

2014 Conference on College Composition and Communication (CCCC)  
March 19-22, Indianapolis, IN

**J.22 (March 21, 2:00-3:15pm)**

**Machine Assessment of the Essays/Essays of the Machine Assessment: Perspectives on the Role of Machine Grading in the Writing Classroom and Its Potential Effects on Pedagogy, Power, and Politics**

**Jason Tham, Jack Hennes, Matthew Barton, & Jaime Heiman**

This session focuses on different perspectives about the potential effects, practical uses, and underlying issues surrounding the use of machine assessment of student writing in the first-year composition college classroom.

---

**Speaker One: Jason Tham**

**“Is There a Robot in This Class?”: The Mechanization of Student, Teacher, and Text in the Writing Classroom**

While mechanizing the evaluation of writing may stem from legitimate concerns about the cost-effectiveness of human graders assessing large numbers of writing assignments, the employment of grading machines to ease these burdens may come at too great a cost. To this end, Douglas Hesse writes that when it comes to the resulting problems of cost and drudgery associated with traditional writing assessment, he worries more about how computer grading will negatively impact “the transferability of skills and about the cost to how students perceive writing (and the resultant cost to human culture).” Despite the present issues that institutions and teachers currently face in regards to cost and time, many teachers believe that their best instruction happens when actively responding to (rather than passively grading) student writing. Speaker One addresses the impact on pedagogy, process, and product that grading machines may have if they replace human readers. More specifically, through a discussion of humanistic pedagogy—particularly in the practice of teaching rhetoric to create better writing and better thinkers—this presentation illustrates how mechanized evaluation of writing may mechanize pedagogy as well as the process to create a written, machine-based product.

---



**“Is There a Robot in This Class?”:  
The Mechanization of Student, Teacher, and Text in the Writing Classroom**

**Jason Tham**

Automated essay scoring (AES) is the use of specialized computer programs to assign grades to essays written in an educational setting. It is a method of educational assessment and an application of natural language processing. Its objective is to classify a large set of textual entities into a small number of discrete categories, corresponding to the possible grades—for example, the numbers 1 to 6. Therefore, it can be considered a problem of statistical classification.

- *Wikipedia*, “[Automated essay scoring](#)”

As we know, the conversation on machine grading isn’t new. It began in 1962 with Walter Reitman’s thoughts on the possibility of machine grading at a conference on needed research of teaching English. Most historical summaries of AES trace the origins of the field to the work of Ellis Batten Page. In 1966, Page argued for the possibility of scoring essays by computer, and in 1968 he published his successful work with a program called [Project Essay Grade \(PEG\)](#).

Multiple-choice exams have, of course, been graded by machine for a long time, but essays are another matter. When it comes to English composition, the question of whether computer programs can reliably assess student work remains sticky. Can a computer program accurately capture the depth, beauty, structure, relevance, creativity, etc., of an essay?

Douglas Hesse writes that when it comes to the resulting problems of cost and frustration associated with traditional writing assessment, he worries more about how computer grading will negatively impact “the transferability of skills and about the cost to how students perceive writing (and the resultant cost to human culture).”

Amid the pressure many composition teachers and their institutions face in regards to the possible adoption of computer-grading programs for assessing student writing, many teachers have strong opinions - most of them disagree with the employment of grading machine in their classroom. In this pressing time, I think it is important for teachers, as well as graduate students, to understand the impact on pedagogy, process, and product that grading machines may have if they replace human readers. Hence, by providing an overview on the appeals of computer grading programs and the ongoing debates around the adoption of automated essay scoring

software for assessing writing, this presentation aims to illustrate how mechanized evaluation of writing may mechanize pedagogy as well as the process to create a written, machine-directed product. I will also offer suggestions for writing instructors in handling the adoption of essay-grading programs in their classrooms.

## **The Appeal**

In Ohio and other states, computer essay graders are on the rise. This is mainly because these states plan to move all their state tests online as part of the transition to a new shared curriculum called the [Common Core State Standards](#). In brief, the Common Core requires students to do more information analysis than regurgitation. The groups developing the new, Common Core tests want them to assess students' writing and reasoning abilities, not just their ability to color in the circles on a multiple-choice answer sheet. The idea is that switching from human to computer graders could make administering these tests cheaper for states nationwide.

Programs that are designed to read and evaluate students are being heavily marketed to K-12 and college administrators and teachers. Criterion, an Educational Testing Service (ETS) product, is recently seen marketed directly to college students through college bookstores. Advocacy and research groups say partially replacing human graders with computer grading programs could help make those tests cheaper, and get results back to students and teachers more quickly. Sure, robots may be cheaper and more efficient. But are they better graders?

## **The New “[Deep Blue vs. Kasparov](#)”**

During the Fourth International Conference on Writing Research held at George Mason University, ETS presented evidence that a pilot test of automated grading of freshman writing placement tests at the New Jersey Institute of Technology (NJIT) showed that computer programs can indeed be trusted with the job. NJIT had freshmen write answers to short essay prompts and used four prompts, arranged in various configurations of two prompts per student, with 30 minutes to write. The testing service compared the results of E-Rater evaluations of students' papers to human grading, and to students' scores on the SAT writing test and the essay portion of the SAT writing test (which is graded by humans). ETS found very high correlations between the E-Rater grades and the SAT grades and, generally, to the human grades of the placement test.

