Writing to a Machine is Not Writing At All
Anne Herrington
Charles Moran
University of Massachusetts Amherst

Ed White has been, as others in this volume and elsewhere have said before us, a giant in the field of writing assessment. In the large corpus of his work on writing assessment he has laid down principles for judging assessment methods and practices, principles that we have found guided our own work in a subset of this field, the machine scoring of student writing.

In the early 1980’s, as McAllister and White have written (16-17), researchers in the field of natural language processing lost their Defense Department funding and turned instead to testing companies to support their work. Testing companies, in their turn, have widened their marketing focus to include both large-scale assessment and classroom-based assessment. White’s writings about computer-based assessment of writing have generally addressed the pros and cons of large-scale writing assessment, used for purposes of placement. In the chapter that follows, we take White’s principles, as articulated in the body of his work on writing assessment, and apply them beyond large-scale assessments used for broad placement purposes to the marketing of these programs as aids to classroom instruction—an application that White seems not to have considered, perhaps because it so obviously violated his principles. White has asked us, eloquently and often, to pay attention, to stay in the mix, to be part of the dialogue (1994, pp. 8, 20; 1996, pp. 11, 23-24; 2006, p. 27). Given the expanding use of the computer-as-reader of students’ classroom writing, writing teachers need to understand what is happening here and take a careful look at its substance and likely consequences, lest we be seen as irrelevant and be “sent out of the room” by the other stakeholders. (White 2006, p. 27).
Marketed for the Classroom

We focus on ETS’s Criterion® although similar claims are also made in marketing Vantage Learning’s My Access!®, both of which are designed and marketed explicitly for classroom instruction. (See Rothermel, 2006, for a review of My Access!) Specifically, ETS’s Criterion is described as a “comprehensive, instructional web-based tool,” aimed to assist both students and teachers (2010). As with all commercial products, these programs are successful only if they meet a perceived need and are purchased. This is the complicity that McAllister and White (2006) write of between “entrepreneurs” and “adopters.” They write, “some teachers and administrators turn reflexively to technological solutions when funding for human labor is in crisis, as has been the case for education at all levels in recent years” (25)—in other words, when there are too few teachers for too many students. ETS targets their marketing on this issue directly, promising, among other things, that Criterion is “a learning tool that does not increase workload and adds value to writing instruction by providing a teachable moment.” The nature of that “teachable moment” is not specified. Students are promised instant feedback that will directly benefit their development as writers:

Students get a response to their writing while it is fresh in their minds. They find out immediately how their work compares to a standard and what they should do to improve it. The Criterion service also provides an environment for writing and revision that capable and motivated students can use independently. This environment, coupled with the opportunity for instant feedback, provides the directed writing practice so beneficial for students.

Essentially, the Criterion materials are promising to address the very real conflict that many teachers face between, on the one hand, the desires to have their students write frequently,
provide them with meaningful feedback, and have them improve as writers and, on the other hand, the problem of having too little time to provide that feedback and individual attention to all of them. The Criterion materials also evoke the discourse of accountability, noting a “standard” against which student work will be compared and promising that the “service provides an effective initial evaluation of student skill levels allowing instructors to benchmark writing, to make placement decisions, adjust instruction and track progress.” If all of these claims are true, then no wonder that these programs would be widely adopted in classrooms at all grade levels, including college. A healthy skepticism about these and other claims for these programs is still warranted, though. Just as McAllister and White (2006) point out in their discussion of “entrepreneurs,” these products are marketed as “a proven, rather than an experimental, technology” (24). Aside from research conducted or sponsored by the companies themselves, there has been insufficient analysis of the claims made in the marketing, specifically claims as to the soundness of the feedback provided by these programs and the impact on students’ development of writing skills. As McAllister and White (2006) argue, “the time has come for reasoned and critical examination” (25). Our aim in this essay is to contribute to this goal by examining the soundness of the feedback provided.

