Can Computers Grade Writing? Should They?

Douglas D. Hesse
Executive Director of Writing
The University of Denver
dhesse@du.edu
303-871-7447
February 2012

A few weeks ago a University of Denver colleague sent me an article about a program that would grade student writing. She asked my opinion, and I'm afraid that, in the rush of time, I dashed off a not-very-useful, "It's an abomination." Well, the question is more interesting and complex than my response, so I wanted to sketch the background of the issue and the reasons why I'm wary. Computer graded writing is especially important as the assessment of the Common Core Standards for writing (the fairly recent standards to which American students will be subjected) will be handled by programs similar to the Educational Testing Service's E-rater or Pearson Knowledge Technology's Intelligent Essay Assessor. A recent article in the popular press imagines "robo-readers" as "the new teachers' helper" (Simon).

The dream that teachers might be "spared" grading student writing (and I'll explain the very intentional scare quotes) has been around for decades. The title of William Wresch's 1993 article "The Imminence of Grading Essays by Computer—25 Years Later" gives some sense of the enterprise's venerability. A landmark essay by Anne Herrington and Charles Moran (the latter one of the founding figures in computers and writing) analyzed with caution "What Happens When Machines Read our Students' Writing." A 2006 book edited by Patricia Ericsson and Richard Haswell encapsulated the state of the art and thought and includes a comprehensive bibliography to that point. Since then, the issue percolates occasionally in the popular and educational press. A story in *Inside Higher Education* a year ago asked, "If a computer can win at "Jeopardy," can one grade the essays of freshmen" (Jaschik)? The answer from scholars including folks at ETS was "yes;" the spirited reply by longtime critic Les Perlman at MIT was "no." A featured session at this year's Conference on College Composition and Communication, the nation's largest and most important gathering of college writing professors, had the provocative title "Automated Essay Scoring: Gateway to Valid Assessment, Effective Learning, or the Twilight Zone?" It featured two senior researchers from ETS and two senior researchers from the field of composition studies, including Professor Perlman.

Sorting all of this out raises two questions: 1) Are computer-generated assessments of writing valid and reliable? 2) Even if so, should they be used?

How Good Are Machine Scores?

The answer to this question makes more sense if you understand how machine scoring works. In a nutshell, these programs use a set of algorithms that look for certain linguistic features in a text, then assign weights to each to produce a composite score. One version of the ETS E-rater scoring engine looks at the following:

- "content analysis based on vocabulary measures
- lexical complexity/diction
- proportion of grammar errors
- proportion of usage errors
- proportion of mechanics errors
- proportion of style comments
- organization and development scores
- features rewarding idiomatic phraseology" ("How")

Most of these are fairly straightforward, but a few merit some explanation. "Content analysis based on vocabulary measures" and "organization and development scores" are fairly related. They depend on chains of words, synonyms, and collocated terms. A couple of examples will illustrate.

1 (High). Dogs are interesting animals. Among pets, they're very supportive. They wag tails and bark.
2 (Low). Dogs are interesting animals. Drill bits are better when they're sharp. Iceberg lettuce isn't very nutritional.
3 (?). Dogs are interesting animals. As pets, drill bits and other canines are generally supportive. They are sharp, wag tails, and bark.

Example 1 gets rewarded because of vocabulary chaining (dogs, pets, wagging, barking). Example 2 gets penalized for the lack thereof. It's important to recognize, then, that "analyzing content" and "development" are measured through word chains and collocations and related syntactic features. The more sophisticated the algorithms are (for example, associating "wag" and "fetch" with "dog") the more reliable these measures can be. I threw in example 3 as the kind of thing that might score relatively highly, even though human readers discern the non sequiter. "Development" is also a function of length and the presence of words and phrases deemed specific or associated with specificity. (Several "for examples" or "for instances" will raise the score.) One problem, as Les Perlman famously and dramatically illustrated, is that the "truth" or logic of essays is undervalued or overlooked altogether.

