12

ALWAYS ALREADY

Automated Essay Scoring and Grammar-Checkers in College Writing Courses

Carl Whithaus

INTRODUCTION

Although Ken S. McAllister and Edward M. White call the development of automated essay scoring "a complex evolution driven by the dialectic among researchers, entrepreneurs, and teachers" (chapter 1 of this volume), within composition studies the established tradition points toward the rejection of machine-scoring software and other forms of computers as readers. This tradition culminates in the Conference on College Composition and Communication's (2005) "Position Statement on Teaching, Learning, and Assessing Writing in Digital Environments," where the penultimate sentence succinctly captures our discipline's response: the committee writes, "We oppose the use of machine-scored writing in the assessment of writing" (789). If, however, we step back from this discourse of rejection and consider the ways in which a variety of software packages are already reading and responding to student writing, we begin to see that outright rejection of software as an assessment and response tool is not a viable, practical stand, because software is already reading, responding, and assessing student writing.

These "on the ground facts" of software's presence in students' writing processes range from the ubiquitous grammar- and spell-checkers in Microsoft Word to the use of Intelligent Essay Assessor to assess student knowledge in general-education courses. Once we acknowledge that software agents are intervening in students' composing processes, and that new more helpful, or more invasive, forms of software will continue to be developed, the questions facing writing program administrators and composition instructors transform from whether or not to use automated essay scoring and other forms of software to what types of software to use and how to incorporate these software features in effective and meaningful pedagogies for composition and writing-in-the-disciplines courses.

As a corrective to categorical rejections of software assessment and response systems, this essay examines the teaching and learning environments at Florida Gulf Coast University (FGCU) and at Old Dominion University (ODU). In the case of FGCU, Intelligence Essay Assessor, the latent semantic analysis based software is used to assess students' short essay question responses (350–500 words). At ODU, Microsoft Word is the default word processor used in open campus labs, the English department computer lab, and on many students' home computers. In both environments, software agents are part of the reading, responding, and evaluation processes for large numbers of undergraduates.

By analyzing writing activities at FGCU and ODU, we come to see that practices of using software as a tool for assessing and responding to student writing are already in place. The use of software agents as tools within students' writing processes, however, does not mean that students are not using these same digital writing environments as media for communicating ideas to their teachers. The Conference on College Composition and Communication's position statement justifies its rejection of machine-scored writing in terms of a dichotomy between human and machine readers, between what I have called software used as a medium for communication and software used as a tool for assessment or correction (Whithaus 2004). Stuart Selber's (2004) work on functional literacy, particularly his examination of computers as literacy tools, argues for a more subtle and nuanced reading of software and the multiliteracies within which students work. Further, the cases of software usage at FGCU and ODU suggest that in practice this either/or formulation does not correspond with the daily realities of students' composing processes.

In practice, software is used as both a medium for communication and as a tool for assessment and response. I am arguing for a conceptual shift within composition studies—if our practices combine software's functions as media and tools, then we need to reformulate our conceptions about machines reading and assessing students' writing. The tradition of rejection, reaching back to Ken Macrorie's (1969) critique of Ellis Page's work (Page and Paulus 1968), needs to be revised in favor of theories and practices of writing assessment that acknowledge the range of software's influence as responsive evaluative agents. Acknowledging this range will make it possible to evaluate the validity as well as the reliability of automated essay-scoring systems, not because the systems are valid in and of themselves, but because—drawing on Lee Cronbach's (1988) notion of validity as argument—the use to which the software agents or other forms of writing assessment are put are appropriate. For instance, the writing component on the new SAT exam is not a valid measure of a high school junior's or senior's overall writing ability, but it is a valid measure of how that student writes on a twenty-five minute timed, impromptu writing exam. Will this exam tell us all we want to know about incoming students' writing abilities? Hardly. But it does give a snapshot of a student's ability for one particular moment and for one particular form of writing. Predications based upon the writing component of the SAT, then, will be most accurate for this form of writing; the scores will have less validity as students move on to other, more complex writing tasks. Similarly, in carefully defined writing activities, software can be effectively used to assess short, close-ended responses from students, to quickly respond to surface features of student writing, and to offer the potential for students to develop metacommentary or reflection on the paragraph level.

