Language Arts Journal of Michigan

Volume 23
Issue 1 *Digital Technology*

Article 14

2007

The Devil in the Machine: Problems with Computerized Writing Assessment

Nancy G. Patterson
Grand Valley State University, Grand Rapids, MI

The Devil in the Machine: Problems with Computerized Writing Assessment

Nancy G. Patterson
Grand Valley State University
Grand Rapids, MI

Years ago when Michigan was still piloting its statewide writing assessment, I volunteered to score the eighth grade tests. All eighth graders were to be tested the following year, and I wanted to get a look at the beast that would descend on my students. Approximately seventy readers scored 18,000 writing samples in four days. I read a lot of mediocre pieces of writing during those four days. Many of them had been written to a formula.

Late in the afternoon on the second day, there was a huge pile of tests in the middle of the round table I was sharing with six other teachers. The tests were bundled into sets of twenty. I could tell that most of the tests in each packet came from a single classroom, either because of the formula that teachers mistakenly thought produced good writing or because the students referred to each other in the pieces they wrote in response to a particular prompt. For piloting purposes, not all participating schools wrote to the same prompt. I was plowing through another packet, and it was obvious that the students in the classroom that the packet represented did not have much experience with writing.

There was one piece of writing that was different. It had a lot of surface errors, but the voice shouted from the page. The writer told a lively story about sneaking out at night and breaking into an "enemy's" house. The narrator and a friend were caught "mid-window" in their attempt to break in, and each began blaming the other for hatching the plan. I remember laughing out loud as I read in this silent room, disrupting the serious business of test scoring. But the writing was wonderful even though there were no indentions for paragraphs, and the spelling and phrasing indicated the writer probably spoke a dialect. There was a perception in the writing about what it meant to be thirteen years old and caught in the middle of a bad decision. The writer had that wonderful ability to step outside herself and observe the world from someone else's perspective, and yet slip back behind her own lenses to report her own joy in deliberate "outlaw" behavior and embarrassment and remorse when caught. I continue to assume the story was true because the writer convinced me that it was. My own joy at reading that piece melted away the writer's deviations from standard writing conventions. I gave the piece the highest score possible at the time, and then peeked to see if the other reader had done the same. Yes, he or she had found the same pleasure I had.

That big room filled with silent readers pouring over thousands of tests has become an annual event, though the readers often do not live in Michigan. And, they do not read for free. *The Detroit Free Press* (Higgins) noted that the MEAP costs $42.40 per student, and those costs continue to rise. Part of that cost goes to pay for the test readers and rental of the space they will use. Some legislators continue to think that cost is too high, and more and more of them believe that machine scoring can make the task of Michigan's standardized writing assessment more efficient, both in time and money. In fact, a representative in the MEAP office told me that Michigan could save up to twenty-five percent the first year it implemented computerized writing assessment. Indeed, machine scored writing was piloted in some schools in 2005, but not for high-stakes tests such as the MEAP. However, as costs increase and as money dwindles, computerized writing assessment becomes more and more seductive.

So, why is this scary?

Giving Hal a Red Pen

Remember Hal, the on-board computer in *2001: A Space Odyssey?* The one with the calm voice that said "Good morning, Dave," and ultimately killed the crew in order to protect the ship. Computerized writing assessment is little bit like Hal. It seems innocuous

at first. Companies that sell this kind of software for classrooms promise instant feedback to students and a relief for teachers from the drudgery of grading essays.

Educational Testing Service (ETS), for example, claims that its *Criterion* online essay evaluation software allows students to work independently on assigned prompts and then provides an analysis of errors and advice based on "basic elements of writing —grammar, usage, mechanics, style, and organization and development" ("ETS's Criterion"). However, the computational linguistics programs that dictate correctness in grammar, usage, and mechanics have been with us for some time in the form of grammar and spell checkers in our writing processor programs. We know that those tools are handy, but we also know they are faulty and those green and red squiggles under words and phrases often alert us to something that is not incorrect or stylistically inadvisable. Indeed, how often have we become annoyed with students who rely on spell and grammar checkers rather than proofreading?

Criterion and other computerized writing assessment programs work by comparing a student's piece of writing to thousands of pieces in its data banks that have been ranked and sorted. They also have been programmed to identify language patterns, writing conventions (such as capitalization and punctuation), and usage issues.