Lexical complexity/diction simply rewards big vocabularies. Thesaurus-loving students can rejoice. Consider the following two sentences.

4 (Lower). Dogs are interesting animals.
5 (Higher). Canines exemplify engrossing vertebrates.

#5 scores higher because the diction is less common, and writers are rewarded for that. Of course, most of us would prefer sentence #4. What's missing is a sense of rhetorical situation and audience. A sentence like #5 would be wholly inappropriate for a target audience of 8 year olds. Now, one can tweak the algorithms to account for this, making the program penalize a too-high percentage of words outside an 8-year-old's vocabulary, but that brings up another issue: the difference between "generic" and "specific" writing tasks and qualities. Computer scores tend to be more valid and reliable—in relation to scores from expert human readers—when the tasks are very carefully defined and limited in length. The SAT writing sample, for example, gives students 25 minutes to write on a very limited task, and lots of corpus analysis and programming energy, can generate a reliable score.

Finally, the style scores tend to reward things like sentence length and type, as in the examples below. #7 rates higher because it's a longer sentence and because frontloading the adjectival and verbal phrases in sentence that suspends its subject and verb is associated with more sophisticated styles.

6 (Lower). Dogs are interesting animals. Dogs are friendly to their owners. Dogs show affection by wagging their tails.

7 (Higher). Friendly to their owners, wagging tails to show affection, dogs are interesting animals.

Now, I'm woefully oversimplifying how these programs work, and I've no doubt that analytic elements beyond the ones I quoted above are built into many algorithms. I realize that I risk caricaturing them in doing so, which would be both unfortunate and unfair. I'm simply trying to suggest the complexity of this undertaking, which ultimate participates in the same challenges that work on artificial intelligence more broadly encounters.

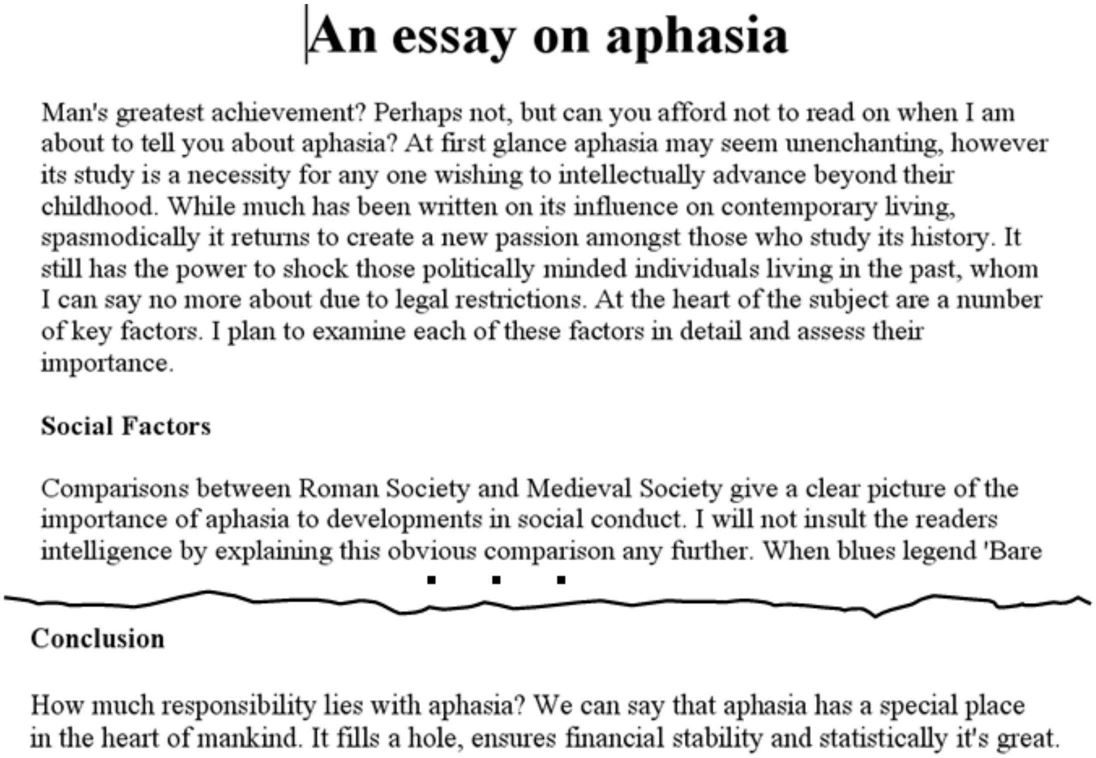
"Beating the System"

