J·T·L·A

The Journal of Technology, Learning, and Assessment

Volume 7, Number 1 · September 2008

Students' Experiences with an Automated Essay Scorer

Cassandra Scharber, Sara Dexter, Eric Riedel

www.jtla.org

A publication of the Technology and Assessment Study Collaborative
Caroline A. & Peter S. Lynch School of Education, Boston College

J·T·L·A

Volume 7, Number 1

Students' Experiences with an Automated Essay Scorer

Cassandra Scharber, Sara Dexter, Eric Riedel

Editor: Michael Russell
russelmh@bc.edu
Technology and Assessment Study Collaborative
Lynch School of Education, Boston College
Chestnut Hill, MA 02467

Copy Editor: Jennifer Higgins
Design: Thomas Hoffmann
Layout: Aimee Levy

JTLA is a free on-line journal, published by the Technology and Assessment Study Collaborative, Caroline A. & Peter S. Lynch School of Education, Boston College.

Preferred citation:

Scharber, C., Dexter, S., Riedel, E. (2008). Students' Experiences with an Automated Essay Scorer. *Journal of Technology, Learning, and Assessment*, 7(1). Retrieved [date] from http://www.jtla.org.

J·T·L·A

Abstract:

The purpose of this research is to analyze pre-service teachers' use of and reactions to an automated essay scorer used within an online, case-based learning environment called ETIPS. Data analyzed include post-assignment surveys, a user log of students' actions within the cases, instructor-assigned scores on final essays, and interviews with four selected students. These in-depth data about students' reactions to and opinions of the ETIPS automated essay scorer help inform the automated essay scoring field about users' perceptions of automated scoring.

J·T·L·A

Students' Experiences with an Automated Essay Scorer

Cassandra Scharber
University of Minnesota
Sara Dexter
University of Virginia
Eric Riedel
Walden University

Introduction

The purpose of this research is to analyze pre-service teachers' use of and reactions to an automated essay scorer as a means of formative feedback on essay drafts composed within an online, case-based learning environment.

Literature Review

Technology holds much promise for contributing to the increased practice of formative assessment in education. Formative assessment evaluates student work as part of a continuum of growth toward increasing quality or degree of expertise rather than on a dichotomous, right or wrong basis (Sadler 1989). The purpose of formative assessment is *for* learning rather than *of* learning, which is typically the purpose of summative assessment (Black & Wiliam, 1998b; Pellegrino, Chudowsky, & Glaser, 2001). The literature makes a strong case for the importance of formative assessment because this type of assessment has been proven to produce specific gains for learning (Barron, et al., 1998; Black & William, 1989a, 1989b; Black & Harrison, 2001; Peat & Franklin, 2002). Black and William (1989b) argue that formative assessment is "at the heart of effective teaching" (p. 140). Clearly, although formative assessment is important, it is often overlooked and little has been written about it in the literature pertaining to formative assessments within online learning environments (Riedel, Dexter, Scharber, & Doering, 2006; Scharber, Dexter, & Riedel, 2005).

One application of technology for assessment purposes has been the use of automated essay scoring software, which is usually employed for summative rather than formative purposes and is designed to reduce costs and increase reliability in writing assessments (Dikli, 2006). Computer-based writing has been the subject of evaluation research for decades (Page, 1966, 1994; Page & Peterson, 1995; Shermis & Burnstein, 2002). Numerous systems and approaches have been developed to measure writing quality since the first system, Page Essay Grade/PEG, including Criterion© and e-rater© (from ETS), MY Access© (from Vantage Learning), and Intelligent Essay Assessor© (from Pearson Knowledge Technologies). The Educational Testing Service (ETS) currently leads the field in terms of effective and accurate automated essay scoring with its e-rater system, which is currently being used for scoring General Management Aptitude Test (GMAT) essays (Burstein, 2003; Dikli, 2006).

Research on these and other available automated essay scorers primarily focuses on the analysis approaches, accuracy of the scores as compared to human scorers, and reliability of automated essay scoring systems (Valenti, Neri, & Cucchiarelli, 2003; Warschauer & Ware, 2006) rather than on students' responses to and experiences with automated essay scorers. A notable exception is the work of Grimes & Warschauer (2006) in which use, attitudes, and usage patterns of secondary students and teachers when using two different AES systems are explored. Despite the growing corpus of AES literature, one area that has been overlooked is users' perceptions and responses to automated essay scoring and feedback. In an attempt to shed light into this area, this manuscript explores students' experiences with an AES that provided formative writing feedback within an online educational context.

Online Environment and Classroom Setting

The data presented here are based upon learners' experiences with an automated essay scorer that provides formative feedback to learners while they work through cases delivered within an online learning environment called ETIPS. The ETIPS cases (located at www.etips.info) respond to the need for improved and new types of assessment as well as the need in teacher education to support pre-service teachers' preparation for technology integration and implementation in their future classrooms.

ETIPS Environment

The ETIPS cases are multimedia, network-based, online instructional resources that provide learning opportunities with embedded assessment features for pre-service teachers to practice instructional decision-making skills related to technology integration and implementation. The topics of the ETIPS cases are correlated with the National Educational Technology Standards for Teachers (NETS-T) and the Interstate New Teacher Assessment and Support Consortium (INTASC) standards and anchored in the research surrounding teaching, learning, and assessment with technology. The ETIPS cases were designed to provide a virtual school context to allow pre-service teachers to practice applying instructional decision-making skills related to technology integration and implementation suitable for use in either methods or educational technology courses. The embedded assessment features are designed to support formative assessment as students work on the case. In this manuscript we focus on one of ETIPS' embedded assessment features, its automated essay scorer, which is discussed in terms of students' uses and opinions.

ETIPS cases ask pre-service teachers to bring together technological, pedagogical, and content knowledge and apply it while imagining themselves in a particular classroom within a particular school. Each case consists of an introduction that frames the decision students must make, a virtual school's web site, and a response page. Technology integration cases, such as the ones students completed for this study, are grounded in a conceptual framework consisting of six principles (Dexter, 2002), which are offered as an explanation for the conditions that are essential for the effective use of educational technology in the classroom.