SOFTWARE AS ASSESSMENT TOOL: INTELLIGENT ESSAY ASSESSOR AT FLORIDA GULF COAST UNIVERSITY

Looking at the use of Intelligent Essay Assessor at FGCU allows us to understand one context within which software could be used to assess short, close-ended written responses from students. Intelligent Essay Assessor is used to assess students' content knowledge and higher-order critical thinking skills through evaluating student writing in a general-education course, Understanding Visual and Performing Arts. This course is WebCT-based with large enrollments (380 in fall 2002, 560 in spring 2003, 541 in fall 2003, and 810 in spring 2004) (Wohlpart 2004b). In addition to assessing short written responses through Intelligent Essay Assessor, students' content knowledge is tested through multiple-choice questions and longer critical analysis essays read by preceptors, paid graders with bachelor's degrees in English.

To understand the impact of Intelligence Essay Assessor on student writing and learning, we need to consider the software within this context of multiple content assessments. Students are not only conscious of having a machine score their writing, they are also aware that a machine is scoring their multiple-choice answers and that human readers are grading their longer critical analysis essays. Students work within a continuum of multiple-choice tests, short essay question responses and longer critical analysis essays. The first two forms of assessment are evaluated by software and the third by a human reader. Students learn not only through video-streamed lectures and reviewing PowerPoint lecture notes but also by preparing for these three forms of assessment. They are given practice multiple-choice tests and analyze short essay question responses; they also develop Web board discussions about the sample essay questions and student responses to these questions. When it comes time for students to take exams with multiple-choice questions and short essay questions, they have already engaged in test preparation, learning activities for those forms of assessment. They are familiar with the concept of the computer as grader for the multiple-choice parts of their exams, and this concept is now extended to their short essay question responses—that extension is not likely to produce the alienation discussed by Anne Herrington and Charles Moran (2001) or the naìveté described by them in chapter 7 of this volume. In addition, the short essay questions on the course exams are contextualized within the course. The structure of the questions reflects the interweaving of course content and the rhetorical forms of students' written responses.

For instance, when a question on the Visual Arts exam asks students to "identify the element of form in Albert Paley's public sculpture *Cross Currents,*" the question is prompting students to focus on the concept of *form* in the visual arts. The second sentence in the question continues the focus on form as a key semantic quality by asking, "How does the form of the work create meaning or experience?" Finally, the third sentence in the question asks for a student interpretation or application of the concept of form in relation to both Paley's sculpture and the student's views: "What do you think this meaning or experience could be?" The short essay question is dictating the form of the student's response: (1) identify; (2) explain form in relationship to meaning or experience; (3) think about the meaning or experience in relationship to your own views. Concepts from the visual arts are invoked in the question, and the student must link these together in writing to demonstrate mastery of the concepts. The format of the writing is formulated in the question. The students must take part in a particular "genre" of writing—the short essay question response—that is not uncommon in high school and college courses. Based on the form of this genre, and its narrowed definition within FGCU's Understanding Visual and Performing Arts, Intelligent Essay Assessor scores the ways in which students link the relevant ideas from the course together.