There's no doubt that computer scoring systems are getting more sophisticated. The same technology that allows people to have "conversations" with the iPhone's Siri, is improving the analysis of writing. Still, just as Siri is not well-equipped to discuss with you whether Nietzsche or Wittgenstein is the better philosopher, so too do computer scoring systems run into difficulty with complex tasks; they can surely score the elements pretty well, but whether these elements add up to a strong or weak piece of writing is another matter.

In 2005, I published an article that, in part, showed how relatively easy it was to beat the Intelligent Essay Assessor, Pearson Knowledge Technology's s computer scoring engine (Hesse).

That article explained how I looked at an online site, "The Essay Generator" which invited you to enter any topic and receive an essay in return. The site wasn't very sophisticated, which is part of the fun. Its database stockpiled sentences: several possible first sentences, several second, and so on. Each essay had three headings: Social Factors, Economic Factors, and Political Factors. Each essay had a graph. Each essay ended with a fictional quotation. Each essay had three references.

At that time, Pearson was allowing people to try out a version of its scoring engine, the Intelligent Essay Assessor (IEA). Online you could enter an essay written on one of three topics, hit send, and receiving some immediate basic scoring. We spent one graduate course I was teaching trying to figure out the algorithms involved by submitting variant essays and seeing what our revisions did to the scores. Playing with the Essay Generator got me thinking. What if you had a computer generate an essay that was then scored by another computer? One of the IEA sample topics was "aphasia," so I plugged this topic into the essay generator, which generated the following essay.



An essay on aphasia

Man's greatest achievement? Perhaps not, but can you afford not to read on when I am about to tell you about aphasia? At first glance aphasia may seem unenchanting, however its study is a necessity for any one wishing to intellectually advance beyond their childhood. While much has been written on its influence on contemporary living, spasmodically it returns to create a new passion amongst those who study its history. It still has the power to shock those politically minded individuals living in the past, whom I can say no more about due to legal restrictions. At the heart of the subject are a number of key factors. I plan to examine each of these factors in detail and assess their importance.

Social Factors

Comparisons between Roman Society and Medieval Society give a clear picture of the importance of aphasia to developments in social conduct. I will not insult the readers intelligence by explaining this obvious comparison any further. When blues legend 'Bare

. . .

Conclusion

How much responsibility lies with aphasia? We can say that aphasia has a special place in the heart of mankind. It fills a hole, ensures financial stability and statistically it's great.

I then cut and pasted the aphasia essay into the IEA and received the following feedback.

