Circle your response.

Not at all useful	A little useful	Somewhat useful	Useful	Very useful

15. What, if anything, did you learn about technology integration from the ETIPS Cases and how they were used in class?

16. What was the *most* helpful aspect of the cases and how they were used in the class?

(Please turn to the next page.)

17. What was the *least* helpful aspect of the cases and how they were used in the class?

18. Did you experience any technical difficulties in completing the cases? If so, please explain.

19. Would you be willing to participate in a short interview (20 minutes) with the researcher about your experiences with the ETIPS Cases and its assessment features?

20. If you answered "Yes" to question #16 and would be willing to participate in a short interview about your experiences with the cases, please list below your preference for how the researcher should contact you.

END OF SURVEY. THANK YOU!

APPENDIX C
Interview Questions

Students' Experience with the Automated Scoring Feature of the ETIPS Cases

NAME OF INTERVIEWEE: _____

Interviews will be conducted in person.

Pull up example of a case on computer!!!

Purpose of interview: The researcher will inform the interviewee the reason for the interview, which is to learn about his or her experience with the ETIPS Cases and the new automated scoring feature. ETIPS Cases is trying to improve its product, so the more candid you are the better.

I need to ask your permission to tape record the interview. Do I have your permission, or would you prefer not to be recorded? _____

1. If you could describe your experiences with the ETIPS Cases in general in one word (adjective), what would that word be? Explain.

2. In regard to your ETIPS Cases assignment you completed for your class, explain for me how you approached responding to the questions asked of you at the end of each case.
    - When did you look at the answer page?
    - Did you answer the three questions one at a time or all at once?
    - Did you go back and forth between the answer page and the school website to get your answers?
    - Were there any differences in how you responded to the first and second cases?

3. Tell me about your response or reaction to the automated scoring feature of the ETIPS Cases.
    a. What was your initial reaction to the automated scoring feature? WHenyou were told about it, AND/OR when (or if) you used it.

    b. Did your reaction to the automated scoring feature change between the beginning and end of the assignment, in other words from the first case to the second case? Why or why not?

    c. How did you respond to the actual automated scores themselves?

    d. Did you feel like you needed to "beat" the scorer, or get a high score before submitting your final answers?

    e. Did you access the explanation of automated feedback during your assignment?

4. The automated essay scorer is a new assessment feature of the ETIPS Cases. Its purpose is to provide formative feedback to users. [DESCRIBE FORMATIVE VS. SUMMATIVE ASSESSMENT] The feedback is formative in nature because the learners receive feedback on drafts of their responses before their final answers are submitted to their instructors. The idea is to help users craft "better" responses.

In your opinion, is the automated essay scorer an effective means of formative assessment? Why or why not?

5. Do you have any ideas about other means the ETIPS Cases could use to supply users with formative feedback about their responses to the challenges and/or performances with the cases? Or, do you have any suggestions about better ways to explain or implement the automated scorer?

6. Computers are being used more and more as assessment tools. Examples are the GRE and PRAXIS tests.
    a. How do you feel about computers being used to assess *learning in general*?

    b. How do you feel about computers being used to assess *writing*?
        - Do you think your position as a preservice English teacher impacts your opinion about computers being used to assess writing?

7. I have asked you all of the questions I had prepared.
    a. Is there anything you care to add?

    b. Do you have any questions for me?