TIME-SHIFTED ONLINE COLLABORATION: CREATING TEACHABLE MOMENTS THROUGH AUTOMATED GRADING[1]

Chapter to appear in Juan, A.A., Daradournis, T., and Caballe, S. (Eds.) 2009. *Monitoring and Assessment in Online Collaborative Environments: Emergent Computational Technologies for E-learning Support.* IGI Global

Edward Brent[2]
Curtis Atkisson
Idea Works, Inc. and University of Missouri, USA
Nathaniel Green
Idea Works, Inc., USA

ABSTRACT

This chapter examines online collaboration in a distributed e-learning environment. We describe the emerging technology of web-based automated essay grading that provides extensive real-time data for monitoring and enhancing e-learning activities. We examine data from student use of this software service in a large introductory social science class. Using information routinely collected by the system, we find that students take advantage of this learning environment to revise and resubmit essays, dramatically improving their final grade by an average of 20% or two letter grades. We conclude the essential components of this learning environment that makes it so successful are its ability to provide detailed, personalized feedback to students immediately after they submit their work along with the opportunity to revise and resubmit. This transforms an automated assessment tool into a powerful collaborative learning environment. Instead of waiting days or weeks for instructor comments, that feedback is time-shifted to occur at the time it can be most effective for students. Changing the timing of feedback creates a powerful teachable moment when students have *motivation*, *information*, *opportunity*, and *feedback*. They are motivated to improve their grade, they are told what they did right or wrong while the relevant information is fresh in their minds, and they have the opportunity to revise and resubmit. The chapter ends with consideration of how those same elements can be, and sometimes already are, used in other instructional strategies such as podcasting to create effective learning environments that take advantage of the teachable moment.

INTRODUCTION: STUDENT – INSTRUCTOR COLLABORATION

The relationship between students and instructors takes many forms. In traditional large lecture course communication is mostly one-way with very little interaction. Most class time is devoted to the instructor lecturing passive students. Then, a few times a semester student performance is evaluated in tests. The role of the instructor is primarily to lecture and evaluate student performance. There is little or

[1] We'd like to thank Matthew Wood and James Barton who helped program the reporting interface for SAGrader that provided much of the information reported here. We also wish to thank the anonymous reviewers who raised important issues that greatly strengthened this work.

[2] Idea Works, Inc., 100 West Briarwood, Columbia, MO 65203 USA. www.ideaworks.com. brente@ideaworks.com.

no two-way communication between instructor and student. In contrast, the constructivist classroom (Jonassen, Peck, & Wilson, 1999) provides a collaborative learning environment in which the instructor's role is guiding students, coaching them to help them learn. In such a classroom there is repeated interaction between instructor and student as the student benefits from instructor feedback to enhance their understanding and, not incidentally, improve their grade. The constructivist environment is a collaborative learning environment in which students learn from repeated instructor feedback.

Increasingly, that collaboration takes place online. Distributed online environments provide the benefits of space-shifting: students and instructors need not be in the same place to interact effectively. This dramatically expands learning opportunities for many, increasing access and reducing travel costs and time. However, some of these new online environments have the ability to provide not only space-shifting but also time-shifting. "Many times learners are more interested in time-shifting capabilities provided by technology-based distance education systems than they are in the location-shifting capabilities of the systems" (Major & Levenburg, 1999). Unfortunately, synchronous collaboration which is the most effective is often impractical, while asynchronous collaboration is much easier to arrange but less effective for learning.

Synchronous collaboration (as exemplified by traditional classrooms, telephone conversations, or chat rooms) requires students and instructors to engage in interaction at the same time. The importance of interaction in learning is widely recognized from many perspectives, though it may often be described as "engagement," "participation," or "collaboration." It is widely agreed that "learning rarely takes place solely through unidirectional instruction. The social process of interaction is required for optimal learning (Lave & Wegner, 1991)." "Many studies have found that some form of participatory interaction by students is critical to their success in face-to-face and in distance education courses (Kearsley, 1995; Sutton, 2001). This engagement is illustrated by what it must have been like for Plato and Socrates to participate in a Socratic dialogue[3]. There the instructor and student are simultaneously focused on the learning task. The student can ask questions and immediately receive a response from the instructor. The instructor can gauge the student's progress and decide when she is ready to go on to more advanced material. If the student makes a mistake, the instructor can immediately point out the problem and explain what they did wrong.

Synchronous interaction encourages such engagement. But it does so at a cost. Merely putting instructor and student in the same room at the same time does not assure effective collaboration. Synchronicity requires students to learn on the same schedule that instructors teach, something that can be inconvenient or impossible for students whose work schedule causes them to miss classes. Worse yet, it may impose a schedule for learning on students that does not fit their optimal learning times (witness students sleeping in class, doodling, or surfing the web with their laptops). Likewise, it can impose a schedule on instructors that conflicts with other activities (e.g., when students all submit papers at the same time just before a holiday, or call the instructor's home at night with questions).

Asynchronous collaboration, typified by email or discussion group postings, does not require that everyone participate at the same time and is often easier to arrange, but may require long waits for a response from the instructor/collaborator. This reduces the interactive dynamic of the collaboration and may render learning less effective. For example, in large classes it often takes days or weeks to grade

[3] Participating in a Socratic dialogue, as discussed here, should not be confused with reading a Socratic dialogue. In the latter case, time-shifted learning is not occurring and there is no feedback from the instructor to students.

papers and students may have forgotten many of the key points in their essay. Often there is a "breakdown in conversational 'flow' due to the lack of continuity in the discussion over an extended timescale" (Hewson & Laurent, 2008); (Bowker & Tuffin, 2004); (Murray & Sixsmith, 1998). Because grading essays is a time-consuming and often onerous process for instructors, students are rarely given the opportunity to revise their essays. Instead of a learning environment in which students learn through collaborative interaction with the instructor, it is primarily an assessment environment in which students get one chance to achieve their grade. There is a temporal gap, and often a large one, between when students are primed to learn and when they have the guidance they need to benefit from that feedback. Even when they receive the feedback there may be no opportunity to capitalize on it through revisions.

This breakdown in interaction between student and instructor threatens learning. This chapter describes how an automated essay grading system, SAGrader™, can be used to implement a collaborative learning environment for e-learning that has the *convenience* of asynchronous collaboration and the *power* of synchronous collaboration. This method of collaboration in effect time-shifts instructor responses and assessments of student work. It allows the instructor to specify assistive knowledge to the student at a time convenient for the instructor. Then it allows the student to interact with the instructor-provided materials through assessments on the student's time frame. In this learning environment, students submit essays and immediately find out their grade, receive detailed feedback, and have the opportunity to revise their work. Students just received their grade and are motivated to improve it. They see detailed feedback to guide that revision. They have just completed the previous draft and have the necessary information in their grasp. For instructors, automated grading makes it possible to permit students to revise multiple times with little or no additional effort for the instructor. Together, these elements—motivation, information, feedback, and opportunity—create a uniquely powerful "teachable moment" analogous to the teachable moment for smoking cessation when someone is diagnosed with lung cancer (Gritz & al, 2006).

In this chapter we describe how SAGrader™ works, compare it to other automated essay grading programs less suitable for this task, and show how it can be used to implement a collaborative learning environment. We examine data from a classroom application of SAGrader and show how it creates a collaborative relationship between instructor and students, examining how students take advantage of that opportunity, showing how their performance improves as a result, and reporting student assessments that reflect on the collaborative aspects of the course made possible by SAGrader. In the process we address two of the themes of this book: how to provide for ongoing assessment in collaborative learning environments and how to effectively and efficiently manage computing resources in the distributed systems over which the collaborative learning system is implemented.