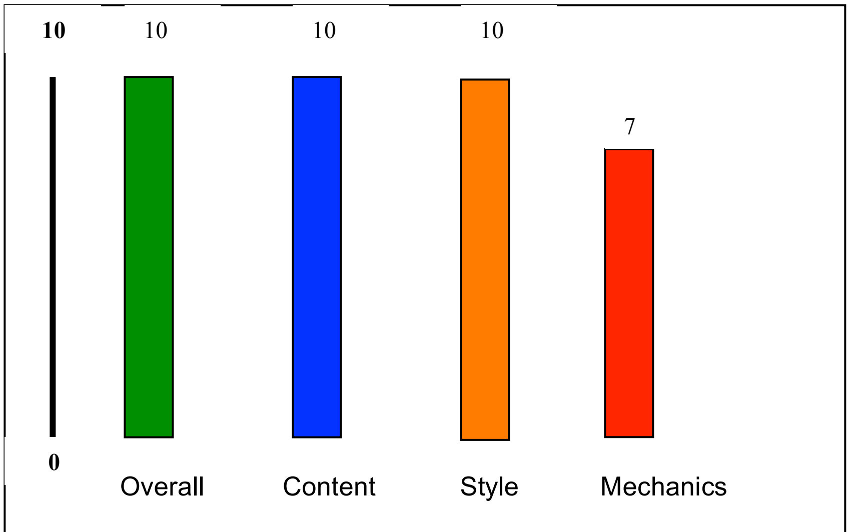


Figure 10: Intelligent Essay Assessor Score of "An Essay on Aphasia"

As you can see, the Intelligent Essay Assessor thought the Essay Generator wrote pretty well, though it could use a little help with grammar and mechanics. Still, it sure knew its aphasia. The point is that the rules the IEA was following for scoring couldn't discern that the content of the essay was nonsense; what they could discern seemed strong.

Now, although there are many examples of this kind gaming the system (so many that the sites took down their test engines), it would be unfair categorically to dismiss all machine scoring. No doubt, further, the programs have gotten more sophisticated over the years; still, the principle holds that they can't read anything like the full range of textual aspects that a person can. Their success depends on the tightness of the correlation between those aspects the programs can "read" and the full complement of features that human readers can.

So, Can Computers Score Writing?

The answer is "yes, but." On the one hand, computers can fairly reliably identify certain features of texts, assign a score to those features, and use a formula to generate an overall rating of the essay. For well-identified and circumscribed tasks, computer scores compare favorably to those generated by people, though there remain problems. In a very large study on specific questions in the GRE exam, only 2% of the essays scored by two human raters needed to go to a third reader for arbitration; however, 41% of the essays scored by a computer and a human needed to go to a third reader (Ramineni 26). The further "but" is that success is more elusive for tasks that are complex, writings whose solutions are not well-circumscribed. Furthermore, the aspects of the text most important to readers, including insight and success for a given rhetorical situation, are beyond the ken of the software. Now, if it's the case that certain textual features are always associated with "insight," for example (and vice versa), then the machine's shortcoming is of no matter. That can't be granted right now, but let's imagine that linguists, psychometricians, rhetoricians, and programmers will keep getting better at it.

Let's suppose, in short, that we computers pass a sort of Turing test , producing evaluations indistinguishable from a trained human rater. What's not to like?

Should Computers Score Writing?

There are two primary reasons given for having computers score writing. One is economic, a savings that accrues in mass testing situations. If you have hundreds of thousands of essays to grade for a national testing situation (as for example with the Common Core standards), hiring and training enough human raters to complete the task in a reasonable time is expensive. So, even massive up-front investments in writing and testing software are recouped in fairly short order, especially if the testing is ongoing and not one-time. This particular advantage disappears, of course, if you deem such mass testing unnecessary or if you look at the "cost" of hiring and training human raters actually as an investment in professional development, with teachers learning from systematically looking at lots of student writing in concert with others. But there it is.

Another justification is that responding to and evaluating writing is “drudgery” from which teachers should be spared. There’s no doubt that reading and thoughtfully responding to papers takes time, and for each of us who find this meaningful work, there are no doubt several who would do this. The dream of machine scoring, then would “free” teachers of responding to writing, allowing them to “teach.”

Now, here’s an interesting notion. I see ***responding*** to student writing ***as*** teaching, in fact, as the most important element of teaching writing. I’ve co-authored a textbook going into its tenth edition that distills my best advice about writing into 900 pages. Standing before a class and declaiming its wisdom (or putting taping lectures to put them online) are the least vital part of my teaching someone how to write. People learn to write by writing and getting feedback. Writing, like playing the piano or playing tennis or painting watercolors, is a skill learned by doing, with feedback and coaching. I can show you how to grip a racquet, but until you step on a court and hit ten thousand balls, my “teaching” (which would take about five minutes) hasn’t taught you how to serve.