Jim Wohlpart describes the students' written responses to the short essay questions as ways in which they demonstrate a greater mastery of their knowledge of the course content and apply higher-order critical thinking skills than they do in the multiple-choice questions (2004a). The short essay question responses are also not the end of the student's writing activities in Understanding Visual and Performing Arts; rather, they are part of a learning and assessment continuum that moves from multiple-choice questions to open-ended, individualized critical analysis essays. As Wohlpart readily acknowledges, these longer critical analysis essays could not be effectively scored by Intelligent Essay Assessor. Rather than directing students to apply course knowledge in a short, relatively controlled form, students are allowed to analyze an artwork or performance of their choice. Because of their complexity and their variability, these open-ended critical analysis essays cannot be scored effectively by essay-scoring software. Wohlpart compares these critical analysis essays to the types of assignments he gives in first-year composition courses. The students need to write multiple drafts and explore concepts discussed in the class in intimate detail. The current version of Intelligent Essay Assessor would be no more appropriate for assessing these critical analysis essays than WebCT's multiple-choice scoring mechanism would be for scoring the short essay question responses. According to Lee Cronbach's (1988) concept of validity as argument and Huot's (2002, 53–56) development of that concept in composition, for a writing assessment to be valid, not only does the scoring mechanism need to be valid but the use to which the results of the scoring are put needs to be valid and appropriate as well. Using automated essay scoring to score the critical analysis essays, or the types of individualized, open-ended essays written in Wohlpart's first-year composition courses, would make the assessment system invalid. When I argued at the beginning of this essay that composition researchers and teachers need to step back from a discourse of rejection, it is in order to make these finer and more accurate distinctions among types of software and their uses.

When software is used as a tool for assessment or response purposes, we need to decide whether the use of that tool is valid. We need to ask: how does the software tool function? Is it accurate for its claimed purpose? And, are the results of the assessment put to valid use within the larger course or institutional context? When software is used as a tool for assessment, response, or revision, it is not necessarily opposed to effective composition pedagogies. While students do need to use software as media to communicate with each other and with their instructors to improve their composing skills, the use of software as a medium for communicating does not exclude the use of software tools as prompts for sentence-level or paragraph-level revision or as an assessment device for content knowledge. Understanding the context within which Intelligent Essay Assessor is used at FGCU provides us with a wider scope within which to evaluate software's influence on student writing and our pedagogies.

At FGCU, Intelligent Essay Assessor is used as a tool, not as a medium for communication. That is, the multiple-choice questions graded by WebCT and the short essay question responses graded by Intelligent Essay Assessor are software agents as tools. When the students submit critical analysis essays for the preceptors to grade, then WebCT and the word-processing software are being used as media for communication. These distinctions are important, because in both the multiple-choice and the short essay question responses the knowledge that is being tested is close-ended and containable, but in the communication-based critical analysis essays the subject matter, what piece of visual or performing art is analyzed, as well as the rhetorical techniques used to create an effective analysis vary from student to student, situation to situation. Still the question remains: is Intelligent Essay Assessor's evaluation of the short essay question responses about writing? To say that having students write within a very specific format is a high-end way of assessing content knowledge and critical thinking strategies is not the same thing as saying that these short essay questions teach the students how to become better writers.

SOFTWARE AS RESPONSIVE TOOL: MICROSOFT WORD AT OLD DOMINION UNIVERSITY

By analyzing how Microsoft Word's grammar-checker and readability features are used in writing courses at Old Dominion University, we will see a narrower example of software used as a responsive tool for improving student writing. Unlike the use of Intelligent Essay Assessor as an assessment tool and WebCT and word processing as media for communication at FGCU, the use of Microsoft Word at ODU combines the functions of software as tool for correction and evaluation and software as a medium for communication in a single software package. If it was important for us to see the use of Intelligent Essay Assessor at FGCU as occurring within a continuum of assessment tools, it is also important for us to recognize that there is a range of software tools for assessment, scoring, and response to student writing. Automated essay-scoring software does not stand alone, especially from students' perspectives. These software packages, particularly when used as described by Jill Burstein and Daniel Marcu (2004), in classroom settings do not exist in isolation for student writers from more mundane, common software tools such as Microsoft Word. Understanding the use of Microsoft Word's grammar-checkers and readability features in ODU writing courses helps us get a better picture of what it means to use Intelligent Essay Assessor as an assessment agent at FGCU; the impact of grammar-checkers and readability features on composition pedagogies makes acknowledging the fuller range of software's influence on writing instruction possible.