USING SAGRADER TO ENHANCE THE COLLABORATIVE LEARNING ENVIRONMENT

Computer-supported collaborative learning is learning that takes place through computer-assisted dialogue among students and between students and instructors (Findley, 1988). While the term is most often used to describe systems in which student peers collaborate to produce a joint product, instructors are inevitably involved to assess learning and the collaboration between student and instructor is of equal importance. When the opportunity for repeated feedback is available this collaboration between student and instructor can be extensive. SAGrader™ is an online automated essay grading service developed by Idea Works, Inc (Brent, Carnahan, McCully, & Green, 2006); (Brent & Townsend, 2007). SAGrader permits students to submit essays at any time over the world-wide-web using standard web browsers. Once their essay is submitted, SAGrader employs a number of artificial intelligence strategies to automatically grade the essay. Students receive immediate feedback indicating their grade along with detailed comments indicating what they did well and what needs further work. Since SAGrader grades essays automatically it is practical to permit students to revise their work and resubmit it for grading as often as instructors will allow. The effect for students is something analogous to immediate collaboration with the instructor while they are writing and rewriting their paper. SAGrader is currently in use in both on-site and online courses in several disciplines and at a range of educational institutions.

SAGrader can interface seamlessly with other more broadly collaborative systems since it can be accessed through an Application Programming Interface or can be "skinned" (made to appear like other applications). SAGrader addresses one of the major concerns of this book -- there are no standard models to monitor and perform efficient assessment of activity in online collaborative working and learning environments to guide and support the development of efficient collaborative projects. SAGrader provides one such mechanism for assessing performance in collaborative environments. It provides assessments of writing and hence can assess higher-level reasoning (Bloom, 1956) instead of the rote memory and recall assessed by multiple choice tests. It is also more objective, faster, and more economical than human scorers (Rudner & Gagne, 2001); (Yang, Buckendahl, & Juskiewicz, 2001), hence suitable for larger populations.

Related work: automated essay grading

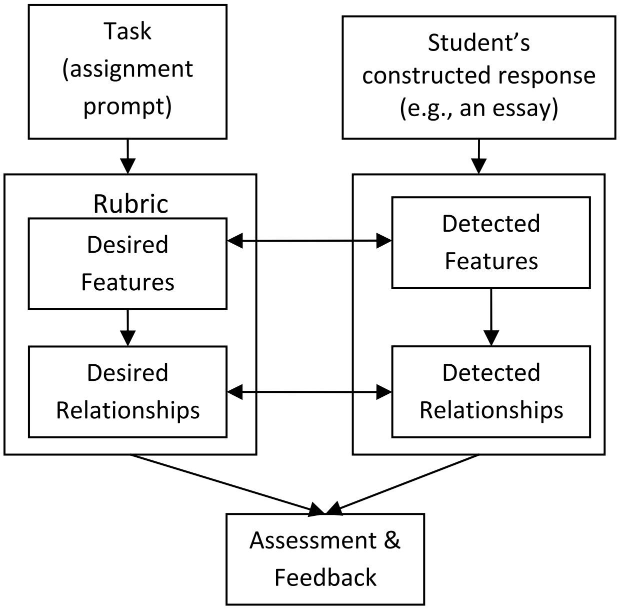
SAGrader is one of several commercially available programs for automated assessment of writing, including the Intelligent Essay Assessor (Landauer, Laham, & Foltz, 2000), the Electronic Essay Rater developed by Burstein and her colleagues at the Educational Testing Service (Burstein, 2003), and the Intellimetric program (Elliot, 2003). These programs are playing an increasing role in assessment and instruction; however, SAGrader differs in important ways from most of those other programs.

Many essay grading programs are based on statistical models, judging essays by how they compare to other good and bad essays. Criterion, Vantage, and IEA all employ a statistical model for developing and assessing the automated grading model. In each case human graders must first grade many (usually several hundred) essays. Those overall grades are then used as the "gold standard" to fit or "train" statistical models predicting scores assigned by human graders from features of essays measured by the programs (Yang et al., 2001). Once trained, the resulting model can then be used to assign grades to papers in the test set without human graders. Unfortunately, while these programs can determine if an essay looks more like good essays than bad ones, they have difficulty providing useful feedback to tell students how they could improve their score. Other early essay grading programs (, 1994) emphasized writing style over substance (Deane, 2006) and often measured proxies that correlated with good writing rather than features of good writing.

However, both of these approaches are often found wanting. Chung and Baker (2003) review the literature assessing the reliability and validity of automated essay grading programs and conclude that where there is a correct answer or a range of correct answers, the sensible "gold standard" for judging good writing is not whether it displays indirect measures that correlate with human readers' scores (Page, 1966; Page, 1994) or whether it matches documents having similar scores (Landauer et al., 1998); (Landauer, Laham, Rehder, & Schreiner, 1997). The important issue is not *consistency* with human graders who, after all, are notoriously inconsistent in their grading (Bejar & Mislevy, 2006) but the *validity* of the scores as measured by the fit of student essays to the knowledge that must be expressed in good essays. Valid measures will have the added advantage of providing *informative feedback* that can be used by teachers and students to improve, and to detect differences in writing as a result of instruction (Chung & Baker, 2003). To be effective at creating a collaborative learning environment an essay grading program must focus on substantive content rather than writing style (Deane, 2006) and must provide informative feedback to guide student revisions. SAGrader does both of these things.

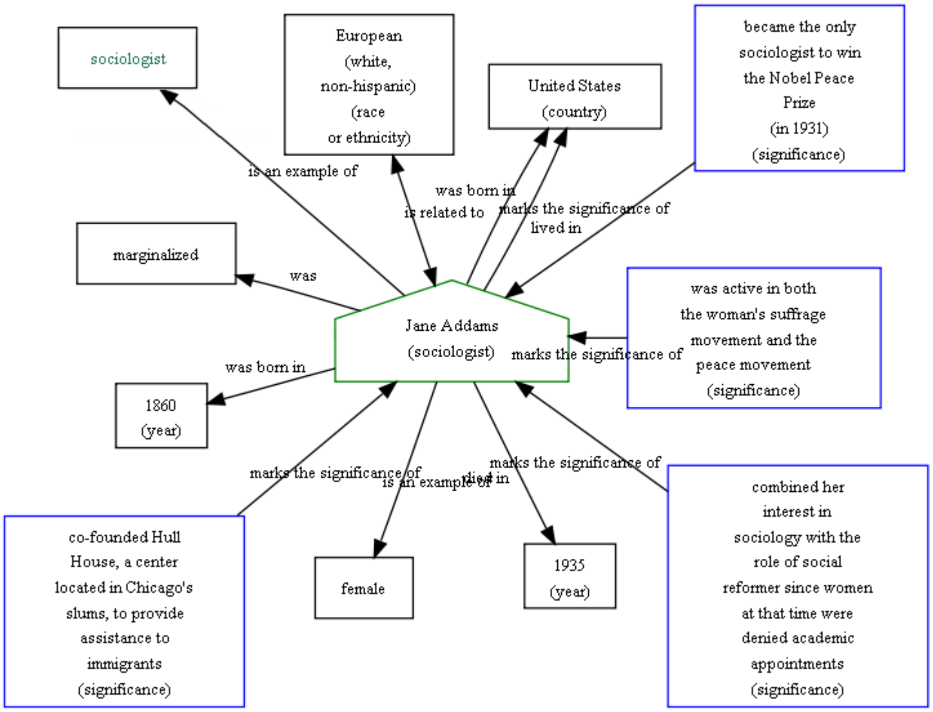
These programs use a wide range of natural language processing strategies (Cole, 1997) for recognizing important features in essays. Intelligent Essay Assessor (IEA) by Landauer, et al. (1998) employs a purely statistical approach, latent semantic analysis (LSA). This approach treats essays like a "bag of words" using a matrix of word frequencies by essays and factor analysis to find an underlying semantic space. It then locates each essay in that space and assesses how closely it matches essays with known scores.   E-rater uses a combination of statistical and linguistic approaches. It uses syntactic, discourse structure, and content features to predict scores for essays after the program has been trained to match human coders. SAGrader uses a strategy that blends linguistic, statistical, and AI approaches. It uses fuzzy logic (Zadeh, 1965) to detect key features in student papers, a semantic network (Sowa & Borgida, 1991) to represent the semantic information that should be present in good essays, and rule-based expert systems (Benfer, Brent, & Furbee, 1991); (Braun, Bejar, & Williamson, 2006) to compute scores based on how well a student's constructed responses match explicit standards of semantic content for good essays.