OK, but can’t computers serve as coach, providing feedback? That’s an interesting question. (To some extent they already do when Microsoft Word squiggles a misspelled word or what it takes to be a grammar mistake.) At some level they can advise “ramp up your vocabulary” or “vary your sentence patterns” or “write more” or “provide more examples.” At this level, improving a text can be like playing a complexly narrative computer game that one masters by repeatedly dying, acquiring some bits of wisdom and some motor skills that eventually allows you to beat the troll, enter the treasure chamber, rescue the prince, save civilization. There’s some research that some student writers revise more with computer feedback on their writing because the feedback cycle can be so quick and, as with the video game, there’s no judgment.

However, and this is an important distinction, the motivation for revision in this circumstance differs significantly from the motivation for “authentic” writing. The million (or $100 million) dollar question is whether skills derived through this process will transfer to authentic writing for authentic readers. The corollary question, of inestimable value in my mind, is how students through a computer-coached process come to understand writing. Do they perceive it, to state the poles with intentional grandeur, as a fundamental act of human activity, making and exchanging ideas between people, or do they understand it as a kind of textual game where they’re creating texts that satisfy a rule-governed system whose satisfaction is the ultimate end and reward?

Because I worry on both counts—about the transferability of skills and about the cost to how students perceive writing (and the resultant cost to human culture)—I’m reluctant to be enthusiastic about computer scoring replacing the “drudgery” of responding to writing. I realize that I could be dismissed as a romantic humanist or misguided luddite on this point, and in a longer piece than I have time for here, I’d address these points, including (Why am I not as concerned with online problem sets in calculus or statistics?) I just want to share caution about what, in the name of efficiency and productivity, we might be doing to the ecology of writing, especially the areas of rhetorical complexity, the role of writing for establishing and shaping relationships, and the diversity of readings and interpretation. (William Condon and Bob Broad have explored these issues further.) What we give up in order to achieve reliability is some nontrivial amount of validity, not for specific well-defined tasks but for the fuller range of writing.

I've simply tried to illustrate why the question of computer response to writing is so complicated, in dimensions technical, philosophical, and social. Perhaps this will start a longer conversation on campuses, one that takes my colleague's initial question as seriously as it merits.

Works Cited

Broad, Bob. "More Work for Teacher? Possible Futures of Teaching Writing in the Age of Computerized Writing Assessment." In Ericsson and Haswell, 221-33.

Condon, William. "Why Less Is Not More: What We Lose by Letting a Computer Score Writing Samples." In Ericsson and Haswell, 211-20.

Ericsson, Patricia Freitag and Richard Haswell. *Machine Scoring of Student Essays: Truth and Consequences*. Logan: Utah State UP, 2006.

Herrington, Anne and Charles Moran. "What Happens when Machines Read Our Students' Writing?" *College English* 63.4 (2001): 480-99. Print.

Hesse, Douglas. "Who Owns Writing?" *College Composition and Communication* 57.2 (Dec. 2005): 335-57. Print.

"How the e-rater Engine Works." Educational Testing Service, 2012. http://www.ets.org/erater/how/. Web. 22 Feb. 2012.

Jaschick, Scott. "Can You Trust Automated Grading?" *Inside Higher Ed*. 21 Feb. 2011. http://www.insidehighered.com/news/2011/02/21/debate_over_reliability_of_automated_essay_grading#ixzz1n8o7EhC4

Ramineni, Chaitanya, et al. *Evaluation of e-rater for the GRE Issue and Argument Prompts*. ETS Research Report RR-12-02. ETS: Princeton, 2012. Web.

Simon, Stephanie. "Robo-readers: The New Teacher's Helper in the U.S." Reuters. 29 March 2012. http://www.reuters.com/article/2012/03/29/us-usa-schools-grading-idUSBRE82S0ZN20120329 Web.

Wresch, William. "The Imminence of Grading Essays by Computer—25 Years Later." *Computers and Composition* 10.2 (1993): 45-58.