*by Lorna Collier*

# Robo-Grading of Student Writing Is Fueled by New Study—But Earns "F" from Experts

*Whether at K-12 or in higher education, shifting to automated assessments of student writing holds many potential perils.*

Can computers grade student writing as well as human teachers can? A new study says yes, and the results are being used by educational testing companies to further market automated grading systems.

Already, such programs have made inroads in schools, and are expected to proliferate further with the advent of online Common Core State Standards testing in the next couple of years. Using so-called "robo-graders" is touted as a way to grade tests more quickly and cheaply than using humans—a big allure for budget-crunched educational institutions.

Yet the e-rating emperor has no clothes, say English professors who have examined such programs. All too often, robo-grading programs are flat-out wrong, marking words and passages as incorrect when they aren't, or giving the highest grades to nonsensical gibberish.

Les Perelman, director of writing programs at the Massachusetts Institute of Technology, created and tested 17 student essays against Educational Testing Service's e-rating program. The program identified errors in his prose that were not errors and credited him with writing well when he copied random, disconnected chunks of text into an essay. The e-rater also seemed enamored of $10 words—"egregious" instead of "bad" or "plethora" instead of "many"—giving him higher scores when he used these words.

"Even if you use 'egregious' incorrectly—I said, 'life is egregious'—it likes it, because the computer is stupid," says Perelman, explaining that the program works not by divining the meaning in words but by counting them and comparing them to lists of infrequently used words.

Similarly, professors Anne Herrington and Charles Moran of the University of Massachusetts at Amherst tested ETS's Criterion program and found that 19 of 20 pieces of feedback on one essay were wrong and one was arguable. The program marked as misspelled 15 correctly spelled words, misidentified a run-on sentence, and misidentified a sentence fragment. (Their findings appear in *Writing Assessment in the 21st Century*.)

Tim McGee, associate director of faculty development at Rider University, tested another computer grading program, the Intelligent Essay Assessor. He created gibberish sentences (but included keywords related to the essay prompt) and earned high marks. (Example: "To effect the detects that Mr.stroke McGeorge had I would several conduct experiments testing ability his communicate to.") McGee reported his findings in a 2006 collection of essays about computer grading, *Machine Scoring of Student Essays: Truth and Consequences*.

Despite these and many other examples of faulty grading by robo-graders, a new study by the University of Akron's dean of education, Mark Shermis, shows that when compared with human graders, nine "compu-grading" programs assessed student writing equally as well. How can this be?

The programs in the study used previously graded student assessments to "learn" how human graders grade certain passages. However, points out Bob Broad, professor of English at Illinois State University, "the human graders they are comparing the program to are human readers trained to read like machines."

The rubric for the material being graded was kept very short and focused, says Broad; human graders were trained to skim the writing as quickly as they could for a few key items.

"I don't think you can take pride in matching human readers who read like machines already," says Broad.

Perelman sees other flaws in the study, which was funded by the Hewlett Foundation. Four of the eight

data sets analyzed were essentially one-paragraph answers to reading questions. "They weren't testing argumentation; they were testing whether certain information was reported back."

The only lengthy answer set (over 500 words, on average) was the one that the computer "did very poorly on."

Shermis himself acknowledged in a recent National Public Radio interview that he ran the Gettysburg Address through compu-grading programs—and it failed, earning only 2s and 3s on a 6-point scale.

Indeed, innovative writing is a challenge for machine scoring programs. Chris Anson, English professor at North Carolina State University, says that the only computer grading systems that work "are ones that severely constrain the genre. Only narrowly defined kinds of texts are used; you can't have a ton of variety."

"Computers can't understand what students have said; they don't have any capacity to interpret meaning. You can present a computer scoring system text that has certain features of organization, but all the information is completely wrong—and the computer will give a high score."

Most educational testing companies do not allow researchers like Perelman to test their software; he praises ETS for making its e-rating program available to him. Other companies are less transparent.

"Their attitude is, 'trust us—but we won't let you examine it,'" says Perelman. "These companies promote educational testing and claim they share the public's concern that schools should be transparent and accountable, and yet they aren't themselves. Toasters are more scrutinized than high-stakes educational tests."

## Why Computer Grading of Writing Is a Bad Thing

Whether at K-12 or in higher education, shifting to automated assessments of student writing holds many potential perils.