FIGURE 1: SAGRADER LOGIC



The operation of SAGrader is illustrated in Figure 1. The instructor first specifies the task or assignment in a prompt. Then the instructor creates a rubric identifying "desired features"—key elements of substantive knowledge that should be included in a good response, along with relationships among those elements.   For example, one SAGrader assignment asks the student to describe important characteristics of two well-known sociologists. The rubric specified that students should include several *desired features* in their essay, including the sociologist's name, the dates they were born and died, the country or countries in which they lived, the name of a theory for which they are known, one or more key concepts they first defined as part of that theory, and the title of an important publication of theirs related to that theory.   The *desired relationships* students should express include correctly linking each sociologist to the correct theory, concept, publication, and so on.   That rubric is then expressed in SAGrader as a semantic network or concept map. A partial semantic network for this essay showing information for a single sociologist is illustrated in Figure 2.

Figure 2: A Partial Semantic Network



A number of intelligent computational strategies, noted above, are used by SAGrader. Fuzzy logic permits the program to recognize literally thousands of word combinations that can be used to detect desired features in the student's essay, including not just concepts but also forms of reasoning and argument important for that discipline. Next, desired features based on the assignment and represented in the semantic network are compared to those detected in the student essay. Finally, an expert system using procedural rules scores student essays based on the similarities and differences between the desired and observed features. Detailed feedback for the student indicates what they did right and wrong, and provides a detailed breakdown of how points were assigned along with their overall score for the assignment.

SAGrader is suitable for a wide range of problems, problems with more than one good answer, and problems of varying difficulty. Assignments can ask about virtually any substantive knowledge instructors might want to assess. Even though SAGrader compares desirable features of good answers with features found in student essays, that still leaves great flexibility. These desired features can assess more abstract rhetorical goals as well as specific substantive knowledge. These features can be extensive, with many options, and considerable diversity. Assignments can permit students to include any of several possible correct answers, and question difficulty can be varied by giving students more or fewer options and requiring only some of the many possible correct answers. SAGrader has been used in sociology, psychology, English, biology, business, journalism, and several other disciplines. It could be used in any discipline where the instructor can specify the kinds of learning objectives students should meet in their essays.

Different instructors and different institutions use SAGrader in different ways. Not all of these create collaborative learning environments that take advantage of the teachable moment. Assignments may be required for credit or optional self-study exercises. Instructors may permit only a single draft, a specific number of drafts, or unlimited drafts of an assignment. The program's grade may be the final student grade or instructors may review the final draft and have the option of changing the student's grade. Feedback can be configured to provide informative guidance to students for further revision, or it can provide the final detailed summary suitable for self-study. How instructors configure SAGrader to work has a good deal to do with the kind of learning environment created. This chapter examines the use of SAGrader where students complete writing assignments for credit, receive immediate feedback designed to guide students as they make revisions while not telling them everything they will ultimately need to learn, and students are given the opportunity to revise their essays as many times as they wish. Together these three conditions fulfill the requirements for a teachable moment.

Efficiently managing computer resources

A second major concern of this book is the efficient management of computer resources in the distributed systems over which collaborative groupware learning systems are implemented. This poses a significant issue for SAGrader because in order to effectively time-shift it is important that feedback be immediate. The specific time required to provide feedback in one recent class ranged from between six and eight seconds for five simpler assignments to between 19 and 26 seconds for the five most complex assignments. The overall mean response time for all assignments was 19 seconds.

However, students often wait until the last minute to submit their assignments and in very busy times students might have to wait until other essays are graded before theirs is graded. Even when that happened in this course 99% of all essays were graded within one minute and the longest wait time was 81 seconds. So at this time the program is able to provide a very quick response even during peak time periods that effectively time-shifts instructor feedback to students to capture the teachable moment.

As use of SAGrader scales up, bursts of activity can be handled by the Amazon Cloud, a grid technology permitting SAGrader to automatically adjust the number of available microprocessors devoted to grading as a function of load. Only a few processors are in use all the time for essay grading and as student submissions increase additional processors are cloned and put into operation.

The role of feedback in creating a collaborative learning environment

A key aspect of SAGrader that makes it a collaborative learning environment is the design of feedback. This feedback is assistive, not exhaustive. Rather than showing students a complete answer, this feedback tries to point students in the direction of a complete answer, leaving it up to students to review the study materials and determine how to address the critique. In effect, the feedback is designed so that students could paste the feedback into their essay and still not improve their score.

This feedback strategy can be illustrated with a specific example. In this assignment students were asked to (among other things) read a hypothetical life history, identify six concepts related to social stratification, and define each concept.[4] Here is an example of the kind of feedback one student received regarding identifying the concepts:

[4] Admittedly, this assignment does not involve the highest level of reasoning we like to encourage in students (Bloom, 1956), but we want to keep this example simple enough to clearly illustrate feedback.

> *You received partial credit for identifying 4 items that are illustrated in Janice's autobiography: "working class," "welfare," "meritocracy," "working poor," and "structural mobility." You should also identify 2 more.*

Notice the student is told the specific items they got right, while they are only told the kind of items they missed, but not the specific items. The student must identify the missing items themselves.

Next the student received detailed feedback regarding definitions of the concepts. Note that feedback is only provided for concepts students have already identified.

> *You received no credit for identifying 0 items that define meritocracy. You should identify 1. You received full credit for identifying 2 items that define structural mobility: "changes in a society's occupational structure including the disappearance of some jobs and the appearance of other new ones" and "mobility resulting from a change in the occupational structure or stratification system."*

Here again, students are told the definitions they identified correctly, while they are only told which concepts lack definitions, not the definitions themselves.

As assignments become more complex and students are asked to identify a series of interrelated items the feedback continues to provide supportive information reinforcing correct answers while providing guidance for incorrect or missing items. As students improve their essays and correctly identify more items the feedback shifts accordingly, always trying to guide them in that next step to successfully completing the assignment. This feedback is designed to help students throughout the learning process while keeping the responsibility for learning squarely on the student's shoulders. In this manner the program provides a collaborative learning environment in which instructors (assisted by the program) collaborate with students to improve student learning.