They cannot, however, read for meaning. That became evident when Dayton, Ohio *Daily News* reporter Scott Elliott decided to test the machine. He wrote a piece of gibberish about a purple imaginary friend that, though grammatically correct, made no sense, and though he used the appropriate transitional words at the beginnings of paragraphs, the paragraphs themselves made no sense. The program, used in a middle school in Dayton, gave him a six, the highest score possible. When Elliott fed a well-crafted essay he had written into the computerized assessment program, he received a one, the lowest rating. Elliott then showed his two essays to an English teacher who gave the nonsense essay a one, and Elliott's well-crafted essay a six (Patterson 57).

In other words, Elliott fooled the machine.

But that's not all.

A Virginia parent active in the anti-testing movement decided to test the ETS scoring engine and typed, word for word, the opening paragraphs of Stephen King's novella *Hearts in Atlantis*. The machined burped back a three, the next to the highest rating. She then typed those paragraphs again, pasting them to the first document she uploaded. This meant she submitted the opening paragraphs, typed twice, on a single document so that the revised piece had twice as many paragraphs and the original, but the second half was a duplicate of the first half. ETS's software gave it a perfect score of four. The woman's son uploaded a piece of writing, too, and received a four. The piece contained eight paragraphs—with one sentence repeated over and over again: "I just want to see if this computer program worked" (Patterson 56).

Tim McGee found similar problems with Pearson's *Intelligent Essay Assessor* (IEA). Pearson's "brain" behind its online writing assessment is its "Knowledge Analysis Technologies (KAT)" and claims that it "evaluates the meaning of text by examining whole written passages" (http://www.pearsonkt.com/prodKAT.shtml). Tantalized by Pearson's claims, McGee took a sample essay on the circulatory system and typed it backwards, meaning that the last sentence became the first, etc. He reversed the order of the thirteen sentences in the essay. IEA awarded the backwards essay with the exact same score it had assigned the original, even though the backwards essay made no sense, and, indeed, because the sentences were in the wrong order, offered inaccurate information (87).

Despite the claims of marketers, machines cannot read for meaning, nor can they provide a human response to human language.

But there is more. McGee entered another essay on the Great Depression into IEA and earned a top score of five. The original essay began with "There were many problems facing the nation in 1938, following the stock market crash in 1929 and in the midst of Franklin D. Roosevelt's New Deal." But the second version of the essay that McGee entered into the program reversed the facts. So, the first sentence said, "There were **few** problems

facing the nation in **1929**, following the stock market crash in **1938** and at the **end** of Franklin D. Roosevelt's New Deal." The essay continued in this fashion, reversing or altering all of the facts, and the altered essay earned the same high score as the original (89).

McGee then tried one more trick. He took a psychology essay that discussed the effects of a stroke on a victim. The original earned a seven out of ten score. McGee "revised" the essay, writing grammatically correct gibberish. The original essay included this sentence: "To detect the effects that Mr. McGeorge's stroke had I would conduct several experiments testing his ability to communicate"; this section of the altered version read: "To effect the detects that Mr. stroke McGeorge had I would several conduct experiments testing ability his communicate to" (89).

The altered essay earned a six rather than a seven.

I might add that the grammar checker in my word processing program saw no reason to underline that sentence, except to question the spelling of "McGeorge." Neither Intelligent Essay Assessor nor Microsoft could do what a human being could. Despite the claims of marketers, machines cannot read for meaning, nor can they provide a human response to human language.

Nothing points more dramatically to the need to maintain the human response to writing than Maja Wilson's account of her experience with Criterion. In her essay "Apologies to Sandra Cisneros," Wilson talks about her students' reactions to Cisneros' "My Name." Wilson loved to read Cisneros' piece aloud and listen to her students' reactions, to the ways in which they identified with Cisneros, questioned her, and tussled with the memories that Cisneros' powerful writing brought about in them. They responded to Cisneros' voice in very human ways.

Wilson decided to test Criterion's ability to respond to Cisneros' writing, and so applied for a guest account on the ETS website. When it came through, she typed "My Name" word for word into *E-Rater*, Criterion's assessment engine. ETS markets Criterion and E-Rater as a way to provide immediate feedback to student writers. Criterion and E-Rater didn't like "My Name." The first thing Wilson noted was that the computer offered no praise to the writer. Wilson, a high school writing teacher, always finds something positive to say to her students before she guides them into rethinking their writing. Criterion faulted Cisneros for using too much repetition, and for problems with sentence fragments and organization.

So Wilson decided to revise Cisneros' work using Criterion's feedback. She combined shorter sentences to make longer ones. And, at Criterion's suggestion, she created a thesis statement. In fact, she ended up writing a five-paragraph essay, something Criterion said was the foundation of good writing. But did Wilson improve on Cisneros' work? According to Criterion, yes.