A Scenario

Let’s imagine that we two, Anne and Charlie, are teachers who are considering the use of Criterion in our writing classes, college classes in Advanced Expository Writing taken by sophomores from a range of disciplines. In our scenario, we are considering this course of action not because we really like the idea but because we have been given 35 students in each of our classes, these numbers, in our judgment, making it impossible for us responsibly to read the several drafts of each of the seven required writing projects assigned in the course. Criterion
gives us two options. We could persuade our institution to subscribe to the Criterion Online Writing Evaluation Service, which would give us and our students one semester’s access to a web-based package that includes not only faculty and student access to the machine-scoring programs but writing-class-management tools such as record-keeping and sorting capabilities. The Criterion web site tells us that the cost of this option to our institution would be $15.20/student, or $1,064 for our two classes of 35, and in addition $300 for a mandatory on-line training program that we as the responsible teachers would have to take before the course began.

Our second option is to have our college bookstore order a sufficient number of Criterion Registration Cards and in our syllabus tell our students to buy these cards at the bookstore along with their textbooks. ETS sells Criterion directly to college bookstores, which it charges $10.36/card; the bookstores may charge the students the ‘suggested retail’ price of $12.95, for a 24% markup, though bookstores are free, according to the terms of the contract, to charge as much as $15.00/card. This way, as the web site tells us, our students would get “practice made easy and convenient….at no additional cost to you!”

The advantages of having an institutional subscription to the Criterion Online Writing Evaluation Service are two: we would be working within the web-based class management system, which would make the students’ writing easier to organize and track; and we would have greater control of our students’ use of the machine-scoring program. The first of these advantages disappears for us, because our institution already has a class management system supported by our IT center. The second of these advantages seems slight, since the additional control we’d have would be in turning off some or all of the automated feedback that the program gives to our student writers—and that feedback is all the program really has to offer us. So the bookstore option seems the best to us.
But in either case, whether we lobby for the institutional subscription or require our students to purchase their *Criterion* cards, we need to know that the use of *Criterion* will advance our goals for our students’ learning. As Ed White has reminded us, “assessment of writing can be a blessing or a curse, a friend or a foe, an important support for our work as teachers or a major impediment to what we need to do for our students”(1994, 3). So the questions we have to answer are these: “Will the assessment and feedback given by the program to student writers be helpful to them as writers? Will it support our work with them as their teachers? Or will it become something we have to work around, or work against—an impediment to our teaching and our students’ learning?”

Now we leave the world of our hypothetical situation and report on a recent trial session we conducted to determine the nature and value of *Criterion*’s feedback.

**Testing Criterion**

In January of 2010, in a trial session provided us by ETS, we set up a “class,” English 112, and a “student”, Carol Santos. We posted an assignment, selecting College, Second Year as our level, and chose this topic from the list provided:

> “Educators should encourage the current popularity of email and instant messaging among teenagers, because this trend is sure to improve their students’ reading and writing skills. Discuss the extent to which you agree or disagree with this opinion. Support your views with specific reasons and examples from your own experience, observations, or reading.”

Charlie wrote a response to this question, arguing, as he believes, that educators should not consider encouraging or discouraging email or texting for two reasons: educators don’t have the ability to affect their students’ behavior outside of their classrooms and schools, particularly in
the use of such attractive social networking tools; and that the effect of email and texting on extended reading and writing was unknown but likely neutral. For the full essay, please see Appendix A. *Criterion* gives the writer two scores, a “holistic” score on a scale from 1 to 6, and a “trait analysis,” with separate feedback on five traits: grammar, usage, mechanics, style, and organization/development. Charlie received a holistic score of 5, which pleased him—but why not a 6? The holistic feedback was not very helpful in answering this question: the difference between a 5 and a 6 seems based on judgments an automated program cannot make: for example, “clear” versus “insightful connections” and “relevant” versus “well chosen evidence.” The “trait feedback analysis” gave him the answer: two grammar errors, 15 mechanical errors, two usage errors, and, under organization/development, a missing introductory paragraph. Taking these in order, *Criterion* identified two errors in “grammar.” The program marked the following sentence as a fragment:

“If the analogy between cycling and reading/writing holds, then training on email/texting will not transfer to extended reading or writing.”

