Running Head: Automated Essay Scores

# Automated Essay Score Predictions as Formative Assessment Tools

Cassie Scharber
Center for Applied Research and Educational Improvement
University of Minnesota

Sara Dexter
Department of Curriculum and Instruction
University of Nevada, Las Vegas

Contact/Correspondence: Cassie Scharber
Mailing Address: 270 Peik Hall
159 Pillsbury Drive
Mpls, MN 55455
Ph: 612 626 7261 Fax: 612-625-3086
e-mail: scha0809@umn.edu

3/3/04

Paper session for the fifteenth international conference of the
*Society for Information Technology and Teacher Education*, Atlanta, Georgia

## Introduction

In this paper we present using automated essay scoring on network-based learning objects developed through a 1999 Implementation and a 2001 Catalyst grant from the U.S. Department of Education's Preparing Tomorrow's Teachers to Use Technology (PT3) program. These learning objects are the ETIPS (Educational Theory Into Practice Software) cases, and they were designed to develop preservice teachers' instructional decision-making skills about technology integration and implementation. In particular, the automated essay-scoring feature of the ETIPS cases is discussed in terms of users' experiences and its potential as a formative assessment practice facilitating the learner making connections between theory and practice

Automated essay scoring is a new feature of the ETIPS cases designed to capitalize on the capabilities of technology to aid assessment. It is employed as a formative feedback mechanism instead of its more typical use as summative feedback tool. The implication of a reliable automated essay scoring tool in use within network-based online learning environments is that it makes possible formative feedback, and presumably improved performances, on complex sets of skills.

Before it can be determined if the new assessment feature of the ETIPS cases is an effective means of formative feedback, students' initial responses to the automated essay scorer need to be documented in order to help better inform the ETIPS cases project about students' reactions to the assessment feature; suggest future implementation methods for ETIPS cases' instructors for using the automated essay scorer as a means of formative feedback for students; and provide insight as to the reliability of the current essay scorer so that changes can be made to enhance its effectiveness and/or usefulness.

## Literature Review

### ETIPS CASES

The Educational Technology Integration and Implementation Principles, or ETIPS, were developed by Sara Dexter and provide the conceptual framework for the cases. Correlated with the NETS-T and INTASC standards, and anchored in the research surrounding teaching, learning, and assessment with technology, the ETIPS are offered as an explanation for the conditions that are essential for the effective use of educational technology in the classroom (Dexter, 2002).

The six eTIPs are organized into two groups; classroom and schoolwide. The principles are an elaboration of two premises: 1) the teacher must act as an instructional designer, planning for the use of the technology so it will support student learning; and 2) the school environment must support teachers in this role by providing adequate technology support. The first three ETIPS are focused at the *classroom-level*, emphasizing the role of teacher as instructional designer. Technology does not possess inherent value; its value is designed into the educational environment by teachers: eTIP 1: Learning outcomes drive the selection of technology; eTIP 2: Technology provides added value to teaching and learning; eTIP 3:Technology assists in the assessment of learning outcomes.

The three remaining eTIPs are focused at the *school-level*, emphasizing the supportive conditions necessary for teachers' integration efforts. The level of technology support at a school is what makes teachers' technology integration feasible: eTIP 4: Ready access to supported, managed technology is provided; eTIP 5: Professional development targets successful technology integration; eTIP 6: Teachers reflect on, discuss, and provide feedback on educational technology.

Not only are the eTIPs themselves standards-based, the ETIPS cases are anchored in the research of teaching, learning, and assessment. The collective research of two National Academy of Sciences reports, *How People Learn: Brain, Mind, Experience and School* (1999) and *Knowing What Students Know: The Science and Design of Educational Assessment* (2001), document that an essential component of effective learning environments is that they are assessment-centered. The ETIPS Cases employ several assessment features, including a new assessment feature -- an automated essay scorer.

## Effective Learning Environments

The National Academy of Sciences report, *How People Learn: Brain, Mind, Experience and School*, suggests that learning environments need to be, among other things, assessment-centered (Bransford, Brown, & Cocking, 1999). Effective assessment-centered learning environments provide multiple opportunities to make students' thinking visible so learners can receive feedback and be given chances to revise and learn about their own learning. Consequently, the report notes that what are needed are formative assessments, which provide students with these opportunities to revise and improve the quality of their thinking and understanding (p. 24).