You be the judge (see Table 1).

Table 1

Cisneros' Original First Paragraph	***Criterion's* Revision**
In English my name means hope. In Spanish it means too many letters. It means sadness, it means waiting. It is like the number nine. A muddy color. It is the Mexican records my father plays on Sunday mornings when he is shaving, songs like sobbing.	Names mean different things in different languages. (http://www.rethinkingschools.org/archive/20_03/apol203.shtml)

Criterion's suggested revision is grammatically correct. But it is hardly lyrical. It cannot move students to respond in a personal way to the experience of a little girl whose name has too many letters and whose father shaves to sad tunes he plays on an old record player.

I cannot help but wonder what Criterion would have done to the piece of writing I read years ago when I was scoring MEAP writing tests. I wonder if that student's voice would have leaped from the page into my ear and lingered through the years had Criterion gotten its computer chips involved in her writing.

Good Marketing Does Not Mean Good Teaching

Julie Cheville, in her 2004 *English Journal* article warns that private interests are threatening the foundation of meaningful classroom writing practice. Programs like ETS' Criterion and *Intellimetric*, billed by Vantage Learning as the "Gold standard for automated essay scoring," cannot recognize good writing. They can recognize "inappropriate words or phrases, sentences with passive voice, long sentences, short sentences, sentences beginning with coordinating conjunctions" (ETS qtd. in Cheville 48). But who decides what is inappropriate? And is it always bad to use a sentence fragment? Or passive voice? Or to begin a sentence with a coordinating conjunction? Cheville challenged ETS and its concepts of style, and was told that "computational linguists had not yet developed an analytic capability beyond parts of speech and simple phrases. The scoring engine was unable to identify clausal structures central to stylistic maturity" (48).

So, the programs themselves are flawed, and it is unlikely that computational linguists and artificial intelligence developers will be able to create programs that duplicate human cognition.

Still, publishers and creators of machined-scored essay software point to the reliability of their products. They have neatly provided statistics that show the close correlation between human and machine scorers. But the writing that is scored by humans in these instances is formulaic. E-Rater, marketed by ETS as a college entrance placement assessment, claims a ninety-four percent accuracy rate. They base this on the fact that E-Rater agreed with two university professors who rated thousands of tests in 1997 (Enbar).

Patterns and Drudgery

Because these programs can only recognize patterns, they will privilege formulaic writing over writing that falls outside the prescribed pattern. When students write to a formula, they have sacrificed the power of process to the gods of product. Writing is the act of addressing audience and purpose. It is the art of decision making. Formulas remove that decision-making process from students, the very people who should be getting as dirty as possible in the mud of writerly decisions. Formulas are not training wheels for inexperienced writers, but prisons that limit writing development and the ability to learn through and with language. The rise of standardized writing assessment and machine scoring has carved a deeper space for formulaic writing in the classroom. This must stop.

Computerized writing assessment promises to relieve teachers from the "drudgery" of grading papers. Indeed, this was one of the claims that promoters used to sell *My Access!*, part of Vantage Learning's writing assessment "environments," to a western Michigan school district. Teachers in that district report that while their drudgery may have decreased, student engagement in writing has plummeted, primarily because the program privileges writing to a formula.

My Access!, like the other computerized writing assessment programs, promises an increase in standardized writing assessment scores through immediate feedback and substantive comments. But at what cost? If reading student writing is drudgery, what does this say about the kinds of writing teachers assign and the learning environment in which that writing takes place? And what must it be like to create that kind of writing? What are we teaching students about writing if what they produce is drudgery for teachers? And for themselves? As a profession we need to ask ourselves why we want students to write, and the answer has to be better than "They need this for college." Or, "They need this for a job." They also need to write to discover who they are and how they fit in a world that changes as they mature. They need to think within the context of written language, not according to some rotting definition of a literary criticism essay or formulaic weak facsimile of an argument. They need to write in order to learn more about themselves and the world around them.

Vantage Learning promises that its software, Intellimetric, will make writing instruction more efficient and provide teachers with more time. But Bob Broad argues that new technologies that bill themselves as time-saving devices never quite fulfill that promise (223). Broad compares computerized essay assessment to cooking technologies. Wood burning stoves were hailed as a time-saving device for women, and indeed, they used less wood which meant that less time went into chopping wood and

maintaining an open hearth. But when the next technology of a wood-burning cook stove came the expectation that a woman should cook more than single pot meals. She had the technology to create more dishes and the expectation to make sure all those dishes were done at the same time and delivered to the table.