Collaborating to improve instruction through student challenges

Clearly, the collaborative learning environment created by SAGrader assists students with feedback time-shifted from instructors. Less evident, but also important, students assist instructors with their own feedback in the form of challenges facilitated by SAGrader's built-in challenge and communication system. When students view feedback for their submission, if they believe the program graded them incorrectly they can challenge their grade. For example, here is a challenge one student entered for an assignment where they were asked to use sociological concepts and principles to interpret a description of a hypothetical community.

> *My example of sector theory is: "With Interstate 494 running north to south through the center of the city, Hwy. 55 bisecting the city east to west and Hwy. 169 running along the eastern border, people who live and work in Rockford have easy access to Twin Lakes and area suburbs."*

When challenging their result, students are asked to identify what item or items they believe the program missed and indicate how they expressed those items. The instructor then reviews the challenge and may resolve it in any of several ways, including a) explaining to the student why they are wrong and the program is correct, b) overriding the program's score and acknowledging the student's correct response, or c) acknowledging the student's response is correct and at the same time sending a communication to a developer to correct the program. For example, here is a response by the instructor to that student challenge.

> *The quotes you used were a good start, but think about what the terms mean, and what they are associated with. Sector theory is associated with growth, and urban decline is about the decline, not just the current state. --- This challenge was resolved by explaining to the student.*

Challenges help instructors further collaborate with students to help students learn and assure they are graded fairly. Equally important, challenges help students collaborate with instructors to improve the grading of the SAGrader program by pointing out any potential weaknesses in the program. Instructors can also, at their own initiative, review a student's assignment and then provide additional helpful feedback or even override the grade assigned by the program.

ASSESSING STUDENT COLLABORATION USING SAGRADER

To get an idea of how SAGrader contributes to a collaborative learning environment and how it affects student performance, we examined data from student essays submitted to assignments in a large introductory sociology course offered at the University of Missouri at Columbia. During Spring Semester of 2008, one hundred and seventy two (172) students submitted essays as a major part of the requirement for their grade.   This course had 16 required writing assignments over the semester along with 7 optional extra-credit assignments. Assignments varied from several short-answer questions to moderate sized (1 or 2-page assignments) to a full-blown 3-part 15-page term paper.   The SAGrader™ computer program graded all of these assignments.   Students were permitted to revise their essays as many times as they wished based on this feedback from the SAGrader program.

Students Use the Opportunity to Revise

To make the case that a collaborative learning environment facilitates learning it must be shown that there is collaboration and it results in improved learning. In the SAGrader learning environment collaboration occurs as students use feedback to improve their learning and increase their grades. Students submitted a total of 1,193 distinct essays to the assignments and 2,863 separate submissions, for a mean of 2.4 submissions (one first draft and 1.4 revisions) per assignment. When given the opportunity to revise their essays, students clearly took advantage of the opportunity for collaboration, revising their essays one or more times for 71% of the assignments. In most cases a relatively few revisions are made. In ninety percent of the essays, students submitted three or fewer revisions for an assignment. Eighty-four percent made two or fewer revisions.

Revisions Improve Grades Significantly

Given that students take advantage of the collaborative learning environment provided by SAGrader to revise their papers, the next question is "Does that collaboration lead to improved learning and higher scores?" The table below compares initial score and final score means and standard deviations for all essays, essays having first drafts not followed by revisions, and essays including one or more revisions.

Table 1. Mean Score and Standard Deviation for Initial Draft and Final Draft

	Number	**Initial Score (%)**		**Final Score (%)**	
	Number	**Mean**	**Standard deviation**	**Mean**	**Standard deviation**
All essays	2,866 (100%)	69	27.76	90	18.08
Essays with first drafts only	820 (29%)	87	20.92	87	20.92
Essays including one or more revisions	2,046 (71%)	61	26.72	91	16.67

For all essays across all assignments, the average performance for the first draft is 69% and the average for the last draft is 90%, for a change of 21% or approximately two letter grades on the 4.0 grading scale typically used in American universities. This is quite an improvement. It is statistically significant (t=34.42, df=5730, p<.0001) and produces an effect size (Glass, McGaw, & Smith, 1981) of 0.77 (the difference in means is .77 times the standard deviation)[5]. To put this in perspective, Cohen (1969) describes an effect size of .8 as “grossly perceptible and therefore large” and comparable to the differences between the average heights of 13 year old and 18 year old girls (Coe, 2002). In comparison, most educational interventions have effect sizes that are often around 0.2 ;Coe, 2002) or what Cohen calls “small” effect sizes.

However, in 29% of the cases (N=820) first drafts were not followed by revisions. The initial and final performance for those essays is the same -- 87%. So those 820 essays that show no change make the improvement due to revisions appear lower than it really is. To better understand the impact of revisions we look only at essays having one or more revisions. For essays having one or more revisions, the average performance on the first draft is 61% and for the last draft, 91%, resulting in a change of 30%, or a three-letter grade jump from a D- to an A-. This is an even more impressive improvement. It is statistically significant (t=42.84, df=4090, p<.0001) and produces an effect size of 1.12 (this being the most conservative estimate of effect size). That is, by using feedback from SAGrader™ and revising their work, students improved their grade by 30 percentage points or three letter grades—even before the instructor examined their essay. In fact, students who revise, on average, ultimately outperform students who do not revise, even though the non-revisers began with an average of 87% compared to a beginning average of 61% for revisers.

How Successful is this Collaboration?

In a successful collaborative learning environment we would expect students to continue with the learning process until they reach a level of performance that meets their personal standard. In an

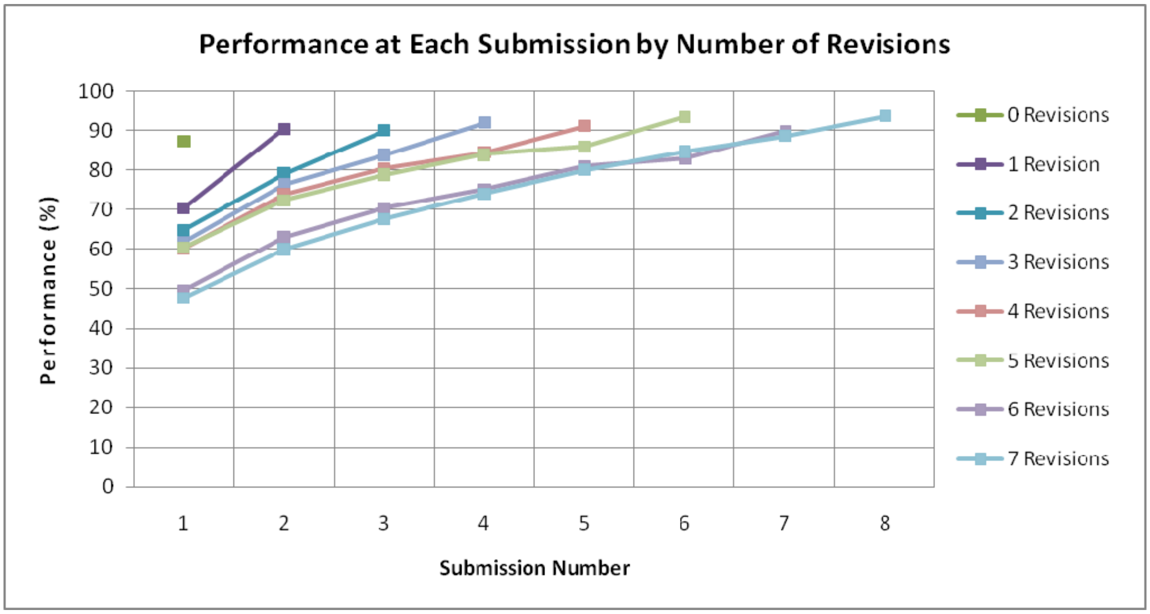
[5] Effect sizes were estimated conservatively by dividing mean differences by the largest standard deviation of the two groups.

unsuccessful collaborative learning environment we might expect to find large numbers of students who quit in frustration before reaching an acceptable level of performance. In the figure below, the performance scores for students for each essay are broken down by the number of revisions made by that student for the assignment. For example, there were 820 essays where students submitted no revisions and had an average score of 87%. There were 611 essays where students submitted one revision. Their initial score was 70.6% and their final score was 90.4%.