The National Academy of Sciences report, *Knowing What Students Know: The Science and Design of Educational Assessment*, addresses principles of assessment design for learning situations that are aligned with the findings of *How People Learn* (Pellegrino, Chudowsky, & Glaser, 2001). Pellegrino et al. (2001) write that networks (defined as an integration of the Internet and intranets offering learning objects, resources, and new forms of instruction and assessment), new media, and new methodologies have much to offer us in the way of support and enhancement of assessment. The authors outline several advances in the science of assessing thinking and learning that can be gained through network-based assessment systems. Also, network-based assessments can build a long-term record of documentation, showing how learners change over time. In addition, network-based assessments can include statistical analysis and displays of information to assist learners and teachers in making inferences about performance. This synthesis of recent research on assessment systems in *Knowing What Students Know: The Science and Design of Educational Assessment* is used as a loose set of criteria to guide the continual development of the ETIPS cases, a network-based assessment system, and its assessment features.

## Formative Assessment

For the most part, assessment that takes place in classrooms is summative, meaning that it happens at the "end" of a unit, lesson, etc., illustrating what students know and do not know at the conclusion of an education experience. Students rarely experience formative assessment or feedback that is integrated as part of, rather than at the end of, the learning experience. However, there is a shift occurring in assessment paradigms; in the past the focus has been on summative assessment whereas, "there is now a realization that the potential benefits of assessing are much wider and impinge on all stages of the learning process" (Dochy & McDowell, 1997, p. 279). Birenbaum (1996) has called this a shift from a "testing culture" to an "assessment culture", with the new assessment culture emphasizing the integration of assessment and learning.

The literature makes a solid case for the importance of formative assessment as it has been proven to produce specific gains for learning (Barron, et al., 1998; Black and William, 1998). Black and William (1998) argue "formative assessment is an essential component of classroom work and that its development can raise standards of achievement" (p. 148) and that formative assessment is "at the heart of effective teaching" (p. 140). Additionally, the Center for Innovative Learning Technologies argues for the importance of embedded ongoing, formative assessments that support the teaching and learning process rather than high-stakes assessments that come after one is supposed to have learned something (Ravitz, 2002, p. 295).

## **Computer-Based Formative Assessment**

Technology holds much promise for contributing to the increased practice of formative assessment in education. Ravitz (2002) believes that "one way technology can be a substantial help to teachers and learners is by improving the ability to offer formative assessments of a learner's knowledge and skills…" (p. 293). There is mounting evidence to suggest that formative computer-based assessments can produce improvement in student learning (Clariana, 1993; Peat & Franklin, 2002). Furthermore, there are pedagogical advantages of computer-based assessment in providing feedback on student work (Charman, 1999). The pedagogical advantages of computer-based formative assessment include repeatability, immediate feedback to students, immediate marks to staff, reliability and equitability, diversity of assessment, content vs. presentation, timeliness, flexibility of access, student interest and motivation, student-centered skills and learning (Charman, 1999, pp. 87-90).

Literature on students' responses to computer-based formative assessment is quite limited. A study conducted by Sambell, K., Sambell, A., & Sexton (1999) considers students perceptions of learning benefits of computer-assisted assessment in the context of undergraduate engineering courses. In this study, generic assessment software was used to provide both summative and formative feedback to students. Similar to the ETIPS cases, the formative measures were voluntary, not timed, and could be repeated as often as students chose. Although the assessments in this study were not essay-based in format, student response to the formative feedback was positive, and encouraged the staff to continue to develop such materials as they "promised considerable learning benefits (Sambell, et al., 1999, pp. 188-189). Alexander, Truell, & Bartlett II (2002) report that students' perceptions of online testing were positive in a survey they administered to 200+ students in a business-information systems classes; however, these students used online testing in the classes for summative purposes rather than formative.

Research on the impact of computer-based formative assessment on student learning and performance is also scarce in the literature. Two cases studies that explore the impact of computer-based formative assessment include a micropaleontology course in Liverpool (Boyle et al., 1997) and a statistics course in geography (Charman & Elmes, 1998). Both of the studies concluded that student learning is positively impacted by computer-based formative assessment. These results need to be supported with additional research.