The program marked the following sentence as run-on:

“I’ve encountered more than one student who is a whiz on email, belonging to multiple discussion groups and contributing thoughtfully to these discussions, but who is entirely unable to put together an extended argument.”

From this response we conclude that *Criterion* cannot reliably identify sentence boundaries. *Criterion* identified two errors in “usage”: one “missing or extra article” and one “preposition error.” The program identified this sentence as containing a “missing or extra article,” and we bold-italicize the allegedly-offending article:

“I take as an analogue here cell-phone use and driving.”
Charlie supposes that one could write “I take as analogue here,” but he resists *Criterion’s* suggestion, “You may need to remove this article.” This is, to us, not a matter of usage, but of style.

In the following sentence, *Criterion* identified the word *in* as a “Preposition error.” Rolling over the marked alleged error produced the pop-up, “You may be using the wrong preposition.” Here is the sentence, with the preposition in question in bold italics:

“As a cyclist, I know that cycling is an activity that includes many different activities: long road-rides, hill-climbing, descending, riding close *in* groups (peloton), time-trialing, and sprinting.”

There is no problem with this preposition. A peloton is riders riding close in a group.

In Charlie’s essay *Criterion* identified fifteen errors in “mechanics,” all alleged spelling errors, and all wrongly identified as such. Charlie used the word “texting” nine times; *Criterion* identified each of these uses as a mis-spelling. Charlie used the abbreviation *e.g.* three times, and again each use of this common term was identified as a mis-spelling. Charlie used the word *peloton*, identified as mis-spelled, and when he used a dash, the two words connected by the dash were identified as a spelling error.

Charlie’s essay received no feedback on “style.” Under organization/development, however, *Criterion* told Charlie that he needed to add an introduction to his essay. *Criterion’s* comment was “No introductory material detected,” and when, as instructed, he ‘rolled over’ this comment, *Criterion* amplified: “*Criterion* has not identified an *introduction* in your essay. An introduction is very important because it provides an overview of the essay, provides background information on the subject, and tells the reader what the author will be discussing. For advice about adding an introduction to your essay, consult the Writer’s Handbook.” Charlie tried a
second version of his introductory paragraph, this time adding a topic sentence at its end, but *Criterion* came back with the same response. Apparently inductive paragraphs (specific to general) don’t fit the model.

So how did *Criterion* do overall with its feedback on the five traits? It gave Charlie 20 pieces of feedback, 19 of which were wrong, and one of which was arguable. It misidentified a sentence fragment, it misidentified a run-on sentence, it gave bad advice about articles and prepositions, it marked as spelling errors fifteen perfectly-good words; and it told Charlie to add an introduction which the essay didn’t need. If we knew that we had in our Writing Program a teacher who mis-advised students as drastically as *Criterion* has mis-advised Charlie, we’d fire that teacher.

Imagine the effect of this 95%-wrong feedback on the student writer and on the teacher. The feedback would at the least confuse the student writer, leaving the teacher somehow to counter the confusion—although if the student were using her *Criterion* card on her own, as purchased from the bookstore, there would be no teacher to intervene. If the student accepted the feedback, here are some of the lessons that would be learned: do not use *e.g.*, or *texting*, or *peloton*, in any of your writing; do not use the dash as a mark of punctuation; shorten and simplify all sentences so that the program will be able to parse them accurately; do not use inductive, specific-to-general, sequences, but stick with deduction—topic sentence first. Among our goals as writing teachers are these: help students discover and use their voices; help them take risks with their writing; help them master the grammar, usage, mechanics, and styles of written English. In this trial, and in earlier trials we have reported on (see Herrington and Moran 2009, Herrington and Stanley), *Criterion* has proved not a useful assessment tool but, to quote Ed White again, “a major impediment to what we need to do for our students” (1994, 3).
Skills, not holistic focus