Table 2: Average scores for all submissions with final submissions highlighted

Number of	Number of	Submission Number

Figure 3. Mean scores by number of revisions and submission number



Notice that the average of scores for the last submission, whether it is the second, the eighth, or any in between, is essentially 90% or a bit higher. Thus it appears that for most essays, students seem to quit revising when their average approaches 90%, regardless of how many revisions it takes to achieve that average. Some students achieve a personally acceptable score with their first draft; others take one, two, or more revisions to reach that score. If final scores for essays with more revisions were substantially lower than the scores for essays with few revisions, this would suggest some of the students were giving up on some of the essays rather than being satisfied with their score. In fact, this does not appear to happen for most essays. This suggests students are getting the help they need during this collaborative learning process to enable them to perform up to their standard.

		1	2	3	4	5	6	7	8
820	0	87.2*	.	.	.	.	.	.	.
611	1	70.6	90.4*	.	.	.	.	.	.
425	2	65.1	79.3	90.1*	.	.	.	.	.
314	3	61.8	76.4	83.8	91.9*	.	.	.	.
187	4	60.2	73.7	80.3	84.4	91*	.	.	.
127	5	60.4	72.3	78.8	83.9	85.9	93.4*	.	.
97	6	49.7	63.1	70.5	75.1	80.9	83	89.8*	.
75	7	47.7	60	67.6	74	80	84.7	88.6	93.6*

While the final scores on essays are very much the same regardless of numbers of revisions, at each step there is a large gap between scores for students who stop revising and those where students continue revising. Generally, there is a difference of at least 12 percentage points and as much as 26 percentage points. Students who continue revising have mean scores on their latest submission from 61% to 79%, while students who decide to quit have mean scores between 87% and 94%. These results suggest students are both rational and tenacious, continuing to work on their papers when they are not satisfied with their score, quitting only when they achieve what they regard as a “good” or at least “acceptable” score. The decision about when to stop working on their papers appears to be determined primarily by their current grade and has little or no relationship to the number of revisions they have submitted.

Are they Really Learning, or Are they Just Gaming the System?

This improvement in scores from 61% to 91% for essays that are revised based on SAGrader’s feedback is impressive. But could it be overestimated? It might be argued that the improvement by students from first to last drafts is overestimated to the extent that students submit very poor first essays and then use feedback from SAGrader™ to modify their essay. That is to say, perhaps they are letting SAGrader do more of the work for them. It is possible that students do sometimes exercise less care with their first drafts and rely more heavily on SAGrader’s feedback to improve their final drafts. However, if this was widespread, one would expect most early drafts to be poor ones. Instead, the average score on first drafts (including students who do not do subsequent drafts) is 69%. The fact that, for 29% of essays, students stop after the first draft is also consistent with the view that students make a reasonably good effort on their first drafts and do not rely exclusively on advice from SAGrader to improve their scores. Finally, it is also possible that the improvements made by using SAGrader are underestimated to the extent that original essays receive good scores. For essays that are less difficult where initial average scores are already high, there is little room for improvement, no matter how helpful the SAGrader program’s advice might be. So any biases that might lead to overestimating learning with SAGrader may be countered somewhat by biases that underestimate learning.

A CASE STUDY

Further evidence that students are indeed learning is provided by examining a specific case. We can illustrate this general trend of dramatic improvement from first draft to final draft, when essays are revised, by examining a specific case. This student submitted her essay three times. The first submission was at 7:59 PM on February 23rd, and the student received a score of 27%. The second submission occurred 6 minutes later and the student received a much-improved score of 71%. She made her third submission for this assignment at 8:09 PM and received a perfect score of 100%. In the course of three submissions and ten minutes, this student went from a score of 27% to a score of 100% – a 73% improvement – all occurring long before an instructor could have graded and provided feedback on any of the essays. In fact, all of this improvement by the student occurred before an instructor even saw her essay.

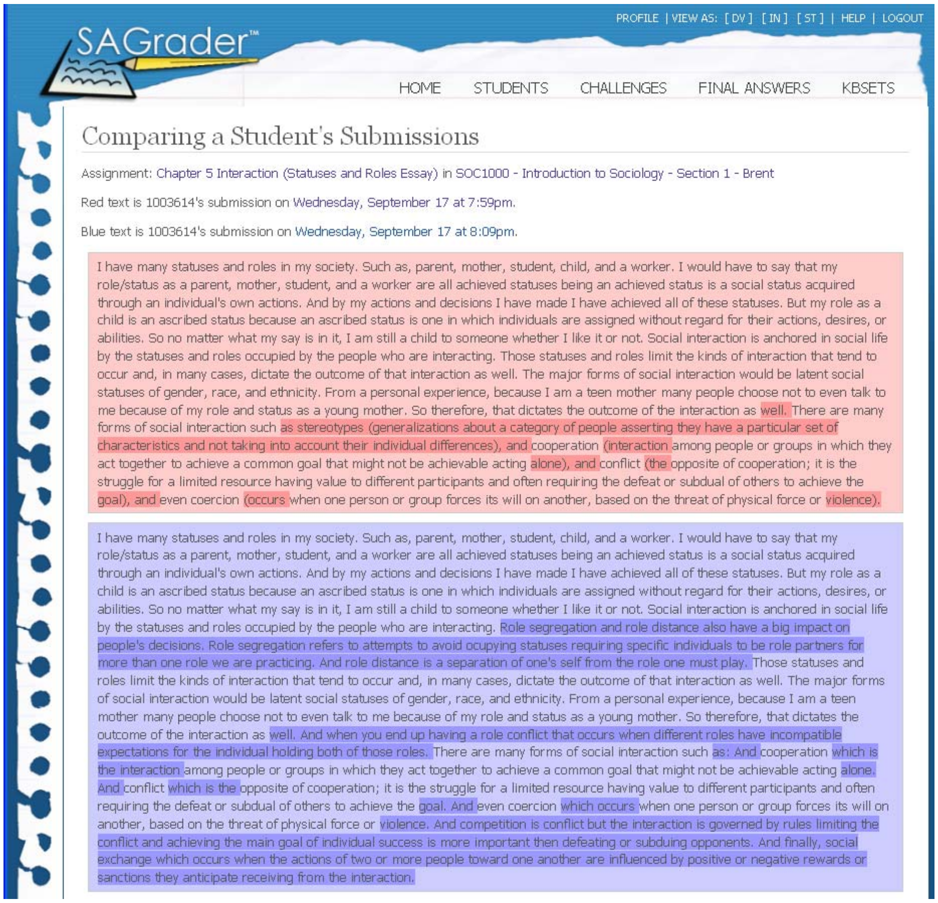
This case illustrates the general trends found above. In this case the student submitted three times (the average is 2.4 submissions). The student's initial grade was less than 70% (27% in this case). The student stopped revising the paper when they achieved a grade near 90% or higher (100% in this case). We believe that this is due to a student's natural tenacity and desire to maximize her grade. It is interesting to note that dramatic improvements such as this are facilitated by systems that provide an opportunity for students to continue working in a collaborative learning environment to improve their understanding.

