Can computers make the grade in writing exams?

The new Common Core writing exams may use computer algorithms to assess student work.

By Samina Hadi-Tabassum

Last year, I presented several professional development workshops on the Common Core State Standards to school districts across Illinois. Inevitably, we arrive at the Core’s writing standards, and we discuss the new writing requirements for the PARCC Assessment (Partnership for Assessment and Readiness for College and Careers) that is being implemented this year throughout Illinois and relies heavily on a student’s ability to argue effectively in writing.

SAMINA HADI-TABASSUM (shadi@dom.edu) is an associate professor of education and director of the bilingual/ESL program at Dominican University, River Forest, Ill.
To put it bluntly, writing assessment has a checkered past in Illinois. In 2005, state officials eliminated the writing test to save $6 million (Rado, 2005), becoming the only state to eliminate standardized assessment of writing for students along with the social studies test. The state was also under pressure to meet the No Child Left Behind federal mandates for reading and mathematics but not necessarily for writing. Lawmakers like State Sen. Miguel del Valle (D-Chicago) led efforts to reinstate the writing test in 2006 along with a $2 million appropriation for designing a new writing assessment. However, in 2010, Illinois eventually eliminated the writing tests for 3rd, 5th, 6th, and 8th grades in order to save just $3.5 million dollars, even though only 54% of 5th graders had mastered the writing standards in 2009, and students performed worse on the standardized writing tests than the reading and math tests (Rado, 2010). The 11th-grade writing test was kept intact for the sake of college entrance requirements.

As colleges were demanding better writing skills from incoming students, Illinois was eliminating writing assessments across the state, even though college-bound students still had to be assessed on writing when they took the ACT and SAT. Multiple choice test questions, such as those on the ACT Writing Test, can measure students’ understanding of writing conventions. But these tests do not align with what will be expected of students in college where they will need to write well-crafted academic essays.

Yet the reliability of assessing student essays by humans also can be questioned, which is why states use multiple human scorers to ensure reliability — also called inter-rater reliability — on standardized writing tests. This adds significant costs. In the 1990s, each writing sample from just one student was assessed by at least three professionals in order to ensure inter-rater reliability. These professionals needed to be housed in hotels for several days as they evaluated thousands of writing samples over hundreds of hours and across several grades in order to ensure fairness, equity, and reliability — all on an hourly basis. As they combed through writing samples, they measured a student’s authorial voice, sense of organization, ideas, creativity, and use of conventions in order to provide a final quantitative score.

Enter the algorithm

Today states rely on computer technology to assess student writing. Instead of writing, erasing, and rewriting on a piece of paper with a pencil, students are required to write into a text box that lets them check for spelling, grammar, and word choice while they are writing responses. Word-processing skills are required.

The National Assessment of Educational Progress
Algorithms vs. humans

At the same time, newer algorithms continue to replace traditional algorithms like LSA, and the type of discourse process being asked of the student in the writing assignment matters immensely. For example, if the student is asked to summarize the reading passage, then the algorithm is measuring the synthesis and coherence of ideas. On the other hand, writing short-answer responses to isolated questions is a much easier discourse to assess since the algorithm is measuring the use of targeted words. However, if students were asked to write open-ended essays instead of summaries, applying what they had read to a new situation, the resulting essays might be expected to be less similar than the summaries to the original text. In addition, we would expect less similarity across students, and we might not expect the level of similarity to predict essay quality. In other words, LSA may not be appropriate in open-ended essay questions. The best bet is to match the individual student’s essay to a pool of essays that human scorers have already judged to be excellent.

Research studies have been conducted on the reliability of scores from an LSA algorithm versus human scorers and the reliability between human scorers. Often, there is higher correlation between human scorers. But research also finds the reliability scores between LSA algorithms and human scorers adequate enough to move forward with computerized testing though it definitely is not the ideal method. In addition, some algorithms are more reliable than others. Some algorithms may be better suited for assessing narrative writing than expository writing due to the differences in cognitive demand between these discourse types. Synthesizing a summary from an expository text is inherently different...
from synthesizing the characters and plot for a narrative writing task. According to research, the LSA algorithm is more reliable on essays of more than 60 words; it encounters particular difficulties in the two- to 60-word range (Wiemer-Hastings, Wiemer-Hastings, & Graesser, 1999).

**Polysemy and other challenges**

Algorithms also vary in efficiency depending on whether they are analyzing a student's writing holistically or whether the algorithm is analyzing various components within a text, which is referred to as the componential method that examines multiple features in a student's writing. A holistic method of assessment, on the other hand, uses the cosine measure to determine the semantic relationship between the student's writing against the text holistically as well as perhaps measuring the student's writing against an expert essay or "golden" essay, which is created from a pool of 100 high-scoring essays previously graded by human scorers. The higher the cosine score, the higher the semantic relationship between the student's writing and the original text itself.

We can all agree that algorithms have gotten better over the decades, which is why high-stakes tests such as the ACT, SAT, and now the PARCC are using the computerized testing of writing. W. Kintsch's (2001) study describing the prediction algorithm, which makes the language used in a student essay more context dependent, shows that the prediction algorithm is improving the polysemy problem inherent in LSA. (Polysemy refers to the idea that a word like "bank" has multiple meanings—a river bank as opposed to a money bank—and the prediction algorithm is trying to use statistical analysis to determine which "bank" is being referred to in the student's writing.) Many English words have multiple meanings, and the algorithm must predict the correct meaning based on context. For example, when the student writes "the store owner is a shark," the algorithm has to make predictions with words like aggressive, predatory, and tenacious, as opposed to words like fish, ocean, and fins, which are literal referents. Sharks and store owners also do not have an established semantic network in the English lexicon, and therefore the semantic link is unique to the individual student. Metaphorical extensions and colloquial phrases such as "nerves of steel" therefore are more challenging for an algorithm to make correct predictions since they might not be context-dependent.