A 2012 study led by Mark Shermis, dean of the University of Akron's College of Education, posits that a computer grading program can score a student essay just as well as a human grader.

*The results demonstrated that overall, automated essay scoring was capable of producing scores similar to human scores for extended-response writing items with equal performance for both source-based and traditional writing genre.*

The study compared the results from feeding student essays from six states into nine different essay scoring software programs and compared the programs' scores with those produced by human graders. Eight of the nine programs were commercial; the other program was a free, open-source software package called [LightSide](#), developed at Carnegie Mellon University. Together they represent nearly all of the available automated essay-scoring options, according to the study.

The differences, across a number of different brands of AES and essay types, were minute. "In terms of being able to replicate the mean [ratings] and standard deviation of human readers, the automated scoring engines did remarkably well," Shermis said in an interview with Steve Kolowich, a staff reporter covering technology for *The Chronicle of Higher Education* and *Inside Higher Ed*.

But not everyone agrees with Shermis.

Les Perelman, director of the Writing Across the Curriculum program at MIT, is among the skeptics of machine-driven assessment. Perelman recently tried out a computer essay grading program made by ETS. In an experiment, he intentionally put in some comma errors and found out that the grading program did not notice them.

In his study, a critique of the Akron study, he concludes that "Even with an experimental design that consisted of different measures for human and scorers and that privileged the machines in half the data sets, the study clearly does not demonstrate that machines can replicate human scores."

Perelman argues that any student who can read can be taught to score very highly on a machine-graded test. That's because software developers build the computer programs by feeding in thousands of student essays that have already been graded by humans. By identifying the elements of essays that human graders seem to like, the programs create an algorithm used to grade new essays. If human graders give essays with long sentences high marks, for example, the programs will tend to do so, as well. If human graders like big words, the programs will do as well.

As such, machine scoring design is incredibly rhetorical and political.

Perelman concludes that if a student knows what the elements of an A essay are and is able to integrate them in the machine-graded test, he or she will likely to score well in the test.

If the reader privileges a certain kind of writing (formal academic writing over African-American English, for instance), we can see how a number of minority test-takers might no longer be situated to do well on these tests.

## A Double-Edged Sword

Shermis says that on shorter writing assignments the computer programs matched grades from real live humans up to 85 percent of the time. But on longer, more complicated responses, the technology didn't do quite as well. He agrees that the technology is not perfect for every situation.

He ran the Gettysburg Address through one of the earlier-generation computer grading programs, one usually used to evaluate the writing abilities of college freshmen. Let's just say Abe didn't do very well. On a scale of 1 to 6, Abraham Lincoln, one of the greatest presidents of the United States, was scoring only 2s and 3s.

It is apparent that the program couldn't handle more complicated responses due to the missing context. The Gettysburg Address is a genre far different and complex from the short essay responses that machine-scoring algorithm may be designed for.

There is, of course, an argument for the context of the speech as well. NPR says that a history professor told Shermis he shouldn't worry; the speech is more famous for its context than for the actual words themselves.

Shermis' Arkon study notes that just because the computer programs agree with the humans, doesn't mean they're right:

*Agreement with human ratings is not necessarily the best or only measure of students' writing proficiency (or the evidence of proficiency in an essay)... The limitation of human scoring as a yardstick for automated scoring is underscored by the human ratings used for some of the tasks in this study, which displayed strange statistical properties and in some cases were in conflict with documented adjudication procedures.*

On the one hand, the Arkon study authors note that computer software hasn't yet caught up with humans when it comes to identifying creativity. But while fostering original, nuanced expression is a good goal for a creative writing instructor, many composition instructors might settle for an easier way to make sure their students know how to write direct, effective sentences and paragraphs ([Steve Kolowich](#), 2012). Shermis and his co-author do not think that replicating human intuition for identifying creativity is a huge concern at this point.

"If you go to a business school or an engineering school, they're not looking for creative writers," Shermis says. "They're looking for people who can communicate ideas. And that's what the technology is best at" evaluating.

This is, in my opinion, a rather contentious claim. What does it mean when we say creativity is not valued in the social sciences and business sectors? Should we really compromise creativity

for the sake of efficiency? I don't think we should prepare a generation of students to become simply laborers just so they can contribute to the economic systems as efficient subjects.

In the end, Shermis advises thrift-minded administrators and politicians not take his results as ammunition in a crusade to replace composition instructors with automated essay scoring software. Ideally, educators at all levels would use the software "as a supplement for overworked [instructors of] entry-level writing courses, where students are really learning fundamental writing skills and can use all the feedback they can get."

### **On Mechanizing the Writing Classroom**

In higher education, AES products are still used primarily to grade students' writing on standardized tests and placement exams, and have not yet found their way into many composition classrooms, according to Perelman. But with demand for writing education rising due to a surge in enrollments among non-native English speakers, triumphant studies such as the Akron researchers' might force some overenrolled, understaffed community colleges to consider deploying AES for their composition classes.