In response to the literature surrounding effective learning environments and the call for more research on formative assessment in *How People Learn* (Bransford, et al., 1999, p. 257), the ETIPS cases offer a new formative assessment tool, an automated essay scorer. The feedback given by the automated scorer in the cases is formative in nature as students have the option of submitting their responses to the challenge posed in the case to an automated essay scorer before submitting final responses to their instructors for grading. In doing so, students receive estimated scores on their responses and can chose to rework their responses or submit their answers to their instructors.

Computer-based writing evaluation research has been going on since the mid 1960s. Since then, numerous systems and approaches have been developed to measure writing quality, which have high correlations with human scorers. 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 being used for scoring the General Management Aptitude Test (GMAT) essays. The ETIPS Cases' automated essay scorer is similar in methodology to the e-Rater system.

## Methods and Data

Students enrolled in a post-baccalaureate program at the University of Minnesota – Twin Cities seeking teaching licensure in secondary English/Language Arts 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. During this course, all students completed two ETIPS cases as part of the course requirements. Thirty-four students were enrolled in the course; twenty-seven of these students completed both pre- and post assignment surveys and agreed to participate in this study.

Study participants were given an attitudinal/demographic survey to complete before they began the ETIPS cases' assignment. Then, over the course of two weeks, students worked through two ETIPS cases. ETIPS cases differ from traditional cases in that they are not linear; instead, each case is broken down into parts identified as website sections/pages, which forces students to select information categorically rather than linearly. Each ETIPS "case" consists of an introduction that frames the challenge, a virtual school's website, and a response page. On the response page, the challenge is broken down into three smaller parts, which students respond to. Students in the course were presented with the same challenge for both cases: "The principal was pleased with your first observation. For your next observation he/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". Students responded to this challenge by answering three questions at the end of the case on the response page:

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) *State your justified recommendation:* What recommendation can you make for implementing a viable classroom option to address this challenge?

Although the challenge was the same for both cases, the context (the school website) changed per case; for this study, the first case took place in a hypothetical, yet realistic school called Cold Springs Middle School and the second case took place in Stromburg High School.

Each time students completed a case, they had opportunities to 1) complete multiple drafts of the three small answer parts that comprised the case challenge; 2) receive automated essay scores predicting how a human might score their responses (formative feedback); 3) return to the cases to gather more information to inform revised responses; and 4) submit their final responses to receive feedback and scores from their instructor. Also during each case, students had online access to the rubric their instructor used to evaluate their responses as well as a visual display of their actual search of a case (Search-Path Map). While students worked on the assignment, the ETIPS cases software collected data on individual students' case use using a log file that tracked the pages students viewed inside the cases, recorded all drafts students submitted for automated feedback and the automated scores, and documented the instructor's scores on their final answers.

After the assignment was completed, students were given a second survey to complete regarding their experiences with and their opinions about the ETIPS cases and the formative feedback they received from the automated scorer during the cases. Based on data collected from both surveys and the software, five students were selected and invited to participate in individual interviews to further detail their individual experiences with the ETIPS cases and its automated essay scorer.

Thirteen of the twenty-seven participating students volunteered to be considered for the interview portion of the study. The thirteen possible interviewees were sorted into eight possible typologies (see Table 1) developed 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 did per response section for both cases. Five of the eight typology categories were populated (Typologies A,B,C,D,E,F), and from each populated category one student was chosen to be interviewed. One of the populated categories, Typology B, contained only one student, who initially agreed to be interviewed, but then declined. So, four interviews were individually conducted, taped, and transcribed which represent four of the five typologies found in for this study.

Table 1
Student Typology

| **Typology A** | **Typology B** | **Typology C** | **Typology C** |
|---|---|---|---|
| Positive opinion of scorer<br>Received scores of 2<br>2+ drafts per response | Positive opinion of scorer<br>Received scores of 2<br>0-1 drafts per response | Positive opinion of scorer<br>Received scores of 0-1<br>2+ drafts per response | Positive opinion of scorer<br>Received scores of 0-1<br>0-1 drafts per response |
| *5 students* | *1 student** | *2 students* | *(none)* |
| **Typology D** | **Typology E** | **Typology F** | **Typology G** |
| Negative opinion of scorer<br>Received scores of 2<br>2+ drafts per response | Negative opinion of scorer<br>Received scores of 0-1<br>2+ drafts per response | Negative opinion of scorer<br>Received scores of 2<br>0-1 drafts per response | Negative opinion of scorer<br>Received scores of 0-1<br>0-1 drafts per response |
| *(none)* | *3 students* | *2 students* | *(none)* |