Ed White has argued steadily throughout his long career that writing needs to be read whole and not disassembled into a set of discrete parts. In 1984 White wrote in defense of holistic, as opposed to analytic, scoring of essays, “Holisticism says that the human spirit and its most significant form of expression (writing) must be seen not in parts, but as a whole, face-to-face as it were, almost sacramentally. Even the meanest bit of halting prose, even the most down-trodden of our fellow creatures, deserves to be taken as a living and vital unit of meaning, an artistic and human whole, not merely as a collection of scraps and parts” (409). Our experience with Criterion, however, tells us that machine-reading programs focus their feedback on the parts: word- and sentence-level errors in grammar, usage, and style.

Indeed, the Criterion online demonstration represents this focus on isolated skills—the trait feedback function—as “the key reason why Criterion is such a valuable remediation tool” (2010). As we have shown above, not only does this feedback focus on isolated skills; it is often wrong. So we proceed with two parallel theses here (doubtless Criterion would object!): that the program’s feedback calls the writer to attend to word-level and sentence-level concerns; and that the program’s feedback is sometimes—apparently most of the time—monstrously wrong.

After they have written and submitted their essay to Criterion, students will view and work with the analysis the program provides, calling up each category, e.g. “Summary of Grammar Errors.” On the screen, a left hand sidebar then lists all possible types of errors in grammar and in the center of the screen a bar graph displays the number of each type of grammar-error that has occurred, showing, for example, a short bar for 1 fragment or missing comma and a taller bar for 4 subject-verb agreement errors. The presentation, with numbers and graphs, gives an aura of objectivity and mathematical power to the analysis. Further, it presents
the revising task to come as one of moving through categories and eliminating errors, not of re-
thinking the argument or adapting to a particular rhetorical situation.

After seeing the bar graph of all errors for a trait, students can call up each type of error, one by one: for example, fragment or missing comma. The center screen then shows the full essay with the error highlighted. Roll-overs then explain the error in a very limited way (e.g., “You may be using the wrong preposition”). The task for revision is evidently to reduce the flagged errors. For example, the following sentence from one student essay that we entered for analysis was flagged for two alleged problems: “fragment or missing comma” and “missing or extra article” (before “different”):

“Some of my bracelets are different color of crystals; jade; silver and metals where she got it from China.”

The explanation regarding the sentence punctuation is: “This sentence may be a fragment or may have incorrect punctuation.” While this feedback may be accurate (though note the program’s steady hedging with the subjunctive), it seems a misplaced priority, focusing as it does on punctuation, rather than directing the writer’s attention to the complexities of the final clause— which is not flagged at all. For the flagging of the article, the explanation is “You may need to use an article before this word. Consider using a.” Well, if anything, the change should be to make “color” plural, “are different colors of crystals.” In another place in the same essay, the following sentence was flagged as a run-on sentence:

“Most of the girls have at least one jewelry box, I am no exception to this rule.”

Yes, grammatically speaking it is a run-on or, to be precise, a comma splice; but rhetorically speaking, the sentence could be viewed as an effective construction, syntactically setting up a comparison of this writer, “I,” with “most girls.” For Criterion, however, this is a sentence with
an error, not a sentence seen as the result of a writer’s choice. In this same sentence, Criterion flagged jewelry box as an error in the category “Style,” and provided this feedback: “You have repeated these words several times in your essay. Your essay will be stronger if you vary your word choice and substitute some other words instance. Ask your instructor for advice.” In our experience, this is the most frequent comment generated under the category “Style.” Again, the flagging and the feedback draws the student writer’s attention to word-level concerns. Further, as we’ve so often experienced, the feedback is dead wrong: In this instance, the essay was about a “jewelry box,” and as readers we require the repetition of key words. Here again, Criterion considers an aspect of language use—in this case repetition of words—in terms of what it can count—the number of times a word appears in an essay—as opposed to what writers do: repeat words for rhetorical effect.

