The New York Times

EDUCATION | ON EDUCATION

Facing a Robo-Grader? Just Keep Obfuscating Mellifluously

By MICHAEL WINERIP APRIL 22, 2012

A recently released study has concluded that computers are capable of scoring essays on standardized tests as well as human beings do.

Mark Shermis, dean of the College of Education at the University of Akron, collected more than 16,000 middle school and high school test essays from six states that had been graded by humans. He then used automated systems developed by nine companies to score those essays.

Computer scoring produced "virtually identical levels of accuracy, with the software in some cases proving to be more reliable," according to a University of Akron news release.

"A Win for the Robo-Readers" is how an Inside Higher Ed blog post summed things up.

For people with a weakness for humans, there is more bad news. Graders working as quickly as they can — the Pearson education company expects readers to spend no more than two to three minutes per essay— might be capable of scoring 30 writing samples in an hour.

The automated reader developed by the Educational Testing Service, e-Rater, can grade 16,000 essays in 20 seconds, according to David Williamson, a research director for E.T.S., which develops and administers 50 million tests a year, including

the SAT.

Is this the end? Are Robo-Readers destined to inherit the earth?

Les Perelman, a director of writing at the Massachusetts Institute of Technology, says no.

Mr. Perelman enjoys studying algorithms from E.T.S. research papers when he is not teaching undergraduates. This has taught him to think like e-Rater.

While his research is limited, because E.T.S. is the only organization that has permitted him to test its product, he says the automated reader can be easily gamed, is vulnerable to test prep, sets a very limited and rigid standard for what good writing is, and will pressure teachers to dumb down writing instruction.

The e-Rater's biggest problem, he says, is that it can't identify truth. He tells students not to waste time worrying about whether their facts are accurate, since pretty much any fact will do as long as it is incorporated into a well-structured sentence. "E-Rater doesn't care if you say the War of 1812 started in 1945," he said.

Mr. Perelman found that e-Rater prefers long essays. A 716-word essay he wrote that was padded with more than a dozen nonsensical sentences received a top score of 6; a well-argued, well-written essay of 567 words was scored a 5.

An automated reader can count, he said, so it can set parameters for the number of words in a good sentence and the number of sentences in a good paragraph. "Once you understand e-Rater's biases," he said, "it's not hard to raise your test score."

E-Rater, he said, does not like short sentences.

Or short paragraphs.

Or sentences that begin with "or." And sentences that start with "and." Nor sentence fragments.

However, he said, e-Rater likes connectors, like "however," which serve as programming proxies for complex thinking. Moreover, "moreover" is good, too.

Gargantuan words are indemnified because e-Rater interprets them as a sign of lexical complexity. "Whenever possible," Mr. Perelman advises, "use a big word. 'Egregious' is better than 'bad.' "

The substance of an argument doesn't matter, he said, as long as it looks to the computer as if it's nicely argued.

For a question asking students to discuss why college costs are so high, Mr. Perelman wrote that the No. 1 reason is excessive pay for greedy teaching assistants.

"The average teaching assistant makes six times as much money as college presidents," he wrote. "In addition, they often receive a plethora of extra benefits such as private jets, vacations in the south seas, starring roles in motion pictures."

E-Rater gave him a 6. He tossed in a line from Allen Ginsberg's "Howl," just to see if he could get away with it.

He could.

The possibilities are limitless. If E-Rater edited newspapers, Roger Clemens could say, "Remember the Maine," Adele could say, "Give me liberty or give me death," Patrick Henry could sing "Someone Like You."

To their credit, researchers at E.T.S. provided Mr. Perelman access to e-Rater for a month. "At E.T.S., we pride ourselves in being transparent about our research," Mr. Williamson said.

Two of the biggest for-profit education companies, Vantage Learning and Pearson, turned down my request to let Mr. Perelman test their products.

"He wants to show why it doesn't work," said Peter Foltz, a Pearson vice president.

"Yes, I'm a skeptic," Mr. Perelman said. "That's exactly why I should be given access."

E.T.S. officials say that Mr. Perelman's test prep advice is too complex for most students to absorb; if they can, they're using the higher level of thinking the test seeks to reward anyway. In other words, if they're smart enough to master such sophisticated test prep, they deserve a 6.

E.T.S. also acknowledges that truth is not e-Rater's strong point. "E-Rater is not designed to be a fact checker," said Paul Deane, a principal research scientist.

"E-Rater doesn't appreciate poetry," Mr. Williamson added.

They say Mr. Perelman is setting a false premise when he treats e-Rater as if it is supposed to substitute for human scorers. In high stakes testing where e-Rater has been used, like grading the Graduate Record Exam, the writing samples are also scored by a human, they point out. And if there is a discrepancy between man and machine, a second human is summoned.

Mr. Foltz said that 90 percent of the time, Pearson's Intelligent Essay Assessor is used by classroom teachers as a learning aid. The software gives students immediate feedback to improve their writing, which they can revise and resubmit, Mr. Foltz said. "They may do five drafts," he said, "and then give it to the teacher to read."

As for good writing being long writing, Mr. Deane said there was a correlation. Good writers have internalized the skills that give them better fluency, he said, enabling them to write more in a limited time.

Mr. Perelman takes great pleasure in fooling e-Rater. He has written an essay, then randomly cut a sentence from the middle of each paragraph and has still gotten a 6.

Two former students who are computer science majors told him that they could design an Android app to generate essays that would receive 6's from e-Rater. He says the nice thing about that is that smartphones would be able to submit essays directly to computer graders, and humans wouldn't have to get involved.

In conclusion, to paraphrase the late, great Abraham Lincoln: Mares eat oats and does eat oats, but little lambs eat ivy.

A kiddley divey too, he added, wouldn't you?

E-mail: oneducation@nytimes.com

A version of this article appears in print on April 23, 2012, on Page A11 of the New York edition with the headline: Facing a Robo-Grader? No Worries. Just Keep Obfuscating Mellifluously.