One can conclude that anomalous writing in general would not benefit from being assessed by algorithms. Educators also doubt whether an algorithm can read paragraphs, process sentences in their entirety, reason with flexibility, and understand tricky language the way humans can. Therefore, an algorithm must be expansive enough to capture a wide range of semantic relationships between words and their meanings and in unique contexts—otherwise known as increasing its semantic memory. Mathematically, more terms must be added to the cosine vector so that instead of 50 target "lemmatized" words the algorithm now looks for 75 lemmatized words with strong links in meaning to the original text. (Lemmatization is the algorithmic process of determining the lemma for a given word. For example, in English, the noun "night" may appear as "night," "nightly," "nightingale," etc. The base form, "night," is called the lemma.)

**Effect on scores**

School administrators are concerned with how the computerized writing testing will affect overall school scores, whether they have the equipment and space to even assess all students using computerized testing and whether accommodations such as extended time can benefit student writing scores. The ratio of students to computers is closer to 3-to-1 at the national level, and access to fast wireless services may not be possible for all. At the same time, school administrators are happy that student writing can be assessed instantly when it is computerized. What's more, there are inherent gaps between students who have access to technology at home versus those who still do not, possibly widening the academic gap even further between student populations.

Colleges are already using computerized software to let students write their essays, submit their essays, receive their grade immediately, and then choose to rewrite the essay based on the grade. EdX, a writing assessment software company developed by Harvard and MIT engineers, uses artificial intelligence to evaluate essays and provide immediate feedback (Markoff, 2013). EdX has human scorers first evaluate 100 essays on a topic, then creates a pool of high-scoring essays, and finally evaluates incoming essays against this pool using algorithms. Students
can choose to continually rewrite their essay and not wait weeks for feedback from their professors — almost making it into a game of perfection. Students often want and need instant feedback on their writing, but EdX offers mostly general feedback, such as whether the student was on target and on topic.

Many states, such as West Virginia and Indiana, are starting to incorporate the automated assessment of writing in their secondary schools. Yet how reliable and specific is the machine feedback? Les Perlman, a former writing director and current researcher at MIT, claims the software is not reliable, and he has started a group of MIT educators — including Noam Chomsky — who oppose machine scoring of student writing because, they maintain, computers cannot read and reason like human beings.

How do you hold an algorithm responsible?

Computers cannot look at the evidence in the writing and make good sense of it. Computers cannot determine if the arguments the student poses are convincing. Computers cannot determine whether the student essay is truthful and clear in its language. When it comes to high-stakes testing and writing assessment, PARCC may need to brace itself for opposition and maybe even lawsuits. There will be social and political consequences for allowing computers to automatically grade student essays. Yet how do you hold an algorithm responsible? If we erect algorithms as our ultimate judges and arbiters, we face the threat of difficulties not only in enforcement but also in culture. At the same time, human scorers take too much time to assess writing and cost much more.

Outsourcing humans

Deep in our hearts, there is also the growing hysteria that educators will be outsourced by algorithms. The National Council of Teachers of English has rejected the computerized assessment of writing because “machines cannot judge some of the most valuable aspects of good writing . . . including logic, clarity, accuracy, style, persuasiveness, humor, and irony” (Berrett, 2013, p. 2). Creative exercises that require imagination may not be analyzed correctly by a machine since it cannot provide feedback on rhetorical choices and many aspects of style. Therefore, machine grading is argued to be superficial with a focus on mechanics over meaning. Algorithms can only assess the complexity of word choice, the variety of sentence construction, the rhetorical dimen-
sions of an essay, including the presence of evidence, and the syntax used in simple argument. If teachers do teach to the Common Core writing test, then, dissenters contend, the scripted writing curriculum will emphasize mostly mechanical forms of writing: a correct number of paragraphs, use of transition words, intact thesis statements, three main details, parallel constructions — “this is what they said” and “this is what I say” — and the use of supporting quotes from authority figures.

The PARCC Writing Rubric for Grades 6-11 has these evaluation criteria: comprehension of key ideas and details, development of ideas, organization, clarity of language, and knowledge of language and conventions. Nowhere can one find a criterion that assesses the student’s authorial voice or writing style. In order to receive the highest grade using the PARCC writing rubric, the student must cite specific details from the text itself, use domain-specific vocabulary, develop coherent paragraphs, and show grammar usage — all of which can be evaluated easily with an algorithm. What is lost are original ideas in writing that carry conversations, engage in debate, play with words, and pose questions that have not been asked.

Heads down

At the same time, the question of validity has been challenged by others like Marc Bousquet, a blog writer for The Chronicle of Higher Education, who argues that machines can score essays well because so much of human scoring is also mechanical (2012): “It’s reasonable to say that the forms of writing successfully scored by machines are already-mechanized forms — writing designed to be mechanically produced by students, mechanically reviewed by parents and teachers, and then, once transmuted into grades and the sorting of the workforce, quickly recycled.”

Supporters of machine grading say human graders are not necessarily spending large amounts of time giving meaningful feedback to students on their writing and that human graders are not always invested in the student’s academic success. Upon reflection, I think back to 2004 when I was a part-time evaluator for the state of Illinois and was housed in a hotel for several days where I quickly scanned thousands of students’ writing, knowing I was overlooking individualized aspects of each student’s writing style, grading robotically with a mass of others like me — all with our heads down.

References


“IT’s my model of an atom. I made it actual size.”