In Figure 5 is a screen shot from SAGrader showing the student's first draft at the top and the last draft at the bottom for an essay addressing the following prompt:

> *In your readings about interaction, you found that the elements shaping most social interaction are social statuses and roles. Each of us has a multitude of statuses and roles that direct the ways we interact in various contexts. In this assignment, use your own sets of statuses and roles to discuss some of the major issues associated with the two concepts. For example, which of your statuses are ascribed or achieved? Have you ever experienced role segregation, distance, or conflict? Be sure to define all of these terms in your answer. In addition, you should discuss the major forms of social interaction. We have all experienced these varying forms of interaction to some degree, so make sure to pick examples that illustrate each type. Explain what each type is.*

Differences between the two drafts are highlighted. The display has been anonymized and displays a number instead of the student's name. The first draft score was 27%, for an effort that was clearly a serious attempt to answer the prompt. This shows that getting even a 27% on an essay is not a trivial matter and requires some effort and knowledge. We can see clear improvements in the essay. The student's initial draft had 330 words. The last draft was considerably longer, with 452 words. The final draft was more developed, with greater coverage of important topics and more precise language. Among other changes, it adds definitions of role conflict, role segregation, and role distance; and it adds several common types of social interaction not discussed earlier.

Figure 4: Comparison of First and Third Drafts of a Student's Essay



Comparing a Student's Submissions

Assignment: Chapter 5 Interaction (Statuses and Roles Essay) in SOC1000 - Introduction to Sociology - Section 1 - Brent

Red text is 1003614's submission on Wednesday, September 17 at 7:59pm.

Blue text is 1003614's submission on Wednesday, September 17 at 8:09pm.

I have many statuses and roles in my society. Such as, parent, mother, student, child, and a worker. I would have to say that my role/status as a parent, mother, student, and a worker are all achieved statuses being an achieved status is a social status acquired through an individual's own actions. And by my actions and decisions I have made I have achieved all of these statuses. But my role as a child is an ascribed status because an ascribed status is one in which individuals are assigned without regard for their actions, desires, or abilities. So no matter what my say is in it, I am still a child to someone whether I like it or not. Social interaction is anchored in social life by the statuses and roles occupied by the people who are interacting. Those statuses and roles limit the kinds of interaction that tend to occur and, in many cases, dictate the outcome of that interaction as well. The major forms of social interaction would be latent social statuses of gender, race, and ethnicity. From a personal experience, because I am a teen mother many people choose not to even talk to me because of my role and status as a young mother. So therefore, that dictates the outcome of the interaction as well. There are many forms of social interaction such as stereotypes (generalizations about a category of people asserting they have a particular set of characteristics and not taking into account their individual differences), and cooperation (interaction among people or groups in which they act together to achieve a common goal that might not be achievable acting alone), and conflict (the opposite of cooperation; it is the struggle for a limited resource having value to different participants and often requiring the defeat or subdual of others to achieve the goal), and even coercion (occurs when one person or group forces its will on another, based on the threat of physical force or violence).

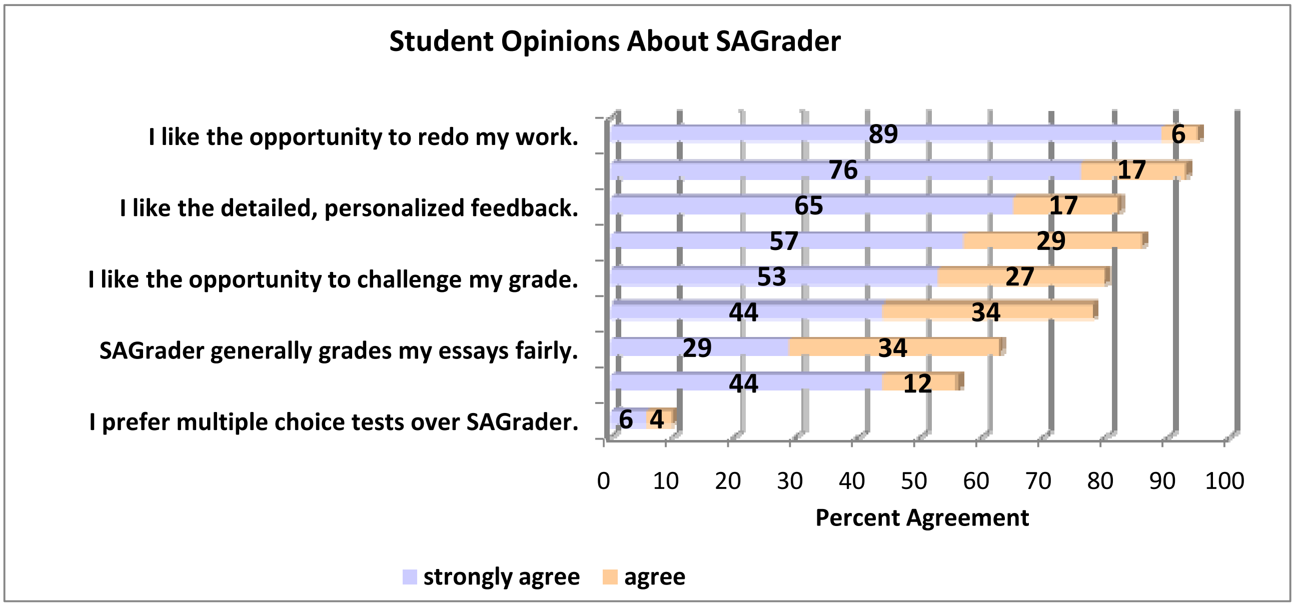
I have many statuses and roles in my society. Such as, parent, mother, student, child, and a worker. I would have to say that my role/status as a parent, mother, student, and a worker are all achieved statuses being an achieved status is a social status acquired through an individual's own actions. And by my actions and decisions I have made I have achieved all of these statuses. But my role as a child is an ascribed status because an ascribed status is one in which individuals are assigned without regard for their actions, desires, or abilities. So no matter what my say is in it, I am still a child to someone whether I like it or not. Social interaction is anchored in social life by the statuses and roles occupied by the people who are interacting. Role segregation and role distance also have a big impact on people's decisions. Role segregation refers to attempts to avoid ocupying statuses requiring specific individuals to be role partners for more than one role we are practicing. And role distance is a separation of one's self from the role one must play. Those statuses and roles limit the kinds of interaction that tend to occur and, in many cases, dictate the outcome of that interaction as well. The major forms of social interaction would be latent social statuses of gender, race, and ethnicity. From a personal experience, because I am a teen mother many people choose not to even talk to me because of my role and status as a young mother. So therefore, that dictates the outcome of the interaction as well. And when you end up having a role conflict that occurs when different roles have incompatible expectations for the individual holding both of those roles. There are many forms of social interaction such as: And cooperation which is the interaction among people or groups in which they act together to achieve a common goal that might not be achievable acting alone. And conflict which is the opposite of cooperation; it is the struggle for a limited resource having value to different participants and often requiring the defeat or subdual of others to achieve the goal. And even coercion which occurs when one person or group forces its will on another, based on the threat of physical force or violence. And competition is conflict but the interaction is governed by rules limiting the conflict and achieving the main goal of individual success is more important then defeating or subduing opponents. And finally, social exchange which occurs when the actions of two or more people toward one another are influenced by positive or negative rewards or sanctions they anticipate receiving from the interaction.