**• English language learners and others with only tentative language mastery may be further confused and set back when e-graders tell them what they are doing is wrong—when it isn't.**

In one essay, Perelman quoted an Oscar Wilde sentence: "I can resist everything except temptation." The e-rating program informed Perelman that "except" was an incorrect use of a preposition, which is not correct.

"What it often identified as errors were preposition errors or article errors that weren't errors," says Perelman. "What's scary is that for the bilingual or second language learners, this false feedback not only destroys their self-confidence but subverts emerging notions of how English works. It does harm."

Grading writing has benefits for faculty, according to Les Perelman and Bob Broad. They argue that when teachers get together and discuss the attributes of good writing, they are actually experiencing a form of professional development.

**• Computer grading may be inversely linked to privilege, resulting in a two-tiered educational system.**

Herrington and Moran note in their essay that customers of e-grading programs tend to be community colleges, not top Ivy League or elite schools.

"This is bifurcating education," says Perelman. "The student who needs the most help with writing will get negative help. Computer grading will be a justification for allowing them to be in writing classes with 40 to 60 students and teachers teaching four to five classes a semester, as opposed to an elite school where the average number of students in first-year writing classes ranges from 10 to 18."

**• Computer assessments deprive teachers of the opportunity to discuss students' writing—the type of professional development experienced in the National Writing Project's summer assessment conferences, says Perelman.**

"If you have real teachers grading the essays you are getting a 'two-fer': not only are the essays getting graded, but having teachers get

*Continued on page* 8

 *Continued from page 7*

together and discuss the attributes of good writing is a valid form of professional development."

Broad agrees: "Grading writing is good for the faculty. It gives them an opportunity to collaborate and read student writing."

**• Reading and writing are diminished and devalued.**

The concept of writing as communication is upended when a student is writing only for a robot. Students may be less motivated to write creatively or with passion when they know their audience is a machine, not another person; the communication cycle is broken.

"Automated assessments distort the nature of writing," says Broad. "A computer can't successfully reconstruct what happens when one person reads another person's writing."

Students writing for computers will be more likely to simply add in what the computer can count: grandiose words, keywords from the question, extra words to pad the word count.

**Automated assessments distort the nature of writing. A computer can't successfully reconstruct what happens when one person reads another person's writing.**

—*Bob Broad, professor of English at Illinois State University*

For example, says Broad, his daughter wrote an essay for high school and ran it through Microsoft Word's readability analyzer. She discovered her essay had been written at a 9th-grade level; she reworked the text to get a 12th-grade ranking.

"I pointed out that the new version was not as good," says Broad. "She had done things to her sentence construction and style that were more pleasing to the computer program that were not as pleasing to the human reader."

While Broad's daughter was changing her text to please Word and not an e-rating program, it's easy to see the same scenario playing out if she had had access to a robo-grader instead.

**• Machine graders are easily "gamed."**

Perelman says he could train any student with basic reading knowledge to do well on these exams. "It's easier to 'game' a robo-grader than it is a multiple-choice test."

Broad notes that this leads to student cynicism: "'How can I beat the computer? How can I game the system?'"

**• Curriculum will be driven by computer-gradable content.**

Instead of "teaching to the test," teachers may find themselves teaching to the test-grading robot. As they see what students need to do to succeed with online, automated assessments, they may constrain their assignments accordingly.

"When people figure out what the machine is looking for, that is going to drive curriculum," says Anson. "That's a serious problem."

## Are There Any Good Uses for Computer Assessment in Writing?

Some educational software companies offer digital content using built-in formative assessments aimed at helping students write better by providing instantaneous feedback. Perelman suspects this is something of a "bait and switch," and that testing companies' real agenda is to provide placement and other high-stakes, mass-market (and therefore highly profitable) testing tools.

In any event, he says, feedback can be handled more accurately and cheaply by Microsoft Word. Perelman tested sample essays against both an e-rater and Word, and found that Word's grammar checker—flawed as it sometimes seems to be—far out-performed the computer grading program.

Using error-plagued robo-graders for feedback, he says, "does more harm than good."

Anson says software that can mine text has value to research—blazing through thousands of pages in a second to count, say, gender pronouns or analyze political statements.