Writing instruction at ODU involves three writing courses for all undergraduates: first-year composition, a second-semester composition course or a science and technical writing course, and a discipline-specific writing-intensive course at the junior or senior level. In addition, the English department offers courses in advanced composition, technical writing, management writing, and a variety of journalism and creative writing courses. Over two thousand students are enrolled each semester in the first- and second-semester composition requirements. In this panoply of writing courses the default word-processing program is Microsoft Word. Eleven percent of the sections meet at least twice a semester in the English department computer lab for writing workshops, hands-on activities in Blackboard, or Web-based research assignments. While the students use Microsoft Word during the writing workshops, most of the instructors do not explicitly address how to use or respond to Microsoft Word's grammar-checker. It is common for instructors to advise students not to blindly trust the grammar-checker; however, more detailed discussion of Microsoft Word's green squiggly lines are not a required part of the curriculum and often do not occur. To be able to explain when to follow Microsoft Word's advice and when to ignore it requires an understanding of both grammatical concepts and software's (mis)application of these concepts. To further complicate matters, it is not only instructors but also the students who need to understand these issues. Within a labor system where 98 percent of the courses are taught by graduate students, adjunct faculty members, or lecturers, the time to focus on grammar and software's application of grammar does not exist. The general sentiment is that composition instructors are teaching writing, not word-processing skills or software usage.

What is funny is that the interface of the word processor, particularly Microsoft Word, is so prevalent in writing instruction at ODU, yet it is infrequently addressed or discussed as an explicit class lesson. The tool exists, but writing instructors are more interested in the computer as a medium through which their students communicate rather than as a tool for correction. Yet, for students, and even for teachers, Microsoft

Word's green squiggly lines often interrupt or at least influence the writing process. Very few go into Word's Tools > Spelling and Grammar menus and deselect the "Check grammar" box. The software is a low-level reader of form and a response agent, but it is untheorized in composition studies and unaddressed, and perhaps underutilized, in our pedagogies.

While it is possible, and useful, to critique Microsoft Word as "the invisible grammarian" as McGee and Ericsson (2002) have done, another response would be to run into the teeth of the machine. In other words, by working with the features in Microsoft Word such as readability and by explaining to students exactly how the grammar-checker works in terms of their language, we make the students' interaction with the software into teachable moments rather than rote acceptance of the software's authority.

For instance, in a junior-level technical writing course, I had students use Microsoft Word's readability feature as a tool for paragraph-level revision. On Blackboard's discussion board, I asked students to:

> 1. Select a paragraph from your proposal or from your current draft that you would like to rewrite. Paste that paragraph into the discussion board space. 2. Score that paragraph according to reading ease and grade level using MSWord. Paste that paragraph into the discussion board. 3. Revise that paragraph. Score the revised paragraph according to reading ease and grade level in MSWord. Paste that paragraph into the discussion board. 4. Explain why you think the revised paragraph had "better" scores. (Or if the revised paragraph did not have better scores, explaining why you believe it is more effective despite the readability and grade-level scores.)

In this assignment, the software is a response agent to encourage revision. The assignment also contains a prompt to respond to the software by developing metacommentary about the revised paragraph and the software's reading of that paragraph. A student who was working on a technical report about the future of U.S. space exploration posted the following material:

> UNEDITED: The security of our nation domestically, internationally, and economically will be ensured through research and developed skills that may help detect and deflect asteroids that may threaten Earth. Since U.S. military strength and economic security rests on our technology leadership, implementation of the space exploration vision will drive technology