*Student declined interview

Interviewees were asked six questions:

1) If you could describe your experiences with the ETIPS cases in one word, what would that word be? Explain your response.
2) In regard to your assignment, explain how you approached responding to the questions asked of you at the end of each case. Were there any differences in how you approached the second case from the first case?
3) Tell me about your response(s) or reaction(s) to the automated scoring feature of the ETIPS cases.
4) The automated essay scorer is a new assessment feature of the cases. Its purpose is to provide formative feedback to users to help them craft "better" responses. In your opinion, is the automated essay scorer an effective means of formative assessment?
5) Do you have any ideas about other ways the ETIPS cases could use to supply users with formative feedback during a case?
6) How do you feel about computers being used to assess learning? Writing?

## Results

### Adam

Adam represented the group of students who expressed a somewhat low opinion of the automated scorer, received lower scores from the instructor, but did multiple drafts in at least one case. On his survey Adam indicated he was an A student, and that he had taken three college-level courses in writing and writing assessment either outside of or within his teacher licensure program. He also responded that he was very confident in using computers to do email and the Internet, that he was comfortable with using computers and agreed that students should learn about how to use computers and that they could instruct and assess students.

Adam, who had a bachelor's degree in English, said in his interview that while he thought technology could aid formative assessment of students' written work it would require first that they understand how the technology tool worked, and its 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." Overall, he indicated that he found the assigned cases "useful" tools for learning about technology use in education. He said he learned that he learned there were more challenges to incorporating technology than just planning it or not; he added an insight about the specificity required of teachers to consider technology-integrated lessons.

In each of the two cases Adam was to draft three short responses. Across these six opportunities to receive formative feedback on these answers from the automated essay scorer Adam did so three times, all in the first case (see table 1). He spent 29 minutes writing and revising answers for case one. He began by drafting an answer for each of the three answer parts for case one and submitting it for an automated score; for each of these responses the prediction was a 0. He revised part one again by taking away a summary statement at the end of his answer, but again his predicted score was a 0. The log shows that in his third round of revisions he made changes to all three answer parts, adding either examples or a summary statement. While the nature of the changes were that he used key words from the rubric and or case question, his predicted score remained a zero. In round four of revisions he focused on his part one answer. He revised its wording completely, although its ideas were similar to the previous draft; for it his predicted score remained a zero. In his fifth and final round of revisions he once again revised part one by adding an explanatory and summary statement and also added to his part two answer some specific examples from the case and phrases from the rubric. His predicted score for part one remained a 0 but his part two predicted score went up to a 1. In summary, Adam made a total of five rounds of revisions, focusing mostly on the part one case answer. Only in one round of revisions did his predicted score improve. He submitted his final answer with predictions of 0, 1, 0 for parts one, two, and three, respectively. The automated scores did not match up with the instructor assigned scores on any of the three answer parts. Scoring against the rubric, his instructor assigned a 1, 0, 1, respectively. The comments the instructor provided and the rubric criteria indicate that he did not think Adam provided enough specific case information in his answers, while acknowledging that Adam was, in general, addressing the topic of the case.

Table 1
Overview of Drafts and Results for Adam, Case 1

| **Draft Round** | **Essay Parts Submitted and Nature of edit** | **Scores** |
|---|---|---|
| 1 | C1 initial response<br>C2 initial response<br>C3 initial response | 0<br>0<br>0 |
| 2 | • C1 Removed a summary statement | 0 |
| 3 | • C1 Added an example, but not from case information<br>• C2 Added an example, but not from case information<br>• C3 Added a summary statement with key words from the rubric and case question | 0<br>0<br>0 |
| 4 | • C1 Completed reworded answer, while making same points | 0 |
| 5 | • C1 Added an explanatory and summary statement, using case information<br>• C2 Added an example from the case | 0<br>1 |
| 6 | C1 final answer<br>C2 final answer<br>C3 final answer | 1 (Human<br>1 assigned<br>1 scores) |

Adam overall approached case two quite differently than case one (see table 2). He requested an automated predicted score only once but did not revise those answers. He first wrote up answer parts one and two and submitted them for feedback. His automated score predictions for these two parts were 1's. He then wrote a response for part three and submitted it, and received a predicted score of a 0. He then made no revisions to any of the three parts and submitted his responses as his final answers. In summary, Adam took a streamlined approach to the second case. In his interview he described it as a more focused approach, and that he did not try to make the essay scorer predict a high score for him. For the answers he submitted he settled for predictions of 1, 1, and 0 on parts one, two and three of the case, respectively. Scoring against the rubric, his instructor assigned a 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 part one, and that his part three answer was "creative", but that he did not score a 2 because it still lacked discussion of a particular idea.