However, he says, artificial intelligence isn't yet capable of truly assessing human writing—and he thinks it will take many years for this type of advance to occur.

"I don't think in our lifetimes we're going to see improvements in artificial intelligence great enough where computers can replace humans."

Some English instructors may see automated grading as a way out from under piles of papers to grade; they say that having such a tool makes them more likely to assign more writing to students.

However, the inaccuracy of such tools makes this use, tempting as it might be, counter-productive, says Broad. Instead, to cope with a heavy writing load, he suggests that teachers use other techniques.

"Peer response among students is a helpful thing. Teachers can be providing their responses while students are still working on texts, not just at the end of a project. If a teacher gives a response first and then makes evaluations at the end, the evaluation may be quicker.

"There are ways that teachers can make their evaluations more efficient without resorting to giving that job to computers."

## What To Do If Your District Is Considering Robo-Grading

Perelman suggests showing administrators, perhaps through trial use, how such programs are ineffective, potentially harmful, and easily gamed.

"I'd tell the administrator that our students who are ELLs will get all this false information. I'd get examples and show how this is going to do damage.

"I'd say, 'Look at how I can tell my students to get high scores on my essays: just memorize lists of big words and string them together and take sentences from the newspaper and paste them into paragraphs in the middle and they'll get a high score.'"

Perelman says a "bottom line" argument is the best way to convince administrators, rather than a focus on the way computer grading subverts the notion of writing-as-communication.

"In a time when people are worried about school budgets being crunched, being philosophically wrong is not going to get you much; but if it doesn't work, that could change minds."

Besides talking to administrators, teachers should go to meetings, write op-eds for local newspapers (focusing on facts and data), and experiment with compu-grading software if they can gain access to it, suggests Perelman.

Nonetheless, warns Broad, it's not going to be an easy battle to win.

"The combination of technology and apparent—though maybe illusory—cost savings is going to be irresistible for many educational administrators," he says.

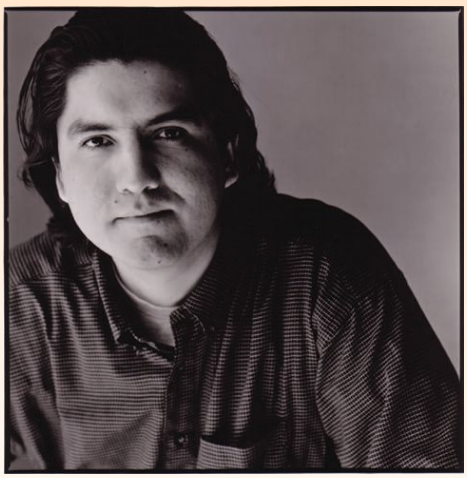
"They love technology and they love saving money. It's going to be a struggle to keep humans in writing assessment. We want educators to keep their hope and to not give up—to keep up the struggle. It's a worthy one."

---

***Lorna Collier*** *is a freelance writer and author based in northern Illinois.*

By *Trisha Collopy*

# Sherman Alexie's Many Tribes



Sherman Alexie hoped that his first young-adult novel, *The Absolutely True Diary of a Part-Time Indian*, would be a literary "gateway drug" to students of color, who don't usually see themselves in YA fiction.

He didn't expect it to hit a nerve with readers across the country, and from a wide range of backgrounds and ethnicities.

"It's been astonishing," he said of the reaction to the 2007 book. "I get letters from students in classrooms almost every day. And wherever they're from, they so closely identify with Junior. It's weird for a rez (reservation) Indian boy to become an archetype."

The 2007 book turned up on bestseller lists and won a National Book Award for Young People's Literature. It has become a widely taught book in high school classrooms, drawing censorship challenges and a flood of fan letters.

"The ones that really get me are the letters that come from prep schools, privileged kids," Alexie said. One student identified with Junior, the novel's main character, because he wanted to be a journalist, but his dad was sending him to a military academy.

"So this rich kid was feeling trapped by his tribe. And it hadn't occurred to me to think a rich kid could be trapped by his tribe's expectations," Alexie said.

Alexie burst onto the literary scene in the early 1990s with his first collection of stories, *The Lone Ranger and Tonto Fistfight in Heaven*, an irreverent and bleak look at life on the

*Continued on page* 10