> related disciplines such as medical research, biotechnology, computing, nanotechnology, composite manufacturing, and many others. The report presents the argument that as international leaders, the U.S. that should forge ahead into space exploration, rather than sitting idly by. This competitiveness will require a skilled workforce, and the space exploration vision will work to create a needed re-focus on math and science education in the United States.
>
> EDITED: Our nation's domestic, international, and economic security will benefit from the research and skills developed through space exploration. For example, we may discover a way to detect and deflect asteroids that threaten Earth. Since U.S. military strength and economic security rests on our technology leadership, implementing the space exploration vision will drive technology related disciplines such as medical research, biotechnology, computing, nanotechnology, and composite manufacturing. The Commission's report poses the argument that as international leaders, the U.S. should forge ahead into space exploration rather than sitting idly by. This competitive approach will necessitate developing a skilled workforce. Thus, the space exploration vision will work to create a needed re-focus on math and science education in the United States.
>
> CONCLUSION: The first paragraph scored a 5.0 for ease of reading and earned a rating of grade level 12. I revised the paragraph by removing some passive sentences and nominalizations. It then scored an 11.5 for ease of reading and remained at the grade level 12 (although I got a 0 percent for passive sentences, down from 25 percent). Since the reading ease scale calculation utilizes average sentence length and average number of syllables per word, a piece with longer sentences and bigger words, such as a technical piece with scientific wording and terminology, will score lower and earn a higher grade level rating. I found that the best way to increase readability was to break long sentences into shorter ones and to make sure there were transitional phrases, such as "however," or "thus," to improve the logical flow of the information.

Her commentary is fascinating because it shows an attention to the stylistic details of sentence length and number of syllables. She articulates an awareness of the software's limitations for scoring "a piece with longer sentences and bigger words, such as a technical piece with scientific wording and terminology" and is still able to use the software to increase the readability of her report.

Although this student was able to implement the changes suggested by Microsoft Word's readability scoring, others were resistant to the software. One student submitted three different paragraphs. Her metacommentary is worth quoting: “Well, finally! The revisions used shorter words and shorter sentences in order to increase the reading ease score and lower the grade level. I am not sure, however, if I will keep this paragraph in my presentation. The second one does flow better than the first, but the third one seems a little too “dumbed down” to me. Maybe I have been in college too long . . .”

She is ranking samples of her own work, using Microsoft Word’s scores as one filter and her own sense of audience as another filter. Another student, whose views I believe would be echoed by many composition instructors and researchers, wrote, “Although I do agree that concise writing is more effective, I think this method of scoring readability is too simplistic.” On one level this student is surely correct—if writing teachers were only to take Flesch Reading Ease and Flesch-Kincaid grade-level scores into account when judging student writing, then those writing assessments would be far too simplistic. However, when students are communicating multiple complex ideas and using software as both media and tools, then the use of simplistic readability scores as useful abstractions in order to help students see their writing through a different screen becomes more appropriate. In first-year composition courses, in the writing of the analytic essays at FGCU, and in this technical writing course at ODU, I would suggest that the limited use of software as an assessment and response tool is valid and appropriate.

CONCLUSION

Microsoft Word can work as a tool, as a prompt for revision on the sentence level or the paragraph level. Within a sequence of assessment tools, Intelligent Essay Assessor can function as a device for building higher-level critical thinking skills and testing content knowledge. In both cases, the software is a tool, not a medium. However, the ultimate goal of the writing activities is a communicative agenda that involves using software as media for communication as well. I would respectfully want to argue that the Conference on College Composition and Communication’s committee on Teaching, Learning, and Assessing Writing in Digital Environments has made a mistake by continuing composition studies’ tradition of rejecting software as a reader, responder, and assessor of student writing. The uses of software as tools within courses at FGCU and ODU suggests that, as contextualized tools, there are uses for Intelligent Essay Assessor and Microsoft Word as readers, responders, and assessors. What composition studies needs is not a blanket rejection of these systems but rather data-driven studies of how these different software agents are already being used in postsecondary writing courses. When software works well for a particular task, writing researchers should build pedagogies that incorporate these features. When the use of software produces decontextualized, invalid writing assessments, writing researchers need to point out the faults of these systems. In the end, a blanket rejection of automated essay scoring and other forms of software as readers does not serve composition teachers or students; a more nuanced, situation by situation consideration of how software is used and its impact on writing pedagogy provides a clearer picture of the challenges facing teachers and students.