SAGrader is designed to insure that the correspondence between scores and knowledge found in this example is likely to be found in general. To begin with, constructed responses or essays are much harder to "fool" than fixed-choice tests, and the process of constructing essay responses itself often fosters learning. The way in which SAGrader assesses essays makes it hard to "game" the system and get a high score. Good essays must not only include appropriate concepts and terms, but must also express relationships among them consistent with the knowledge underlying the learning objectives. In the student excerpts above, for example, mentioning concepts such as competition or conflict alone is not enough to get full credit. Students must also indicate that those are both examples of social interaction. Definitions must be clearly linked to the correct concepts; authors must be associated with the correct theories; and so on. As one instructor said at an SAGrader workshop, if students are "gaming" the system the game they are playing is the one set up by the instructor, and they are learning whether they realize it or not.

STUDENT OPINIONS

As part of the normal course evaluation, students were asked to evaluate SAGrader. Their responses along with their open-ended comments reflect the collaborative learning environment created by SAGrader. Tops on their list, students like the opportunity to redo their work (95% agree with that, 89% strongly agree), they love the immediate feedback (93% agree, 76% of them strongly agree) and the detailed personalized feedback (82% agree, with 65% of them agreeing strongly). Students overwhelmingly agree that writing essays with SAGrader helps them learn (78% agree, 44% of those strongly agree). These responses reflect the importance to students of SAGrader's ability to take advantage of the teachable moment to provide a collaborative environment that facilitates learning.

Figure 5: Student Opinions



Students overwhelmingly like the fact that SAGrader grades everyone's essays without bias (86% agree, with 57% of those strongly agreeing)—no trivial matter in social science courses where students sometimes express concern instructors are biased and unfair in their grading. Most agree the program grades their essays fairly" (63% agree and 29% of those strongly agreed). Students also like the opportunity to challenge their grade (80% agree, with 53% strongly agreeing). This sends the message to students that we are concerned with fairness and accuracy and gives them an opportunity to voice any concerns they might have about having a computer program grade their work. When whether they prefer to have only SAGrader assignments and no multiple choice tests in the course, 56% of students agreed (44% strongly). In contrast, only 10% (6% strongly) preferred having only multiple choice tests and no SAGrader assignments.

While these results suggest an effective collaborative learning environment is created by SAGrader, there were a few cautionary results. In open-ended comments students occasionally say things like "Sometimes I feel that my examples do fit the description of the definition, but SAGrader says that it does not." "SAGrader was extremely picky about answers. I would say the right thing, just not the perfect way." or "You had to have pretty much the exact definition for SAGrader to understand it." At least some of the students sometimes think the program is asking for greater precision from them than is necessary. Of course that level of precision may be appropriate. Every discipline has technical terms students are expected to employ correctly, and a student's use of the phrase, "differentiation of labor" is *not* equivalent to the widely recognized sociological term, "division of labor". This is a mistake by the student, and the TAs or instructor would mark it wrong as well.

DISCUSSION AND CONCLUSIONS

When instructors grade essays, each submission incurs a substantial commitment of time, money, or both to grade. This encourages instructors to limit writing assignments and to restrict the number of times students can revise their work or prohibit revisions altogether. As a result, writing assignments are few and far between and when they do occur they are used primarily to evaluate students rather than as a learning experience. Because of the time required for grading, that evaluation often comes long after the essay is written. Writing-across-the-curriculum programs at many universities try to overcome this by making more resources available for selected courses to make it possible to have more writing assignments, to have enough graders to shorten the wait for grading, and to permit multiple revisions and learning through writing. Those writing programs have often not been extended to online learning environments at least in part because of the difficulties of achieving the level of interactivity and collaboration so crucial for effective writing assessment.

In contrast, SAGrader automates the grading of essays, creating an environment in which assessments of student revisions are nearly free. Once the initial essay assignment is constructed, the incremental cost of grading multiple drafts by each student is hardly more than the cost of grading a single draft (Brent, Carnahan, & McCully, 2006a). Thus, SAGrader provides both an assessment tool and a collaborative learning environment where students can submit essays, receive detailed, personalized feedback immediately while the issues are still fresh in their minds, and revise and resubmit their work. We have shown in this study that, when SAGrader is used in such an environment, students take advantage of this learning opportunity to revise their work based on feedback from the program, often dramatically improving their performance.

This analysis of students' use of SAGrader finds it to be a very effective learning environment in which revised essays improve by an average of 30 percentage points or three letter grades. Students work hard when given the opportunity to do so in this supportive collaborative learning environment. Students submit each essay an average of 2.4 times and make considerable changes between the first draft and last draft. Students display considerable tenacity and rationality, taking advantage of this learning opportunity by continuing to revise and resubmit their essays until they achieve an average grade of roughly 90% or higher, even when their initial grades may have been quite low. Students who revise many times generally end with average scores much like those of students who submit only once or a few times. This suggests that these students are not finally giving up and settling for lower scores, but instead are persisting with revisions until they perform up to the standard they have set for themselves. The dramatic improvement in scores from first draft to final draft does not appear to be an artifact of artificially low first draft scores or artificially high last draft scores. Students are making changes in their essays that show evidence of learning and improved writing. The case study illustrates how this large change in scores chronicles a true improvement in understanding and communication that is reflected in increased length, quality of writing, coverage of topics, and writing precision. Remarkably, students make all this progress largely unattended by instructors.

Students may be unattended, but they are not unassisted. By providing the feedback and opportunity to revise at that crucial time, SAGrader in essence time-shifts interaction with the instructor to fit the student's schedule. The feedback to the student is crucial for this process. By being able to view detailed comments and advice on how to improve their essay immediately after each revision, students are in essence collaborating with the instructor. This is not synchronous collaboration because the instructor does not participate at the same time as the student. Nor is it asynchronous collaboration in which the student must wait days or weeks for feedback from the instructor. Instead, this is time-shifted collaboration in which the strength of synchronous collaboration is delivered without requiring student and teacher to be actively engaged at the same time. It is as though students had an instructor looking over their shoulder responding to their work, providing encouragement and direction for improvement, permitting students to take advantage of the teachable moment to achieve significant learning.

It is interesting to note that there are a number of other pedagogical solutions that provide approximations of time-shifting – even though they don't use that terminology. There are also many learning environments that fail to incorporate time-shifting. We know of no existing literature that examines the time-shifting aspect of pedagogical solutions. It is our hope that this article will stimulate such work so that we can better understand how time-shifting can be used to enhance existing educational strategies. We argue that those educational environments that fail to provide time-shifted interaction are missing out on important ways to facilitate student learning.

Some teaching tools capture many of the aspects of time-shifting but not all aspects. One of the oldest and most widely used educational tools – the textbook – in effect time-shifts information. What the textbook does is take the knowledge of the author, and time-shifts that knowledge to the moment in time that the student desires to learn from the author. The textbook fulfills a number of the requirements to be a student-directed teachable moment, but fails in one respect – feedback. In order to take maximum advantage of the moment when students are most ready to learn, the textbook would need to provide feedback on how the students are performing. Some learning environments that do a better job of capturing the teaching moment are computer-based multiple choice assessments and fill-in-the-blank lecture notes. Some common practices that fail to take advantage of the teachable moment are hand-graded essays, lectures, and optically scanned tests, each of which typically provides feedback to students days or even weeks after they complete their work. On the positive side, there are some technologies that have started moving closer to taking advantage of teachable moments including podcasts, clickers, and online chat and discussion groups.