Table 2
Overview of Drafts and Results for Adam, Case 2

| Draft Round | Essay Parts Submitted and Nature of edit | Scores |
|---|---|---|
| 1 | C1 initial response<br>C2 initial response | 1<br>1 |
| 2 | C1 initial response | 0 |
| 3 | C1 final answer<br>C2 final answer<br>C3 final answer | 1 (Human<br>1 assigned<br>1 scores) |

During his work formulating an answer for each case Adam indicated that he drew upon the rubric, the search-path map, and the automated essay scores, in relative order of importance. On a questionnaire he rated the automated scorer as only “a little useful” and marked that he was “not at all confident” of it and that it was “not at all accurate.” His assessment of the tool warranted given that it its prediction did not ever match his instructor 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. He 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 it was the human’s score that would be recorded. His experience with the essay scorer in the first case quickly made him frustrated when it did not return to him predictions of high scores. This triggered his competitive spirit such that his efforts to “beat” the automated scorer distracted him from focusing on the case, its question, and the overall purpose of the assignment. He described how he took on the attitude that it 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 did not receive high scores predictions either. He concluded then he would not let the score predictions bother him and recognized he should “just let it go.” He described that after he did so, his experience with the second case was more enjoyable. In the second case he said he explored nearly every menu item and focused more on the case specifics.

## **Mitch**

Mitch is representative of the group of students who had a positive opinion of the scorer, received lower scores on average (0s and/or 1s) from the instructor, and completed multiple drafts (two or more) for most of his responses. He is comfortable working with computers, but expressed that he does not have “a lot of self-confidence” when it comes to working with computers; he is confident sending email and using the internet, but is not as confident in using spreadsheets or saving documents as different formats. Mitch has taken one course on writing mechanics and/or assessment and is an “A” student. On the pre-assignment survey, Mitch shared that he does not think that computers can help students improve their writing or aid teachers in assessment of students. He also believes that assessment should take place both during and at the end of a project.

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 “ it encourages you to craft answers…I think the scorer has potential as long as it is accurate”. Overall, Mitch expressed that he found his experience with the ETIPS cases assignment “frustrating” but “a little useful”. He said, “there seemed to be a disconnect between the challenge and the information that was available…I felt like I should be answering the questions as a technology expert rather than a teacher.” He learned that “just having technology available is not adequate” when considering technology integration.

Across the two cases’ three short responses, Mitch took three opportunities to revise his answers in response to formative feedback; all of these were in the second case. He described his approach moving from the cases’ homepage to the links he was interested in, then submitting part one after which he returned to the case to look for more information to answer part two, and then repeating this process for part three. This took him about 49 minutes. The log shows that for case one Mitch submitted his answer part one, received an automated predicted score of 0, then submitted his answer part two, and received a predicted score of 0; he then submitted part three and again received an predicted score of 0. Even with predictions of 0 for all three answer parts he submitted them to his instructor as his final responses. His instructor scored his responses as a 1, 1, and 0 for parts one, two, and three (see table 1).

Table 1
Overview of draft approach and results for Mitch, Case 1

| **Draft Round** | **Essay parts submitted and Nature of edit** | **Scores** |
|---|---|---|
| 1 | C1 initial response | 0 |
| 2 | C2 initial response | 0 |
| 3 | C3 initial response | 0 |
| 4 | C1 final answer<br>C2 final answer<br>C3 final answer | 1 (Human<br>1 assigned<br>0 scores) |

Mitch took a different approach to case two, revising two of answer parts twice and one part he revised once. He took a total of 40 minutes with this 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 (part three of the answer). He revised this once after the automated scoring software returned a “null” result, meaning that it didn’t 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 received a predicted score of 0. He then drafted and submitted part two of the answer and received a predicted score of 0. In round four he not only submitted a first draft of part one of the answer, he added one detail to the second and third parts of his answer with the result being a predicted score of 1 on part one, and the other two parts remaining a predicted score of 0. He did one more round of changes to his first and second answer parts but his scores did not change. He submitted his final answers to his instructor with predicted scores of 1, 0, and 0 respectively, for parts one, two, and three. His instructor scored these parts with point values of 2, 2, and 1.