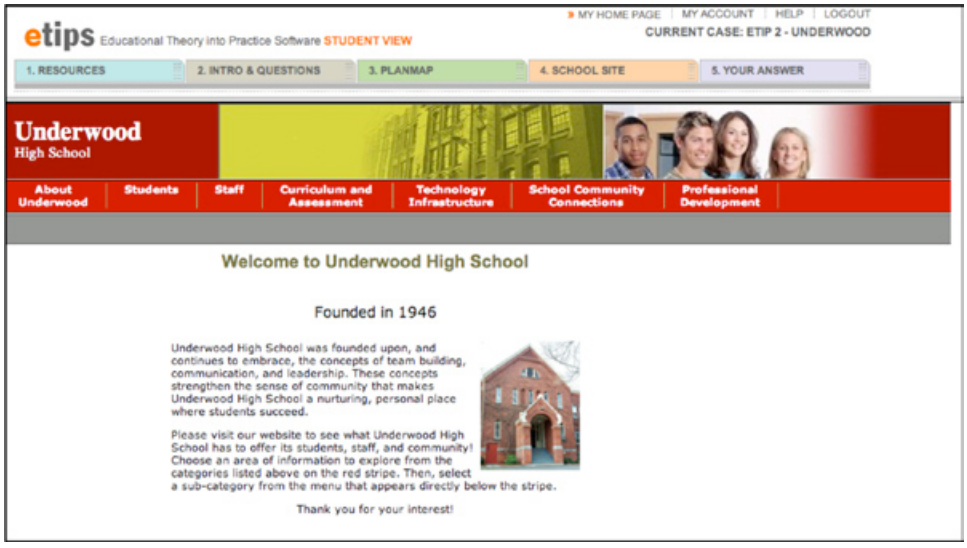
Both cases in this research were grounded in the second principle—*technology provides added value to teaching and learning*. Therefore, in the case introduction, students were assigned the role of a first year teacher with an instructional decision to make about integrating instructional technology in his/her classroom. In this study, the key challenge to which students responded was about how technology could be integrated in his/

her classroom so as to meet learner's needs (Appendix A, page 32, contains the entire case introduction):

> The principal was pleased with your first [classroom] observation. For your next observation she challenged you to 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.

The school setting (i.e., case information) is presented to learners via the virtual school's website (Figure 1). Thus, ETIPS cases differ from traditional cases that are typically presented as a narrative, and in a linear fashion. Presenting the case information in parts identified as web pages forces students to select information categorically rather than receive it linearly. Students can explore any of the school web pages in order to find the information they believe they need to make their instructional decision. Menu item categories at these school sites are labeled About the School; Students; Staff; Curriculum & Instruction; Technology; Community; and Professional Development. Each category has three to seven sub-categories to select from, where specific information appears. It is important to note that not all menu items are relevant for each challenge. The goal of ETIPS is to help foster decision-making skills. Because schema drive decisions, a related goal is that as learners complete multiple cases their processing of input from class discussion and instructor feedback should help them develop schema about making instructional decisions like the one posed in the case.

Figure 1: Screenshot of a Welcome web page of an ETIPS virtual high school called Underwood



Note: School information can be accessed by clicking on the red tabs across the top of the screen.

After investigating the school environment, students provide responses to three questions in the form of short essays. The questions to which students respond and the rubrics used by instructors to evaluate student essays were designed to emulate the decision-making process as outlined by Marzano and Pickering (1997) (Appendix B, page 33):

> Question 1 (Q1)—*Confirm the challenge:*
> What is the central technology integration challenge in regard to student characteristics and needs present within your classroom?
>
> Question 2 (Q2)—*Identify evidence to consider:*
> What case information must be considered in making a decision about using technology to meet your learners' diverse needs?
>
> Question 3 (Q3)—*State your justified recommendation:*
> What recommendation can you make for implementing a viable classroom option to address this challenge?

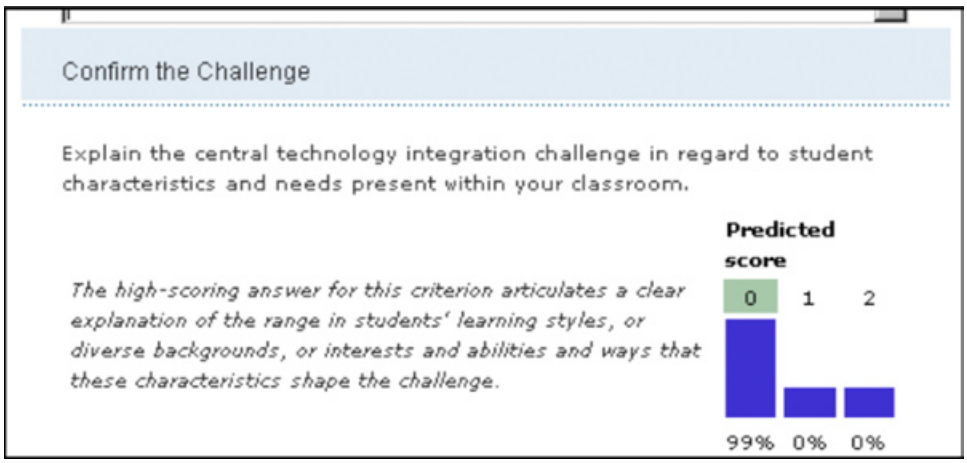
For the two assigned ETIPS cases, students were presented with the same introduction/challenge, but the school setting (i.e., website) changed in each case. The first case took place in a high-performing, urban middle school called Cold Springs Middle School, and the second case took place in a low-performing, suburban high school named Underwood High School. By addressing the same challenge in two very different settings, the cases emphasize how context can influence the decisions teachers make.

Each time students completed a case, they had opportunities to: (a) complete as many drafts as they chose for the three case questions, (b) receive formative feedback via the automated essay scorer, (c) return to the case context (school website) to gather more information to inform revisions to their responses, and (d) submit their final responses to receive feedback and scores from their instructor. Students also had online access to the rubric their instructor would use to evaluate their responses.

Access to the ETIPS' automated essay scorer (AES) during a case is positioned such that students can elect to submit their answer drafts and receive a prediction of their score, according to the rubric criteria. During this study, if students chose to submit their responses to the case's questions for feedback from the AES, they received (a) an estimated score (0–2), which is based on established rubrics (Appendix C, page 34), and (b) a short explanation of a "good" answer (Figure 2, next page) and, for the second case question (Q2), this feedback might also include suggestions about what other parts of the school web site the student might consider going to for helpful information. By providing students access to the AES before they submit their final responses to their instructors, this score and the related feedback were intended as formative feedback that could be utilized by students to improve their responses. While not as detailed or insightful as feedback a human might provide, it nonetheless provides some information about the answer's content and quality.

Once students obtained automated feedback, they could choose to go back into the case's school context to search for additional information to incorporate into their responses, re-draft their responses, re-submit their response to the AES, or submit their responses as final answers to their instructor. When the instructor scored student essays, s/he was able to view the number of drafts a student submitted for automated feedback as well as the estimated score the scorer gave them, though instructors were not able to view the actual drafts. Students were aware that teachers have access to information regarding their drafting attempts and automated scores.

Figure 2: Example of feedback generated by the ETIPS automated essay scorer



The question is stated, followed by an explanation of the characteristics of a high-scoring answer. A bar graph predicts the certainty that the answer will receive the score of a 0,1, or 2 against the scoring rubric's criteria.

Developed in 2003, the ETIPS automated essay scorer uses a Bayesian model to score essays both for content and style, examining various features of essay responses including vocabulary, word usage, specific phrases, and grammatical structure. The software then compares these features in students' essay drafts to those same features in training essays that have already been scored reliably by humans guided by rubrics. By examining the correlations between students' essays and the training essays, the software predicts how likely students working with ETIPS cases are to receive a score of 0, 1, or 2 from their instructors against the rubric provided to assess student responses. The ETIPS automated essay scorer is context-specific; it cannot evaluate essays outside of the ETIPS learning environment or essays that deal with other than ETIPS case -specific questions and topics. At the time of this study, the ETIPS scorer was in its first generation of development and was being used with only selected cases for

experimental purposes; therefore, there were no performance data available regarding the accuracy of the scorer.