The key to determining whether or not a technology will take full advantage of time-shifted interactions is to identify whether the technology has all of the features needed to create teachable moments. In order for these teachable moments to be attained, the student needs to be motivated, have the relevant information in hand, receive feedback on their work, and have an opportunity to revise. All of this needs to be available in a span of minutes or seconds. We argue that only by providing access to those teachable moments, and taking advantage of them, can teachers realize the full potential of their students. The results from this study show that students exposed to such a teachable moment will take full advantage of it.

REFERENCES

Bejar, I., & Mislevy, R. (2006). Automated Scoring of Complex Tasks in Computer-Based Testing: An Introduction. In D. Williamson, R. Mislevy & I. Bejar (Eds.), *Automated Scoring of Complex Tasks in Computer-Based Testing*. Mahwah, NJ: Erlbaum.

Benfer, R., Brent, E., & Furbee, L. (1991). *Expert Systems*. Newbury Park - London: Sage.

Bloom, B. (1956). *Taxonomy of educational objectives: The Classification of Educational Goals*: Susan Fauer Company, Inc.

Bowker, N., & Tuffin, K. (2004). Using the Online Medium for Discursive Research About People with Disabilities. *Social Science Computer Review, 22*(2), 228-241.

Braun, H., Bejar, I., & Williamson, D. (2006). Rule based methods for automated scoring: Application in a licensing context. In D. Williamson, R. Mislevy & I. Bejar (Eds.), *Automated scoring of complex tasks in computer based testing*. Mahwah, NJ: Erlbaum.

Brent, E., Carnahan, T., & McCully, J. (2006a). *Students Improve Learning by 20 Percentage Points with Essay Grading Program, SAGrader*™. Columbia. MO: Idea Works, Inc.o. Document Number)

Brent, E., Carnahan, T., McCully, J., & Green, N. (2006). SAGrader™: A Computerized Essay Grading Program, Version 2. Columbia, Missouri: Idea Works, Inc.

Brent, E., & Townsend, M. (2007). Automated Essay Grading in the Sociology Classroom: Finding Common Ground. In P. Frietag Ericsoon & R. Haswell (Eds.), *Machine Scoring of Student Essays: Truth or Consequences?* : Utah State University Press.

Burstein, J. (2003). The e-rater® scoring engine: Automated essay scoring with natural language processing. In M. Shermis & J. Burstein (Eds.), *Automated essay scoring: A cross-disciplinary perspective*. Mahwah, NJ: Erlbaum.

Burstein, J., & Kukich, K. (1998c). *Computer analysis of essays.* Paper presented at the NCME Symposium on Automated Scoring, Montreal, Canada.

Burstein, J., Kukich, K., Chodorow, M., Draden-Harder, L., Harris, M., Wolff, S., et al. (1998a). *Automated scoring using a hybrid feature identification technique.* Paper presented at the Association of Computational Linguistics, Montreal, Canada.

Burstein, J., Kukich, K., Wolff, S., Lu, C., & Chodorow, M. (1998b). *Enriching automated scoring using discourse marking*. Paper presented at the Annual Meeting of the Association of Computational Linguistics.

Chung, G., & Baker, E. (2003). Issues in the reliability and validity of automated scoring of constructed responses. In M. Shermis & J. Burstein (Eds.), *Automated essay grading: A cross-disciplinary approach*. Mahwah, NJ: Erlbaum.

Coe, R. (2002). *It's the Effect Size, Stupid: What effect size is and why it is important* Paper presented at the British Educational Research Association.

Cohen, J. (1969). *Statistical Power Analysis for the Behavioral Sciences*. NYC: Academic Press.

Cole, R. (1997). *Survey of the State of the Art in Human Language Technology*. Cambridge: Cambridge University Press and Giardini.

Deane, P. (2006). Strategies for Evidence Identification Through Linguistic Assessment of Textual Responses. In D. Williamson, R. Mislevy & I. Bejar (Eds.), *Automated Scoring of Complex Tasks in Computer-Based Testing*. Mahwah, NJ: Erlbaum.

Elliot, S. (2003). Intellimetric™: From Here to Validity. In M. Shermis & J. Burstein (Eds.), *Automated Essay Scoring: A Cross-Disciplinary Perspective*. Mahaw, NJ: Lawrence Erlbaum Associates, Inc., Publishers.

Findley, C. (1988). *Collaborative Networked Learning: On-line Facilitation and Software Support, .* Burlington, MA: Digital Equipment Corporationo. Document Number)

Glass, G., McGaw, B., & Smith, M. (1981). *Analysis in Social Research*. London: Sage Publications.

Gritz, E., & al, e. (2006). Successes and Failures of the Teachable Moment : Smoking Cessation in Cancer Patients. *Cancer, 106*(1), 17-27.

Hewson, C., & Laurent, D. (2008). Research Design and Tools for Internet Research. In N. Fielding, R. M. Lee & G. Blank (Eds.), *Online Research Methods*. London: Sage Publications.

Jonassen, D. H., Peck, K., & Wilson, B. (1999). *Learning with Technology: A Constructivist Perspective*: Merrill.

Kearsley, G. (1995). *The Nature and Value of Interaction in Distance Learning*. Paper presented at the Third Distance Education Research Symposium.

Landauer, T., Laham, D., & Foltz, P. (2000). The Intelligent Essay Assessor. *IEEE Intelligent Systems, 15*, 27-31.

Landauer, T., Laham, D., Rehder, B., & Schreiner, M. (1997). *How well can passage meaning be derived without word order? A comparison of latent semantic analysis and humans.* Paper presented at the Cognitive Science Society, Mahwah, NJ.

Lave, J., & Wegner, E. (1991). *Situated Learning: Legitimate Peripheral Participation*. Cambridge: Cambridge University Press.

Lipsey, M., & Wilson, D. (1993). The Efficacy of Psychological, Educational, and Behavioral Treatment: Confirmation from Meta-Analysis. *American Psychologist, 48*(12), 103-121.

Major, H., & Levenburg, N. (1999). Learner Success in Distance Environments: A Shared Responsibility. *The Technology Source*.

Murray, C., & Sixsmith, J. (1998). Mail: A Qualitative Research Medium for Interviewing? *International Journal of Social Research Mthodology: Theory and Practice, 48*(12), 103-121.

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

Page, E. (1994). Computer Grading of Student Prose, Using Modern Concepts and Software. *Journal of Experimental Education, 62*.

Powers, D., Burstein, J., Chodorow, M., Fowles, M., & Kukich, K. (2001). *Stumping E-Rater: Challenging the Validity of Automated Essay Scoring*. Princeton, NJ: GREo. Document Number)

Rudner, L., & Gagne, P. (2001). An Overview of Three Approaches to Scoring Written Essays by Computer. *ERIC Deigest*.

Sowa, J., & Borgida, A. (1991). *Principles of Semantic Networks: Explorations in the Representation of Knowledge*: Morgan-Kaufmann.

Sutton, L. (2001). The Principle of Vicarious Interaction in Computer-Mediated Communications. *International Journal of Educational Telecommunications, 7*(3), 223-242.

Yang, Y., Buckendahl, C., & Juskiewicz, P. (2001). *A Review of Strategies for Validating Computer Automated Scoring*. Paper presented at the Midwestern Educational Research Association.

Zadeh, L. (1965). Fuzzy Sets. *Information and Control, 8*(3), 338-353.