Table 2
Overview of draft approach and results for Mitch, Case 2

| **Draft Round** | **Essay parts submitted and Nature of edit** | **Scores** |
|---|---|---|
| 1 | C3 initial response | null |
| 2 | • C3 elaborated upon how recommendation met learners' diverse needs. | 0 |
| 3 | C2 initial response | 0 |
| 4 | C1 initial response<br>• C2 added name of school<br>• C3 added detail about a recommended software | 1<br>0<br>0 |
| 5 | • C1 replaced one word<br>• C2 added an additional factor to consider, noting its relationship to other factors | 1<br>0 |
| 6 | C1 final answer<br>C2 final answer<br>C3 final answer | 2 (Human<br>2 assigned<br>1 scores) |

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 didn't necessarily improve.

## Erica

Erica represents the group of students who expressed a low opinion of the automated scorer, received high scores (2s) from the instructor, and averaged one draft per response. An "A" student, Erica highly rated her confidence in her computer skills and her comfort in using computers. On her pre-assignment survey, Erica "strongly agreed" that computers are valuable in "stimulating creativity in students," "helping students improve their writing," and "aiding in assessment"; she also believes that students should learn about computers in school. Erica has taken three courses on writing mechanics and/or assessment, and strongly agrees that assessment [in general] should be formative rather than summative. Despite her knowledge and comfort of computers, 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.”

“Frustrating” was the word Erica used to describe her experiences with the assignment, mainly because she feels like she “spends so much time as a student in front of a computer.” So, for her approach to the assignment, she copied and pasted the information from every webpage to a word document that she printed off so she “could take the print outs somewhere else, like the living room.” Overall, Erica found the ETIP cases “not at all useful”, noting on her post-assignment survey that she felt like she “had knowledge [of technology integration] before the ETIPS and the assignment was more like busy work.” However, in her interview a week after the assignment, Erica stated, “I feel bad because on the survey I know that I wrote, ‘did you find this to be a valuable learning experience”’ and I think I put ‘not at all’. Well, to be honest, I did have the chance to think about if I were a teacher at this school, how would I go about finding how to use technology or whether or not the resources are there.”

Across the two cases Erica only made one round of revisions. In her first case she drafted answer part 1 first and submitted it and received a predicted score of a 0. Next she drafted and submitted parts 2 and 3 at the same time for automated essay scores. She received a 1, and 0, respectively, on parts 2, and 3 (see table 1). Her instructor assigned higher scores on all three parts, a 2, 2, and 1, respectively. He also made very favorable comments back to her about her answers, telling her they were “excellent” and “great recommendations.”

Table 1
Overview of draft approach and results for Erica, Case 1

| **Draft Round** | **Essay parts submitted and Nature of edit** | **Scores** |
|---|---|---|
| 1 | C1 initial response | 0 |
| 2 | C2 initial response<br>C3 initial response | 1<br>0 |
| 3 | C1 final answer<br>C2 final answer<br>C3 final answer | 2 (Human<br>2 assigned<br>1 scores) |

In the second case Erica availed herself of the opportunity to revise her work once, but only on part two of her answer. She spent 21 minutes working on her answers. First, she drafted a response for answer parts 1, 2, and 3 and submitted them together for a predicted score. She received a 1, 0, and 1, respectively (see table 2). She chose to revise only part two of her answer, and did so by adding two significant points to it. First, she referenced information about the computer access and how this plays a part in whether or not technology can be used in an effective way within her classes. She also added a lengthy section about her curriculum, and how she would need to change it in order to make room for technology-enhanced projects. 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 2
Overview of draft approach and results for Erica, Case 2

| **Draft Round** | **Essay parts submitted and Nature of edit** | **Scores** |
|---|---|---|
| 1 | C1 initial response<br>C2 initial response<br>C3 initial response | 1<br>0<br>1 |
| 2 | - C2 added how access impacts use, and how use impacts curriculum | 0 |
| 3 | C1 final answer<br>C2 final answer<br>C3 final answer | 2 (Human<br>2 assigned<br>2 scores) |

On her 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 they only “somewhat impacted” her final responses. She explained that the she thought that “the rubric criteria did not seem to match well with the 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 part two, 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 he would be scoring their work. 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 assessment. The series of events in the computer log contrasts a bit with her recollection, in that her first case shows a prediction of 0, 1, 0---although this does agree with her memory that her part two answer was scored highest. And, they also show 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 unlikely 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 human.