Classroom Context

This study was conducted at a large public Midwestern university that has a post-baccalaureate program for majors in a variety of content areas to earn their teaching certification. These programs require students to take a technology-integration course, which is taught in content-specific cohorts for each of the licensure areas, including sections for elementary education, agriculture, art, business, early childhood, English, family education, math, physical education, science, second languages, social studies, and special education. The cohort invited to participate in this study was the English cohort, which was composed of thirty-four secondary English pre-service teachers enrolled during fall semester of 2003.

Typical assignments in this course include having students (a) learn software that is relevant for the content and ages of students they will teach, (b) develop an example product with the software, and (c) describe how specific software could enhance students' learning. For final projects, students developed a unit or lesson plan that integrates at least two of the technologies used in class, and delivered parts of the unit/lesson to their peers in class. As a part of the English cohort's educational technology course's overall emphasis on developing students' technological pedagogical content knowledge, the course instructor, who had several years of experience using the ETIPS cases, assigned all students to complete two ETIPS cases as a required assignment.

Methodology

Purpose

This study utilized a mixed-methods, multiple case-study approach to learn about pre-service teachers' experiences with and responses to the ETIPS automated essay scorer and its formative feedback. Twenty-five of the thirty-four post-baccalaureate students in the cohort seeking initial teaching licensure in English agreed to participate in the study by completing the two cases as assigned and pre- and post-assignment surveys. Then, based upon their responses in the survey, a sample of these students was invited to participate in individual interviews. Students were allowed two weeks to complete their two assigned cases. For purposes of the assignment and this research, students were not required to use the ETIPS automated scorer, however it was available for them to use during both cases.

Data Sources

Data were gathered using both qualitative and quantitative measures including pre- and post-assignment surveys, a user log of students' actions within the cases, and in-depth interviews with four selected students.

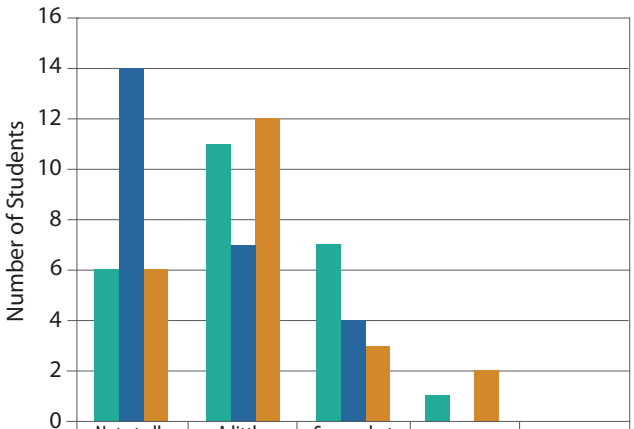
Pre-assignment surveys gathered demographic data from students in addition to asking questions about their experiences with and use of computer technology and their beliefs about the role of technology in education (Appendix D, page 35). Responses to the pre-assignment survey were used in developing the descriptions of the four students' experiences that are detailed later in this manuscript. The post-assignment survey was a paper-based survey administered by one of the co-authors after students completed the two-case assignment. The post-assignment surveys took approximately five-to-ten minutes to complete during class time. This survey consisted of both open-ended and Likert-scaled questions developed by the researchers for the purpose of learning about students' experiences with the ETIPS cases and its assessment features, primarily the automated scorer (Appendix E, page 39). Close-ended items were imported into SPSS 13.0 for analysis. Open-ended questions were coded using a grounded theory approach (Glaser & Strauss, 1967). In addition, while students worked on the assignment, ETIPS software collected data on individual students' case use using a log file that (1) tracked the school web pages students viewed inside the two case contexts, (2) recorded all drafts of answers the students submitted for automated feedback and the corresponding automated predicted scores, and (3) documented the instructor's scores and comments on the final version of the essay the student submitted. The log file data was then imported into an SPSS 13.0 file and merged with the post-assignment survey data.

Thirteen of the twenty-five students completing the surveys and case assignments volunteered to take part in in-depth interviews about their experiences with the ETIPS cases and its automated essay scorer. A purpose sampling method was used to select four respondents based on their opinions of the automated scorer (taken from post-assignment surveys), and their actual essay scores assigned by the instructor (Table 1). This sampling method identified students with diverse experiences with the AES and reactions to those experiences. Specifically, students whose score on an AES Effectiveness Scale, created from their responses to post-assignment survey questions and discussed in detail in the following section, ranged from 3–5 were classified as having a negative opinion while students whose score on the scale ranged from 6–15 were classified as having a positive opinion (Figure 3, next page). Students who received a mean instructor-assigned score of 0–1 were classified as low-performing while students who received a mean instructor-assigned score of 2 were classified as high-performing.

Table 1: Typology used to select interviewees

		Instructor-Assigned Scores	
		High (2)	**Low (0–1)**
Respondents' opinion of AES [AES Effectiveness Scale score]	**Positive** (scale score range 6–15)	Moderate expectations with constructive revisions *N = 6; Interviewee = Stacey*	Focused on understanding how it works *N = 2; Interviewee = Mitch*
	Negative (scale score range 3–5)	Low expectations foreclose use *N = 2; Interviewee = Erica*	High expectations not met *N = 3; Interviewee = Adam*

Figure 3: Student responses to post-assignment questions of AES effectiveness



	Not at all useful/ confident/ accurate	A little useful/ confident/ accurate	Somewhat useful/ confident/ accurate	Useful/ confident/ accurate	Very useful/ confident/ accurate
To what extent were the automated essay scores useful or not useful in composing your responses to the case challenges?	6	11	7	1	0
How confident were you in the accuracy of your scores that the automated scorer assigned to your responses?	14	7	4	0	0
How accurate do you think the automated scorer was in scoring your responses?	6	12	3	2	0

Semi-structured interviews were used to encourage students to provide rich details about their experiences with the cases and the automated scorer (Appendix F, page 42). These interviews were conducted after the assignment was complete when students had access to the instructor's scores and feedback about their responses. The four interviews were recorded, transcribed, and analyzed by the researchers. Responses to questions were coded using a grounded theory approach (Glaser & Strauss, 1967). Finally, log data from the software was accessed after assignments were completed and interviews were conducted to triangulate student responses and help braid together a cohesive case profile for each of the four interviewed students.

Analysis

For the twenty-five students completing the surveys and two ETIPS cases, researchers calculated the average number of drafts per case and per case question (Q1, Q2, Q3); explored the relationships between number of essay drafts and (a) final instructor scores and (b) impact of automated scores on final essays as reported by students on the post-assignment survey.

In addition, the sub-set of four interviewed students' ETIPS essays and log data that contained revision and score records (e.g. 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) were analyzed in greater detail. Four detailed narratives were constructed in order to illustrate these students' use of the ETIPS automated scorer and the nature of their experiences with the scorer and its formative feedback.