As writing instructors, how should we respond when our school system considers to adopt one of these machine-grading programs for our classes?

The following are approaches writing instructors and composition scholars pursue take should they find themselves in such position.

#### **1. Teach students to think rhetorically about literacy and their writing**

Carl Whithaus reminds us that students are already writing on word-processing programs with spell- and grammar-checkers that give them responses to their writing. Instead of condemning these computer-generated responses out of hand, Whithaus argues, we should teach students how to understand the machine-generated feedback. Through the practice of teaching rhetoric, we enlighten our students to create better writing and to become better thinkers.

Dennis Baron wrote in *The Chronicle of Higher Education* 1998 that both the history of reading and our everyday experience confirm that even the most rigorous and attentive of readers will disagree over texts.

*If we've learned anything about reading and writing in the past 20 years, it is that each person brings to the activities such individual perspectives that it's amazing readers can agree on any aspect of the meaning or value of texts at all. Heraclitus, commenting on the vagaries of life in Greece in the 6th and 5th centuries B.C.E., said you can't put your foot into the same river twice, since it is always changing. Some wag amended this to claim you can't even put your foot into the same river once.*

*No two readers approach a text in exactly the same way; even the same person reading a text repeatedly will come away with different feelings each time. Like it or not, students must learn that their human readers will be inconsistent, a hard and frustrating lesson.*

By encouraging human assessment, we also encourage students to become engaged in literacy learning, to reflect on their own reading and writing in productive ways, and to set respective literacy goals. As such it promotes recognition for the complexity of literacy in today's society and reflect that richness through holistic, authentic, and varied writing instruction.

## 2. Propel research on humanistic pedagogy and computer-grading programs

While existing research claims to demonstrate the value of automated grading, current information is insufficient to use in college writing evaluations. As composition scholars, we are in a position to be proponents of continued research in humanistic education and explore the potential effects of machine grading in the writing classroom. Further studies need to be conducted to measure both teachers' and students' attitudes about computer grading and other effects of artificial testing on the development of writing in the classroom. Often, the results of these studies can affect the livelihoods of teachers, the fate of schools, and the educational opportunities for students.

Perelman says the real problem isn't replacing human teachers with computers. It's that the human graders for most standardized writing tests take about two minutes per essay. That's really not much better than having a computer grade them. He says, what you want is real, live human teachers taking their time with each student's work.

In other words, put humans back into the humanities. An ideal, humanistic writing condition should allow students the opportunity to think, read, talk with others, address real audiences, develop ideas, and revise their emerging texts over time. Computer scoring removes the purpose from written communication — to create human interactions through a complex, socially consequential system of meaning making — and sends a message to students that writing is not worth their time because reading it is not worth the time of the people teaching and assessing them.

## 3. Advocate for the true purposes of writing

Ed White - Writing to a machine is not writing at all... In the composition classroom, we teach our students to write for human being, for human purposes. Building a text for a machine-scoring program is not what we aim to teach as "writing" but some other activity, perhaps one that is more closely related to game-playing than to human communication.

Writing is a highly complex ability developed over years of practice, across a wide range of tasks and contexts, and with copious, meaningful feedback. Students must have this kind of sustained experience to meet the demands of higher education, the needs of a 21st-century

workforce, the challenges of civic participation, and the realization of full, meaningful lives ([Valerie Strauss](#), 2013).

Another resource to look to is the [CCCC Position Statement on Teaching, Learning, and Assessing Writing in Digital Environments](#) from the College Composition and Communication Conference (CCCC), a subset of the National Council of Teachers of English (NCTE). This position statement is unequivocal: "Writing-to-a-machine violates the essential nature of writing."

Other research-based guidelines for effective teaching and assessment of writing include:

- Standards for the Assessment of Reading and Writing (rev. ed., 2009)
- Framework for Success in Postsecondary Writing (2011)
- Framework for 21st Century Curriculum and Assessment (2008, 2013)

In conclusion, effective writing assessments need to account for the nature of writing, the ways students develop writing ability, and the role of the teacher in fostering that development. As a proponent of humanistic pedagogy, I think we should not allow machine scoring to be used because it confronts the values of our pedagogical beliefs.

---

## References:

Les Perelman (2012). [Critique of Shermis and Hammer's study](#)  
Molly Bloom (2012). [Computers Grade Essays Fast ... But Not Always Well](#)  
Steve Kolowich (2012). [A Win for the Robo-Readers](#)  
Molly Bloom (2012). [Computers Can Score Student Essays As Well As Human, Study Finds](#)  
Scott Jaschik (2011). [Grading Essays: Human vs. Machine](#)  
Molly Bloom (2012). [The Pros and Cons of Using Computers to Teach Students How to Write](#)  
Digital Is (2014). [Machine Scoring of Student Writing](#)  
Valerie Strauss (2013). [Can Computers Really Grade Essay Tests?](#)  
Dennis Baron (1998). [When Professors Get A's and Machines Get F's](#)

## More Resources:

CompPile - [Machine Scoring of Essays and Computer Analysis: A Bibliography](#)