Criterion and Many Englishes

A further impediment to our teaching and our goals for our students’ learning is Criterion’s bias toward Standard American English and its inability to respond to the ways in which our language is growing and stretching as it is used by students who come at English as a second or third language. As Herrington and Stanley show in “CriterionSM: Promoting the Standard,” Criterion is programmed to value only one dialect of English, Standard American English. Throughout the Criterion materials, “the standard” is repeatedly invoked as the norm against which students will be judged. For example, in response to the one of the Frequently Asked Questions posed on the website, “How can the Criterion service help students?,” the response is: “Students get a response to their writing while it is fresh in their minds. They find out immediately how their work compares to a standard and what they should do to improve it. . . .” That standard is based on two sources. For the holistic judgment, the standard is based on essay rankings from ETS
testing programs like NAEP®, the English Placement Test designed for California State University, Praxis®, GRE®, and TOEFL®. Criterion’s Grammar, Usage, and Mechanics analyses are all error-focused, programmed to read for errors in standardized American English, represented as “general English grammar,” the corpus used for identifying these “violations of general English grammar” being 30 million words of newspaper text (Burstein, Chodorow, & Leacock, 2004, p. 28).

We have no quarrel with a pedagogy that aims to introduce students to standardized academic American English so long as that dialect is recognized as one dialect of English and so long as other dialects are recognized and valued. With Criterion, however, this “standard” is normalized and deviations from the standard are recognized as error—for instance, dialectical differences in verb tenses and ways of organizing and developing ideas (Herrington and Stanley). Because Criterion is programmed to recognize only standardized American English, it also has difficulty analyzing writing by students for whom English is not their native language. When we submitted to Criterion an essay by a student from China and one by a student from Vietnam, we found the program had difficulty correctly identifying article and preposition errors—most unhelpful to students coming at English from Asian languages. Further, as we have suggested in examples above, the program conveys the wrong priorities: focusing attention on surface features of grammar and form instead of substantive development and rhetorical effectiveness, and failing to be able to recognize effective use of language and syntax. Imagine the confusion an English language learner would experience receiving the kind of feedback we have reported.

Writing for Nobody: The Game

Even if Criterion and its competitors could be improved—and given Criterion’s 95% failure rate we experienced on Charlie’s essay, above, they have a long way to go—they should
still be banished from our classrooms, for, as Ed White has long argued, “Writing for nobody is not writing at all” (1969, 167). Machine-scoring programs, in common with skill-and-drill curricula, “deny the humanity of writing. . . . Let us find honest assignments—they are everywhere to be found, in experience, in books—and help students to see writing as a valuable way to gain important knowledge and to define their own thoughts for themselves and for others” (1970, 869). White has called us “to resist and protest against the dehumanizing effect of materials and essay assignments that turn writing into academic gamesmanship. . . . We can support the humanity of our students as writers by insisting on our own humanity as readers. We can refuse to assign and decline to accept writing for nobody” (1969, 168).

From our experience writing essays for Criterion and using their packaged prompts, writing for Criterion does become “writing for nobody.” Criterion and its competitors “turn writing into academic gamesmanship” (White, 1969, 168), a game of learning to write the kind of essays that are used for ETS testing programs since they are the source for the prompts used by Criterion in order to provide the norming capability for the holistic ratings. Part of the sense of game comes from the time limits imposed on the task. If one wishes to use the holistic feedback rating, The Criterion Teaching Guide advises setting a time limit for writing: “Note that it’s very important to assign the same amount of time for writing the essay from instructor-generated and modified prompts as is required by the standard higher education prompts” (ETS, 2007, 28). In another of our trials with Criterion, Anne experienced this pressure when she was writing an essay for Criterion. She was just beginning the first sentence of her third paragraph when a blinker told her she had only three minutes left. Even though she wasn’t sure how she wanted to finish, she hurried to make up a conclusion, taking the easy way out and dropping in a maxim.
There is the game of writing against the clock, followed by the game of revising to get a higher score. Once Anne submitted her essay, she received the promised immediate rating: in this case a 4 with 6 being the highest score. It became clear by looking at the sample, scored essays provided by *Criterion*, that an essay with an agree/disagree thesis and at least three main points—not just the two in her essay—was favored. Knowing that was the game, Anne easily improved her score to a 6 by adding what appeared to be an agree/disagree thesis and adding an additional paragraph with a simplistic example. Her focus was not on what she was aiming to get across—indeed, the revised essay did not represent her view, rather on adding the favored structural features.