Results

Survey Data

Three close-ended questions from the post-assignment survey were used to create an AES Effectiveness Scale of the respondent's overall opinion of the AES (Figure 3, page 13). These questions asked how useful students found the AES, how confident they were in the AES scores, and how accurate they thought the AES scores were. In general, most students did not assign strong positive ratings to any aspect of the scorer. They were more likely to describe the AES as useful in helping them to compose their own responses than to assign confidence in the AES or belief in its accuracy. Responses to each item were added together to form the AES Effectiveness Scale (Cronbach's alpha = .83; range = 3–15; mean = 5.68; *s.d.* = 2.16) which was used as a partial criterion in selecting respondents for semi-structured interviews.

The post-assignment survey also asked respondents several open-ended questions about the automatic essay scorer. The first question asked students to discuss their response to a close-ended question on the degree to which the automated essay scorer impacted their final submitted responses. Twenty-four of twenty-five students responded to this question and up to two responses were coded for each student. The most frequent type of 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 response (n = 9) was that they "used the scorer then gave up."

The second open-ended question asked students to explain their rationales for the number of drafts they composed for their essay responses. The most common response, mentioned by eleven of twenty-four respondents, was that they were responding to the feedback given by the automatic essay scorer and attempting to receive higher scores.

Students used the AES 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 case (Table 2). Nearly all students (23) 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 automated scorer before submitting their second case analysis to the instructor. The AES returned relatively low scores to students, especially on the first case. AES provided the lowest scores for the first question on the first case with 81.1% of scores generated by AES equal to 0. AES provided the highest scores for the first criterion on the second case with 61.0% of scores equal to 1 and 30.5% of scores generated by AES equal to 2.

Table 2: Number of drafts submitted to AES with distribution of AES scores

School	Question	Mean # of drafts submitted to AES	Distribution of AES Scores			Mean AES Score	N
			0	1	2		
1	1	4.88	81.1%	7.6%	11.4%	0.30	132
	2	2.28	26.0%	51.9%	22.1%	0.96	77
	3	3.68	78.1%	13.2%	8.8%	0.31	114
2	1	1.48	8.5%	61.0%	30.5%	1.22	59
	2	2.28	44.9%	37.2%	17.9%	0.73	78
	3	1.92	48.6%	30.0%	21.4%	0.73	70

The final scores assigned by the instructor were typically higher than the range of scores generated by AES (Table 3, next page). Unlike the AES scores, there did not appear to be an increase in instructor scores between the first and second case. If the difference between AES scores assigned to the last draft submitted and the instructor-assigned scores is examined, the difference ranges from 0.7 points to 1.5 points (Table 4, next page). Thus, it appears that the AES systematically undervalued student performance in the cases in comparison to instructor judgments.

Table 3: Distribution of instructor-assigned scores

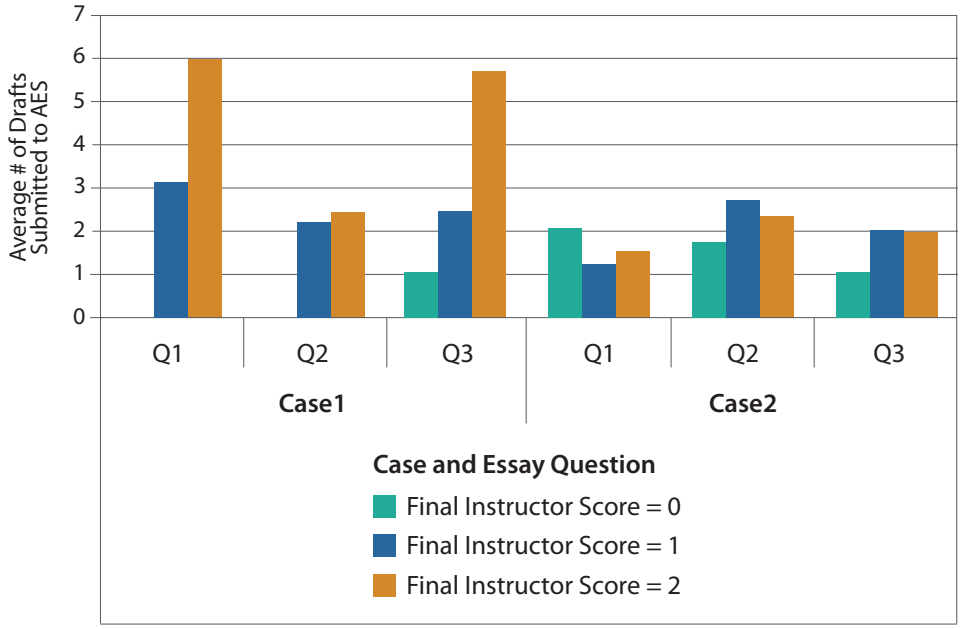
School	Question	Distribution of Human Scores			Mean Human Score	N
		0	1	2		
1	1	—	40.0%	60.0%	1.6	25
	2	—	32.0%	68.0%	1.68	25
	3	4.0%	56.0%	40.0%	1.36	25
2	1	8.0%	20.0%	72.0%	1.64	25
	2	—	44.0%	56.0%	1.56	25
	3	4.0%	36.0%	40.0%	1.56	25

Table 4: Accuracy of instructor versus AES scores

	Accuracy *(human score – last draft if exact duplicate)*		
	N	Mean Difference	Median Difference
Case 1 Part 1	18	1.50	1.50
Case 1 Part 2	15	0.87	1.00
Case 1 Part 3	18	1.28	1.00
Case 1 Overall	**14**	**1.19**	**1.33**
Case 2 Part 1	18	0.67	1.00
Case 2 Part 2	14	1.21	1.00
Case 2 Part 3	15	1.27	1.00
Case 2 Overall	**9**	**1.14**	**1.17**
Overall	**7**	**1.11**	**1.00**

Not only did students use the scorer more often during the first case, but the average number of drafts in the first case is positively related to the average final instructor score assigned (Spearman's rho = .42). The same relationship is not present in the completion of the second case when students reduced their use of the automated scorer in completing their final essays (Figure 4, next page). The number of drafts submitted during the first case was also positively related to the perceived impact students reported the automated scores having on their final essays (Spearman's rho = .46, $p < .05$).

Figure 4: Mean number of drafts submitted to AES by case, essay question, and final instructor-assigned score



Individual Student Cases

The in-depth interviews with four students and the software's log illustrating the series of revisions made to their answers along with how the AES scored each draft provides further insight into the post-survey results. These four students were selected from each of the four typologies depicted in Table 1 on page 12. The following narratives illustrate individual student's use of the ETIPS automated essay scorer and the nature of their experiences with and opinions about the scorer and its formative feedback.

Adam

Adam is a member of the group of students who expressed a somewhat low opinion of the automated scorer and received low scores from the instructor. Adam, who has a bachelor's degree in English, said during his interview that while he thought technology could aid formative assessment of students' written work, it would require first that students understand how the technology tool worked and the tool's limitations. He said that he did not trust a computer to assess writing in a summative fashion, but added that if he did work with a tool long enough to find it a reliable judge of writing he would probably use it. He also shared that "if I had my students spend the time working on an essay or paper, I think that I would owe them the time to look through it and spend some time on it."