**Stacy**

Stacy represents a group of users that had a favorable opinion of the automated essay scorer, received high scores (2s) from the instructor on all six of her responses, and on average completed two or more drafts per response per case. On her pre-assignment survey, Stacy reported an average grade of "A+" in her teacher education courses. Seeking licensure in English education, she has taken four courses on writing mechanics and writing assessment during her studies thus far. Stacy also noted that she is confident in her computer skills, is comfortable with computers, and believes that computers have a valuable role to play in aiding assessment, both in formative and summative fashions.

Stacy noted on her post-assignment survey that she found the eTIP Cases "useful" in learning about technology integration; "I learned the benefit of knowing about the culture and resources of technology in a school that I might work in." In her interview, Stacy shared that she "thought about this project" when she contacted the school she will be conducting her student teaching at; in addition to the classroom teacher, she also contacted the school's "media person" because she was "not going to start planning units if they do not even have [the technology] for me to show iMovie on." Overall, 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 one). She received a predicted score of 1 for parts one and two, and a 0 for part 1. Stacy then proceeded to revise and resubmit her part one answer another seven times. In these drafts she mostly added to her answer. In the first few revision rounds she focused on the conclusion and added more detail by either using words from the eTIP 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 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 parts one, two, and three, 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 1
Overview of draft approach and results for Stacy, Case 1

| **Draft Round** | **Essay parts submitted and Nature of edit** | **Scores** |
|---|---|---|
| 1 | C1 initial response<br>C2 initial response<br>C3 initial response | 0<br>1<br>1 |
| 2 | • C1 added to concluding sentence by referencing to the eTIp for the case | 0 |
| 3 | • C1 added to the conclusion by referencing additional case and challenge detail | 0 |
| 4 | • C1 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 | • C1 deleted a statement about what she would recommend and added a statement referencing case information about students' | 0 |

| | | |
|---|---|---|
| | skills | |
| 6 | • C1 added a reference to facts in the case’s introduction | 0 |
| 7 | • C1 deleted a summary statement referencing her students | 0 |
| 8 | • C1 added a statistic about students’ achievement | 0 |
| 9 | C1 final answer<br>C2 final answer<br>C3 final answer | 2 (Human<br>2 assigned<br>2 scores) |

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 answer parts 1, 2, and 3, respectively (see table 2). She did one further edit of part three of her answer, adding a couple of reference to software by name.

Table 2
Overview of draft approach and results for Stacy, Case 2

| **Draft Round** | **Essay parts submitted and Nature of edit** | **Scores** |
|---|---|---|
| 1 | C1 initial response<br>C2 initial response<br>C3 initial response | 1<br>1<br>0 |
| 2 | • C3 added examples of some software tools. | 0 |
| 3 | C1 final answer<br>C2 final answer<br>C3 final answer | 2 (Human<br>2 assigned<br>2 scores) |

Stacy indicated on her survey that she drew upon her notes, the provided rubric, and the automated essay scores, in that order of importance, when she formulated her answer. The log shows that in both cases Stacy visited 100% of the case’s pages. In her interview she described the assignment as “overwhelming” and that because she was not sure how else to proceed through the available information she did so sequentially. The result was that mid-way through the available information she had “information overload”. Stacy described that she took notes because her instructor indicated one should, and that in the end these really helped her by focusing her attention. She suggested that the volume of information was daunting and that she could not tell if she was over preparing or not. 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.”

### Summary

An unexpected theme in these students’ responses was that the automated essay scorer evoked an emotion. In their interviews the students were rather animated in describing their experiences with the scorer and while we asked them what the scorer made them do with their essay drafting, the students always volunteered information about how it made them feel. Erica was disdainful of the software’s ability to judge writing and so she ignored the score predictions of zero and instead relied upon her own judgment as to whether or not her essay was ready for final submission. Adam got caught up in a self-described competition with the scorer and revised and revised his work in an effort to try and make it return a higher score prediction. Mitch revised his work when he received a zero, but didn’t describe this as stemming from competition; perhaps instead it was from a sense of guilt or his desire to be a good student. Stacy made extensive revisions in one case and was very clear in describing the negative emotions she felt when her scores did not change as she revised her work. Whether disdain, competition, guilt, or anger, the scorer provoked an emotional response from the students.