The title and substance of White’s 1969 essay, “Writing for Nobody,” remind us of what a student said to us after discovering that her placement essay was going to be read and scored by WritePlacer Plus, and not by a teacher. Surprised and even a bit disturbed, she said,

I’m assuming that someone is going to be reading this thoroughly and, you know, thinking about what you wrote, and not really—I mean I know the computer is smart, but I’m not thinking of them as thinking about what I’m writing as the way a human would think about what I’m writing. (Herrington and Moran 2006, 123)

This was a student from a local community college that had adopted WritePlacer Plus for its placement testing and was considering bringing the program into its full curriculum as an aid in its tutoring center and as a pre- and post-test to measure the value added to a student by the two-year course of study.

We note from the *Criterion* website’s long list of institutions using the program that those writing for *Criterion* and, by extension, WritePlacer Plus, are predominantly two-year colleges. There are no Ivy-League names on the list. This raises for us the specter of a two-tiered field,
divided by wealth and connection, where poor and unconnected students write to machines, and wealthy and connected students write to human readers. We understand the very real financial pressures some schools face, pressures that lead them to seek technologies that promise efficiencies of time and cost. But at the same time, we know, as Ed White has said before us, that writing to a machine distorts the very nature of writing itself. Writing to a machine is writing to nobody. Writing to nobody is not writing at all.

References


Appendix A
Essay Written to Criterion Prompt

This assignment makes two assumptions that I have difficulty with. First, it assumes that educators can have an effect on students’ use of email and texting. Second, it assumes that there is a powerful link between students’ use of email/texting and “students’ reading and writing skills.”

The first assumption, that we as teachers can affect our students’ use of email and texting, is difficult to support. We probably don’t want our students to be emailing/texting during our lectures and classes—certainly I don’t. And we don’t have much control over what our students do when they leave our classrooms. In addition, email/texting are tremendously gratifying social media; it is unlikely that educators, or anyone else, can influence their use. I take as an analogue here cell-phone use and driving. We know that cell-phone use distracts drivers and leads to horrific accidents. And legislators and others are, for now at least, entirely unable to control this use of communication technology.

The validity of the second assumption, that there is a close relationship between the use of email/texting and “reading and writing skills,” depends upon the definition of “reading and writing skills.” I’ll assume that by these terms we mean extended reading (e.g. essays, New York Times-length articles, novels) and extended writing (e.g. thousand-word expository essays that are reflective and that maintain logical coherence over their full length.) It seems to me that the ability to compose and read in short bursts is not the same as the ability to compose and read extended documents. I’ve encountered more than one student who is a whiz on email, belonging to multiple discussion groups and contributing thoughtfully to these discussions, but who is entirely unable to put together an extended argument. On the basis of my own experience, both intellectual and athletic, training is specific to the performance task: that is, you learn to do what you practice, and there is very little transference to tasks that may seem related. As a cyclist, I know that cycling is an activity that includes many different activities: long road-rides, hill-climbing, descending, riding close in groups (peloton), time-trialing, and sprinting. To train for one of these, e.g. a century (long road ride) is not training for sprinting, or descending, or any of the other ‘kinds’ of cycling. If the analogy between cycling and reading/writing holds, then training on email/texting will not transfer to extended reading or writing.
So where does that leave us? First, educators can’t affect students’ use of email/texting outside of their own spheres of influence, which are limited to classrooms or, at the most, school buildings. And second, there’s very little likelihood that practice on email/texting will have a positive effect on the very-different skills of extended reading and writing. So, to return to the question assigned, educators should not concern themselves with students’ use of email/texting. If they try, it will be a waste of their time and energy.