In each of the two ETIPS cases, Adam was to draft three short responses to questions (Q1, Q2, Q3) that would comprise his final essay. Across these six opportunities to receive formative feedback from the automated essay scorer, Adam sought an automated score three times, all in the first case (Table 5). He spent 29 minutes writing and revising his essay responses. In the first case he began by drafting a response to each of the three questions and submitting them for automated scores; for each of these responses the predicted scores was 0 (out of possible 2). In subsequent drafting he worked on all three of the question parts in turn, making changes such as taking away a summary statement, adding examples, and using key words from the rubric and/or case question. In summary, Adam made a total of five rounds of revisions, focusing mostly on Q1. Only in one round of revisions and for one part of his essay did his predicted score improve. The comments the instructor provided and the rubric criteria indicate that the instructor did not think Adam provided enough specific case information in his essay, while acknowledging that Adam was, in general, addressing the topic of the case. Adam received scores of 1 from his instructor on all three questions (Table 2, page 15).

Table 5: Overview of Drafts and Scores for Adam, Case 1

Draft Round	Responses Submitted and Nature of Edit	Automated Scores
1	• Q1 initial response • Q2 initial response • Q3 initial response	0 0 0
2	• Q1 Removed a summary statement	0
3	• Q1 Added an example, but not from case information • Q2 Added an example, but not from case information • Q3 Added a summary statement with key words from the rubric and case question	0 0 0
4	• Q1 Completely reworded answer, while making same points	0
5	• Q1 Added an explanatory and summary statement, using case information • Q2 Added an example from the case	0 1
		Instructor Scores
6	• **Q1 final answer** • **Q2 final answer** • **Q3 final answer**	**1** **1** **1**

Adam overall approached case two quite differently than ETIPS case one (Table 6, next page). He submitted his initial responses for all three answer parts and, despite getting predictions of 1s and 0s, did not revise

those responses and submitted his initial responses without further revisions as his final answers. During his interview, he described that in the second case he was more focused and that he did not try to make the automated essay scorer predict a high score for him. Scoring against the provided rubric, his instructor assigned scores of 2, 1, and 1, respectively. The professor's comments to Adam on case two were positive, telling Adam he was "right on in your analysis" for Q1, and that his response to Q3 was "creative" but that he did not score a 2 because it still lacked discussion of a particular idea (see Table 6).

Table 6: Overview of Drafts and Scores for Adam, Case 2

Draft Round	Responses Submitted and Nature of Edit	Automated Scores
1	• Q1 initial response • Q2 initial response	1 1
2	• Q3 initial response	0
		Instructor Scores
3	• **Q1 final answer** • **Q2 final answer** • **Q3 final answer**	**2** **1** **1**

During his work formulating an essay for each case, Adam indicated that he drew upon the rubric and the automated essay scores, in relative order of importance. On the post-assignment survey he rated the automated scorer as only "a little useful" and marked that he was "not at all confident" of it and that the scorer was "not at all accurate." His assessment of the tool was warranted given that its automated predictions rarely matched his instructor's assigned scores. In his interview he described how initially he thought the scorer was "fantastic" and "really cool" and was very curious about how it worked. Adam said he approached the first case with a playful attitude, following the professor's instruction to explore the case, and try out the automated essay scorer, but with the caveat that only the instructor's score that would be factored into his grade for the assignment. His experience with the automated essay scorer in the first case quickly made Adam frustrated when it did not return predictions of high scores. He expressed that his inability to obtain high, automated scores from the automated essay scores triggered his competitive spirit such that his efforts to "beat" the automated scorer distracted him from focusing on the case, its questions, and the overall purpose of the assignment. Adam also described how he took on the attitude that ETIPS was a game he wanted to win and became frustrated when, through his self-described writing talent, he could not do so. After the first case he compared experiences with classmates and learned that they, too, did not

receive higher automated predictions. Adam concluded then he would not let the automated score predictions bother him and recognized he should "just let it go." He described that after he did so, his experience with the second ETIPS case was more enjoyable and that he explored nearly every menu item and focused more on the case specifics.

Mitch

Mitch had a positive opinion of the scorer, but received lower scores on average (0s and/or 1s) from the instructor. When asked in an interview what he thought of computers being used to assess writing, Mitch thought computers could be used to assess writing as long as there were definite answers to the questions being asked: "in theory, you could write a good answer that does not correlate with the computer's answers, so you would receive a bad score, but your answers would not be bad." Mitch further noted, "a computer cannot comprehend." He did concede that the scorer's positive feature was that " it encourages you to craft answers....I think the scorer has potential as long as it is accurate."

Across the two cases' three short responses, Mitch took three of the six opportunities to revise his answers in response to formative feedback, and all of these were in the second case (Table 7). He described his approach to the assignment as moving from the cases' homepage to the links he was interested in, then submitting Q1 after which he returned to the case to look for more information to answer Q2, and then repeating this process for Q3. This process took him about 49 minutes. The log shows that for case one Mitch submitted Q1, Q2, and Q3 in separate draft rounds, and while his predicted scores for each were 0, he then still submitted these initial drafts to his instructor as his final responses. His instructor scored his responses as a 1, 1, and 0 for Q1, Q2, and Q3, respectively (Table 7).

Table 7: Overview of Draft Approach and Results for Mitch, Case 1

Draft Round	Responses Submitted and Nature of Edit	Automated Scores
1	• Q1 initial response	0
2	• Q2 initial response	0
3	• Q3 initial response	0
		Instructor Scores
4	• **Q1 final answer** • **Q2 final answer** • **Q3 final answer**	**1** **1** **0**

Mitch took a different approach to case two, revising two of the three answer parts twice and one part once (Table 8). He took a total of 40 minutes with this second case. In his interview he said he first went to the answer page and looked at what he would be required to do. He began by writing out his justified recommendation (Q3). He revised this once after the automated scoring software returned a "null" result, meaning that it did not have enough information to predict a score. Mitch then added another sentence to that answer part, elaborating upon a part of his answer. He submitted this version for feedback and then drafted and submitted part two of the answer, both times receiving predicted scores of 0. In round four he submitted a first draft of Q1 of the answer and added one detail to the Q2 and Q3. He did one more round of changes to his first and second answer parts but none of his revisions resulted in changes in his predicted scores from the AES. He then submitted his final answers to his instructor who scored Q1, Q2, and Q3 with point values of 2, 2, and 1, respectively (Table 8).