Two of the students, Adam and Erica, had rather negative opinions of the scorer. This seemed to stem more so from their opinion that computers had little, if any, of a role in grading students’ written work than from their scores or their summative opinion of the learning they derived from the cases. Stacy and Mitch were selected for interviews because they had more positive opinions about the scorer. They had somewhat similar opinions of the value of the cases, but varied in their drafting behavior and scores. However they both thought that computers overall had some potential for scoring students’ written work but only in particular situations.

The changes these students made to their essays were positive ones in terms of writing structure: they provided additional examples or details, or referenced either the challenge from the case introduction or the eTIP the case was designed to provide practice thinking about. However, using the provided rubric we scored the first draft they 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 (i.e. they were not yet at the highest level on the rubric, a score of 2).

## Implications and Conclusion

These in-depth data about students’ reactions and opinions to automated essay score predictions suggest several implications for the ETIPS project staff members. First, that the support materials for faculty and students need to address the role of automated essay scoring and discuss the possible emotions that it may evoke in users. Some of the negative feelings expressed by these users are unlikely to aid learning from the cases, and may in fact detract from it.

Secondly, the nature of the feedback given to students needs to be more sophisticated. A simple prediction of a 0, 1, or 2 score evidently does not, in combination with the rubric to which these cores refer, provide enough guidance to students as to how to improve their essays. The feedback instead needs to be more detailed. Stacy, one of the students, made a similar suggestion as she described what a teacher would do. The score prediction data could easily be combined with the ability of the software to track which case information students accessed and return to them hints about specific information they should access in the case and relate to the case questions.

Finally, these data illustrate that the scorer fairly consistently returns inaccurate predictions of scores. The scores are usually one point lower than how we scored the students' drafts. While these low predictions did, in some cases, prompt the students to revise their work, it also provoked some negative responses. In combination with the simple form of feedback mentioned above, the students were not guided as to how to improve their work. And when their scores did not go up, they seemed to lose confidence in the tool's ability to give helpful feedback. Clearly, the accuracy of the scorer is important.

## Conclusion

Of course, the simulated experiences the ETIPS cases provide would never be intended to supplant student teaching and other field experiences. They can supplement them, however, and provide the added value of an inexpensive and easy to arrange learning opportunity. A reliable automated essay scoring tool in use within them, provides even further added value in that the learning opportunity could provide more feedback, and presumably improved performances, on complex sets of skills. Further, this feedback could also be easily stored and retrieved to review change over time. Iterations of timely feedback and revision can foster metacognitive skills in learners on the instructional decision-making process demanded by these learning materials.

## References

Alexander, M., Truell, A., & Bartlett II, J. (2002, Winter). Students' perceptions of online testing. *Delta Phi Epsilon Journal, 44*(1), 59-68.

Almond, R., Steinberg, L., & Mislevy, R. (2002, October). Enhancing the design and delivery of assessment systems: A four process architecture. *The Journal of Technology, Learning, and Assessment, 1*(5), 1-63.

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.

Birenbaum, M. (1996). Assessment 2000: Towards a pluralistic approach to assessment. In M. Bierenbaum & F. Dochy (Eds.) *Alternatives in assessment of achievements, learning processes and prior knowledge* (pp. 3-30). Boston:Kluwer.

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

Boyle A.P., Bryon, D.N., & Paul, C.R.C. (1997). Computer-based learning and assessment: A palaeontological case study with outcomes and implications. *Computers and Geosciences, 23,* 573-580.

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.

Charman, D. (1999). Issues and impacts of using computer-based assessments CBAs for formative assessment. In S. Brown, P. Race, & J. Bull, *Computer-assisted assessment in higher education* (pp. 85-93). London: Kogan Page Limited.

Clariana, R. B. (1993). A review of multiple-try feedback in traditional and computer based instruction. *Journal of Computer Based Instruction, 20*(3), 74-76.

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.

Dochy, F. & McDowell, L. (1997). Assessment as a tool for learning. *Studies in Educational Evaluation, 23*(4), 279-298.

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.

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

Ravitz, J. (2002). CILT2000: Using technology to support ongoing formative assessment in the classroom. *Journal of Science Education and Technology, 11*(2), 293-296.

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.