Table 8: Overview of Draft Approach and Results for Mitch, Case 2

Draft Round	Responses Submitted and Nature of Edit	Automated Scores
1	• Q3 initial response	null
2	• Q3 elaborated upon how his recommendation met learners' diverse needs.	0
3	• Q2 initial response	0
4	• Q1 initial response • Q2 added name of school • Q3 added detail about a recommended software	1 0 0
5	• Q1 replaced one word • Q2 added an additional factor to consider, noting its relationship to other factors	1 0
		Instructor Scores
6	• **Q1 final answer** • **Q2 final answer** • **Q3 final answer**	**2** **2** **1**

Mitch indicated during his work on these two cases that he mostly relied upon the automated essay scores to develop his responses and used the rubric only "a little." He rated both the scorer and the rubric as being "a little useful" in his composition of responses. He was only "a little confident" of its accuracy and concluded it was only "a little accurate" of its accuracy. Before he completed the cases he had indicated that he was both skeptical and curious about it, and wanted to know how it worked. After he used it during two cases he had concluded that while the scorer had

some possibilities, that he hadn't relied upon it much since he didn't find it very responsive:

> ...if I got an 0 then I felt obligated to change my answer. But if I got a 1 then it was OK, and I was not so inclined to change my answer. ...A couple of times I tried to improve my responses to receive a higher score. This was usually not very successful and I didn't worry too much about it. It is somewhat difficult to know how to improve your response. Minuscule changes sometimes change the score while significant improvements would not.

While the log of Mitch's work did not in fact show any score improvements, he developed a positive impression of the scorer, even though when he made improvements to his answer his score did not necessarily improve.

Erica

Erica expressed a low opinion of the automated scorer and received high scores (2s) from the instructor. Erica does not think that computers should be used to assess writing that "is more than technical" because "you need to be thinking in order to assess writing...if you are grading on something that is more than pass/fail."

Across the two cases Erica only made one round of revisions. In her first case she drafted Q1 first and submitted it; then she drafted and submitted Q2 and Q3 at the same time for automated formative feedback (Table 9). While only one answer part received a score higher than 0, she nevertheless submitted these to her instructor as final answers. Her instructor assigned scores of 2, 2, and 1 for Q1, Q2, and Q3, respectively. He also provided very favorable comments back to her about her answers, noting that her responses were "excellent" and that she made "great recommendations."

Table 9: Overview of Draft Approach and Results for Erica, Case 1

Draft Round	Responses Submitted and Nature of Edit	Automated Scores
1	• Q1 initial response	0
2	• Q2 initial response • Q3 initial response	1 0
		Instructor Scores
3	• **Q1 final answer** • **Q2 final answer** • **Q3 final answer**	**2** **2** **1**

In the second case Erica availed herself of the opportunity to revise her work once, but only on part two of her answer (Table 10). She spent 21 minutes working on her answers. First, she drafted a response for Q1, Q2, and Q3 and submitted them together for a predicted score. She chose to revise Q2, the only part that scored a 0, and did so by adding two significant and lengthy points to it. Her instructor again felt she did a good job with her work, and assigned all three of her answer parts a score of 2 as well as writing favorable comments to her about her work.

Table 10: Overview of Draft Approach and Results for Erica, Case 2

Draft Round	Responses Submitted and Nature of Edit	Automated Scores
1	• Q1 initial response • Q2 initial response • Q3 initial response	1 0 1
2	• Q2 added how access impacts use, and how use impacts curriculum	0
		Instructor Scores
3	• **Q1 final answer** • **Q2 final answer** • **Q3 final answer**	**2** **2** **2**

On her post-survey Erica indicated that she used the automated essay scores and the rubrics when formulating her answers, but noted that the automated essay scores were "not helpful." She rated the rubric as most helpful but overall indicated that it only "somewhat impacted" her final responses. She explained that she thought that "the rubric criteria did not seem to match well with actual questions posed, so constructing my answers to fit both the rubric and the questions was a difficult and frustrating task." Her attitudes toward the automated essay scorer were quite negative. She rated it "not at all useful" and that she was "not at all confident" of its accuracy and that it was "not at all accurate." She marked that the scores she received on her drafted responses had "no impact" on her final response and added that "I felt confident that my answers were good. When the scorer told me differently, I did not change them."

In her interview she related how the demonstration of the scorer her instructor gave made her from the outset doubtful of its usefulness because he had only typed in a few words and received a prediction of a 1. She reasoned that since the high score was only one point higher that "this is not a person and it is a machine and it does not know what it is doing." She continued by explaining how it seemed that she recalled on case one having gotten a 1, 2, and 1 from the scorer and that she had felt Q2, the highest scoring part, was the weakest answer part she had given.

This "kind of freaked her out" and so she waited to submit her work until she could compare notes with her classmates about their score predictions and email her instructor about how much credence to give the scores. She felt assuaged by the fact that her classmates had gotten similar predictions and that her instructor reassured her that his scores would "count" in the grading of the assignment. She then concluded that she had "no faith in the system" and that "it would not be worth my time to beat the game."

She said that in the second case she did again submit her scores but because of curiosity more so than to consider it as a formative feedback. The series of events in the computer log contrasts a bit with her recollection, although the log does agree with her memory that her Q2 answer was scored highest. And, the log also shows that in case two she did revise a part of her answer after receiving a score prediction. Her overall opinion of this automated scorer was that "it did more harm than good" because if it was inaccurate it would make students draft more than they needed to, or if it overrated their responses it would encourage them to quit perhaps too early. She also shared that she doubted that any automated essay scorer was likely to be accurate enough to be helpful and that even if it were that she didn't think its score would be as meaningful as one received from a instructor.

Stacy

Stacy had a favorable opinion of the automated essay scorer and received high scores (2s) from the instructor on all six of her responses. Stacy believes that computers have a valuable role to play in aiding assessment, both in formative and summative fashions. Stacy shared that she is open to using computers for assessment, including writing assessment, as "a computer might actually be more fair [than a human scorer]."

Stacy did revisions to parts of her answers in both cases one and two. In case one, which she spent 65 minutes working on in two separate sittings, she initially submitted all three parts for predicted scores (see Table 11). Stacy then proceeded to revise and resubmit her Q1 answer another seven times. In these drafts she mostly added information to her answer. In the first few revision rounds she focused on the conclusion and added more detail by either using words from the principle or the challenge or case facts. In the last few revisions she deleted some statements and added some more details from the case. Across all these revisions her automated score remained a 0. So after 8 rounds of edits when she submitted her final answers to her instructor she did so with predicted scores of 0, 1, and 1 for Q1, Q2, and Q3, respectively. Her instructor assigned her scores of 2 on all three parts and his feedback to her was very positive for each of the three sections.

Table 11: Overview of Draft Approach and Results for Stacy, Case 1

Draft Round	Responses Submitted and Nature of Edit	Automated Scores
1	• Q1 initial response • Q2 initial response • Q3 initial response	0 1 1
2	• Q1 added to concluding sentence by referencing to the principle for the case	0
3	• Q1 added to the conclusion by referencing additional case and challenge detail	0
4	• Q1 included a reference to her students and how technology could help them express what they know and added more to the conclusion statement.	0
5	• Q1 deleted a statement about what she would recommend and added a statement referencing case information about students' skills	0
6	• Q1 added a reference to facts in the case's introduction	0
7	• Q1 deleted a summary statement referencing her students	0
8	• Q1 added a statistic about students' achievement	0
		Instructor Scores
9	• **Q1 final answer** • **Q2 final answer** • **Q3 final answer**	**2** **2** **2**

In case two Stacy took a much more direct approach and spent far less time, a total of 12 minutes. She submitted initial responses for all parts of the answer and received predicted scores of 1, 1, and 0 for Q1, Q2, and Q3, respectively (Table 12, next page). She did one further edit of Q3, adding a couple of reference to software by name. Her instructor scored all three of her answer parts with the highest score, a 2.

Table 12: Overview of Draft Approach and Results for Stacy, Case 2

Draft Round	Responses Submitted and Nature of Edit	Automated Scores
1	• Q1 initial response • Q2 initial response • Q3 initial response	1 1 0
2	• Q3 added examples of some software tools.	0
		Instructor Scores
3	• **Q1 final answer** • **Q2 final answer** • **Q3 final answer**	**2** **2** **2**

Stacy indicated on her post-assignment survey that, in order of importance, she drew upon her notes, the provided rubric, and the automated essay scores when she formulated her answer. Stacy rated the automated essay scorer as "somewhat useful" and while she had only been "somewhat confident" of its accuracy, but that its predictions "impacted a lot" her final responses to the case challenges. In her interview she explained how when her instructor explained the automated essay score her initial reaction was that she was perplexed and wondered how it could work. During the first case she said it brought out different emotions, including animosity and competitiveness:

> I hated it. [laugh] I became a human robot. By the end, I did not care anymore; it was "OK I don't care what you are telling me." I just trusted my own intuition about this, by the end…it kept giving me a 1 and I would rearrange what I said, I would add more data, I would put more case characteristics on that first question….Of course I wanted to [beat the scorer]. Nobody is in this program unless they are competitive [laugh].

Stacy went on to say that she thought the scorer could be helpful, particularly if instead of just a score it would provide "hints of missing information." She reasoned that if it was programmed to see certain key words that it should prompt the students things like "What about demographics? or What about computer technology available?…if I were a teacher giving formative feedback I would give that sort of information. Give the students a clue as to where they could start looking and working as opposed to just a 1 or a 2."

Discussion

Although using an automated essay scorer represented a new type of software experience for these pre-service teachers, the students did not hesitate to use and experiment with the AES as a means of improving their essays—despite not being required to use it. Increased use of the AES was associated with better essays as measured by their final instructor-assigned essay scores. This outcome boosts confidence in the potential utility of automated essay scoring software to aid in formative assessment of writing. While simply making an AES accessible for purposes of formative feedback might result in more drafting and improvements in written work, it does not address the unexpected outcome of this study, which was the way the AES evoked an emotional response from these students.

During the interviews the students were rather animated in describing their experiences with the scorer. While we asked them what the scorer made them *do* in regard to their essay drafting, the students instead tended to focus the conversation on how the AES made them *feel*. By and large, the students used the automated scorer extensively in the first case because they found it engaging as a tool, and also because they hoped to improve their work—they wanted formative feedback on their writing. However, as the interviewed students described, when they believed that the scorer was not accurately measuring their essay improvements they became quite frustrated with it; and so the use of the automated scorer declined during the completion of the second case. Thus, the results of this study point to the importance of the students' subjective experience with automated essay scorers.

The experiences of the four interviewed students illustrate how the inaccuracy of the scorer produced most of the negative emotions students experienced. The data in this research show that the ETIPS automated scorer fairly consistently returned inaccurate predictions of instructor scores. Students' experiences with the scorer were further impacted when they accessed their instructors' scores on their essays and compared them to the automated scores, which were typically one point lower than the instructor's scores. Obviously, this discrepancy in scores only led to heightened frustration with the scorer. While the low automated predictions did, in some cases, prompt students to revise their essays, it also provoked some negative, emotional responses; when automated scores did not improve, students seemed to lose confidence in the automated scorer's ability to give them helpful formative feedback.

The detailed analysis of the four students' series of essay drafts illustrated how the changes they made to their essays were positive ones in terms of writing structure. Most changes were to provide additional examples or details as elaborations on their points, or to reference either the

case's key idea or the principle the case was designed to provide practice thinking about. However, using the provided rubric, the first two authors scored the first draft students submitted and then also scored their successive drafts and found that their scores did not improve, even though in the majority of cases there was room for them to do so; that is, their work was not yet judged to be at the highest level on the rubric, a score of 2. This outcome suggests that the nature of formative feedback an AES supplies is just as important as the accuracy of its predicted scores. The four interviewed students expressed they did not feel guided by the automated feedback as to how to actually improve their essays in response to the case challenges. An automated predicted score of 0, 1, or 2 evidently did not, in combination with the rubric to which these scores refer and a short statement about the qualities of a "good" answer, provide enough guidance to students as to how to improve their essays. In other words, the nature of the formative feedback given to these students by the ETIPS scorer was not sophisticated enough for them to know what specific sort of revision to make to their answers.

Implications

These in-depth data about students' reactions to and opinions of the current ETIPS automated essay scorer provide insight as to factors that should be considered in using computer-based, automated essay scorers to provide feedback on student writing. In any AES used to provide assessment scores, its design, nature, accuracy, and the specificity of the feedback it can provide are all very important in aiding students' better performances on and positive experiences with the AES in support of their learning. Other researchers recognize the attention developers of AESs must give to the ability of that software to provide specific feedback if learners are to grow in their writing skills as a result of its use. In their discussion of the widely used e-rater AES, Burstein, Marcu, Andreyev and Chodorow (2001) note that the feedback must be adapted to the writer and the writing task:

> Unfortunately, providing students with just a score (grade) is insufficient for instruction. To help students improve writing skills, writing evaluation systems need to provide feedback that is specific to each individual's writing and that is applicable to essay revision (p.1).

Therefore, in addition to providing accurate, reliable, and valid automated scores to students on their essays, which is where most of the AES research is focused, ETIPS and other AESs designed to improve writing through automated feedback should consult and explore the literature on the purposes for and function of response to student writing as well as

research on writing revision. Also, students using an AES for a specific task might also be consulted for the type of feedback to design into the AES. For example, one of the four interviewed students (Stacy) made a suggestion as how to improve the feedback given by the ETIPS automated scorer. She suggested that automated, predicted scores could easily be combined with the ability of the ETIPS software to track which case information students accessed and return to students hints about specific information they should access in the case and relate to the case questions.

Now that the technologies behind making AES possible are stable, more attention needs to be brought to users' experiences with AESs. Indeed, design-based development and research, an iterative methodological approach aimed at enhancing learning and teaching processes by means of theory development, research in authentic and naturalistic environments, and the sharing of knowledge amongst practitioners and researchers (The Design-Based Research Collective, 2003), should be incorporated more systematically within the field of automated essay scoring. A limitation of this paper is that it offers a single snapshot of students' experiences with an AES in an educational, online context. In order to more fully inform the AES field, more research is needed in a myriad of contexts regarding both the optimal design and functionality and users' experiences with AESs.

References

Barron, B., Schwartz, D., Vye, N., Moore, A., Petrosino, A., Zech, L., Bransford, J. (1998). Doing with understanding: Lessons from research on problem- and project-based learning. *Journal of the Learning Sciences, 7*(3), 271–311.

Black, P., & William, D. (1998a). Assessment and classroom learning. *Assessment in Education, 5*(1), 7–74.

Black, P. & William, D. (1998b, October). Inside the black box: Raising standards through classroom assessment. *Phi Delta Kappan, 80*(2), 139–148.

Burstein, J. (2003). The e-rater scoring engine: Automated essay scoring with natural language processing. In M. D. Shermis & J. Burstein (Eds.), *Automated essay scoring: A cross-disciplinary perspective* (chap. 7) (pp. 113–121). Mahwah, New Jersey: Lawerence Erlbaum Associates.

Burstein, J., Marcu, D., Andreyev, S., & Chodorow, M. (2001, July). Towards automatic classification of discourse elements in essays. Proceedings of the 39th Annual Meeting of the Association for Computational Linguistics.

The Design-Based Research Collective. (2003). Design-based research: An emerging paradigm for educational inquiry. *Educational Researcher, 32*(1), 5–8.

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.

Dikli, S. (2006). An overview of automated scoring of essays. *Journal of Technology, Learning, and Assessment, 5*(1). Retrieved August 18, 2006 from http://www.jtla.org.

Glaser, B.G., & Strauss, A. (1967). *The discovery of grounded theory: Strategies for qualitative research*. Chicago: Aldine.

Grimes, D. & Warschauer, M. (2006). Automated essay scoring in the classroom. Paper presented at the Annual Meeting of the American Educational Research Association, San Francisco, CA.

Marzano, R.J. & Pickering, D.J. (1997). *Dimensions of learning: Teacher's manual*. Colarado: McREL.

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.

Page, E.B., & Peterson, N.S. (1995). The computer moves into essay grading: Updating the ancient test. *Phi Delta Kappan, 77*, 561–565.

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.

Pelligrino, J., Chudowsky, N., & Glaser, R. (2001). *Knowing what students know*. Washington D.C.: National Academy Press.

Riedel, E., Dexter, S., Scharber, C., & Doering, A. (2006). Experimental evidence on the effectiveness of automated essay scoring in teacher education cases. *Journal of Educational Computing Research. 35*(3), 267–287.

Sadler, D.R. (1989). Formative Assessment and the Design of Instructional Systems. *Instructional Science, 18*, 119–144.

Scharber, C., Dexter, S., & Riedel, E. (2005, April). *Formative feedback via an automated essay scorer: Its impact on learners*. Paper session for the annual meeting of the American Educational Research Association, Montreal, Québec, Canada.

Shermis, M.D., & Burnstein, J.C. (2002). *Automated essay scoring, a cross-disciplinary perspective*. New-Jersey: Lawrence Erlbaum Associates.

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

Warschauer, M., & Ware, P. (2006). Automated writing evaluation: Defining the classroom research agenda. *Language Teaching Research 10*(2), 1–24.

Appendix A
Case Introduction

Imagine that you are mid-way through your first year as a *seventh* grade teacher at *Cold Springs Middle School*, in an *urban* location. A responsibility of all teachers is to differentiate their lessons and instruction in order to accommodate for the varying learning styles, abilities, and needs of students in their classrooms and to foster students' critical and creative thinking skills.

As a new teacher at *Cold Springs Middle School*, you will be observed periodically throughout the first few years of your career. One of the focuses of these observations is to analyze how well your instructional approaches are accommodating students' needs. The principal, *Dr. Kranz*, was pleased with your first observation. For your next observation she challenged you to **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**. She will look for your technology integration efforts during your next observation.

Examine the school web pages to find the information you need about both the context of the school and your classroom in order to address the challenge presented above. When you are ready to respond to the challenge, click "submit answer". On the case's answer page, you will be asked to address this challenge by making three responses:

1. Confirm the challenge:

   What is the central technology integration challenge in regard to student characteristics and needs present within your classroom?

2. Identify evidence to consider:

   What case information must be considered in a making a decision about using technology to meet your learners' diverse needs?

3. Submit your justified recommendation:

   What recommendation can you make for implementing a viable classroom option to address this challenge?

Note: Items in italics change with differing school contexts.

Appendix B
Rubric for ETIPS 2

ETIPS 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

Appendix C
Instructor Scoring Rubric

ETIPS: Scoring Rubric for Student Response

ETIPS 2 Primary Challenge

How can technology 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?

Outcome Expectation for ETIPS 2

Using his/her own words in two to six sentences, the credential candidate's essay clearly demonstrates his/her decision-making process about using technology to foster the critical and creative thinking skills of the diverse group of students in the case.

Criteria	Level 0 = Score 0	Level 1 = Score 1	Level 2 = Score 2
Validation			
Explains the *central technology integration challenge* in terms of case characteristics	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
Evidence			
Identifies *case information* that must be considered in a decision about using technology to differentiate instruction to meet the diverse needs of learners	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
Decision			
States 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, particular technology to differentiate instruction to meet the diverse needs of learners

Appendix D
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: _____

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).					

* 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.

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.					

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 XXXX?

_____ Yes _____ No

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 E
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: _____

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.

* 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.

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 Cases 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	Did not use

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?

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?

End of survey. Thank You!

Appendix F
Interview Questions

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

Name of Interviewee: _____

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.

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 the 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? When you 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 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?

Author Biographies

Cassandra Scharber will be an assistant professor of Learning Technologies in the Department of Curriculum and Instruction at the University of Minnesota. A former English teacher, her research interests include K–12 technology integration, digital equity, and digital literacies. She has expertise in developing online curricula for both K–12 students and teachers. Cassie has coauthored several book chapters on technology integration and has published in journals such as *The Journal of Educational Computing Research* and *The Journal of Adolescent and Adult Literacy*. Cassandra Scharber can be reached at scharber@umn.edu.

Sara Dexter is an assistant professor of technology leadership in the Curry School of Education at the University of Virginia. She has also been a junior high and high school science teacher as well as a district staff developer specializing in educational technology. Dr. Dexter studies effective learning environments for educators about technology integration and implementation through K–12 school and online case-based research. She has been the principal investigator on several sponsored projects, including ETIPS cases and Ed-U-Tech, and the Exemplary Technology-Supported-Schooling Case Studies Project. These are described in further detail and linked to at her website: http://sdexter.net. Dexter has served as Chair and Program Chair of the AERA SIG: Computer and Internet Applications in Education; she is a member of the editorial review board of the Journal of Computing in Teacher Education. Sara Dexter can be reached at sdexter@virginia.edu.

Eric Riedel currently serves as the Executive Director for the Office of Institutional Research and Assessment at Walden University and as team leader for the university's participation in the Higher Learning Commission's Academy for the Assessment of Student Learning. Prior to coming to Walden, Dr. Riedel served as an educational evaluator at the Center for Applied Research and Educational Improvement at the University of Minnesota. Dr. Riedel has participated in research and published in the areas of civic education, the integration of technology into assessment, and the role of social capital in online social interaction. He received a doctorate in political science from the University of Minnesota and a BA from the University of Michigan. Eric Riedel can be reached at eric.riedel@waldenu.edu.

J·T·L·A

The Journal of Technology, Learning, and Assessment

www.jtla.org

Technology and Assessment Study Collaborative
Caroline A. & Peter S. Lynch School of Education, Boston College