Broad points out that classroom labor saving devices like automated writing assessments will not ultimately save teachers time. School leaders will simply add to teachers' tasks (223). Broad also argues that humans need to respond to human writing. Computerized writing assessment "would trivialize and denude [writing] instruction and experience," and he urges educators to fight the use of it. Richard Haswell argues that we must resist the notion that responding to student writing is drudgery; rather, it is a "difficult, complex and rewarding skill requiring elastic intelligence and long experience" (77). He adds, "Good diagnosis of student writing should not be construed as easy, for the simple reason that it is never easy" (77). I am reminded of the Tom Hanks character in the film *League of their Own.* The fact that baseball is hard makes it worth doing. Giving substantive feedback to writers is not easy. If it were, anyone could do it.

The Audience as Hal

For the first time in the history of writing our students can write to an audience that is not human. What implications will this have on their perceptions as writers, on their identities as writers? If the purpose of writing is to address the needs of an audience for a given purpose, even when the audience is the writer, how will machine as audience mechanize writing? How will it stifle good writing and good pedagogy? What would a machine have done to Sandra Cisneros? Or to the anonymous MEAP writer?

Broad writes, "Victory in this struggle will depend on our ability to link the pedagogical (including assessment) practices we promise to a compelling portrait of what [writing] is, why [writing is] important to our society, and what it means to be human and literate, a portrait that clearly demonstrates the necessity of human relationships and interactions" (233).

One afternoon I sat in a windowless room and read a piece of writing that charms me to this day, fifteen years after I read it, in the midst of thousands of pieces of writing. It warms me still. I don't know who the author was or where in Michigan she lived. But I wish her well and hope that she continues to lift her voice. And I hope that the likes of Criterion and Intellimetric and My Access! and Intelligent Essay Assessor, and all the other computerized assessment programs that can only really promise big profits but never good pedagogy and assessment, go the way of the Edsel, video disc players, and eight-track tapes. Wilson points out that automated writing assessment is really about a lack of commitment to smaller classes and the professional lives of teachers: "If they trusted teachers to teach and if they trusted students to think and question, they'd be out of a job."

Good writing is good thinking. And good thinking moves across the waters of imagination and creativity, experience and emotion. These cannot be sorted into hierarchies waiting to be identified and retrieved by a roving bot on transistor-powered motherboard.

It takes a teacher.

Works Cited

Broad, Bob. "More Work for Teacher? Possible Futures for Teaching Writing in the Age of Computerized Assessment." *Machine Scoring Of Student Essays: Truth or Consequences.* Eds. Patrician Freitaq Ericsson and Richard Haswell. Logan, UT: Utah State University Press, 2006.

Cheville, Julie. "Automated Scoring Technologies and the Rising Influence of Error." *English Journal* 93.4 (2004): 47-52. Educational Testing Service. "ETS's Criterion v 7.1 Allows Students to Plan Their Essays Online." (3 Jan. 2007): http://www.ets.org/portal/site/ets/menuitem (see Criterion Online Writing Evaluation: News).

Educational Testing Service. "Phoenix Union High School District to use ETS' Online Essay Scoring Service." (2007). 8 July 2007 <www.ets.org>.

Enbar, Navda. "This is E-Rater. It'll be Scoring Your Essay Today." *Business Week Online* 21 Feb. 1999. Retrieved 8 July 2007 <http://www.businessweek.com/bwdaily/dnflash/jan1999/nf90121d.htm>.

Haswell, Richard H. "Automatons and Automated Scoring: Drudges, Black Boxes, and Dei Ex Machina." *Machine Scoring of Student Essays: Truth or Consequences*. Eds. Freitag Ericsson, Patricia, and Richard H. Haswell. Logan, UT: Utah State University, 2006.

Higgins, Lori. "High Schoolers Act Could Help Kill MEAP." *Detroit Free Press* 5 May 2004: 1.

McGee, Tim. "Taking a Spin of Intelligent Essay Assessor," *Machine Scoring of Student Essays: Truth or Consequences*. Eds. Freitag Ericsson, Patricia, and Richard H. Haswell. Logan, UT: Utah State University, 2006.

Patterson, Nancy G. "Computerized Writing Assessment: Technology Gone Wrong." *Voices from the Middle* 13.2 (2005): 56-57.

Wilson, Maja. "Apologies to Sandra Cisneros: How ETS' Computer-based Writing Assessment Misses the Mark." *Rethinking Schools Online* 20.3 (2006). 8 July 2007 <http://www.rethinkingschools.org/archive/20_03/apol203.shtml>.

About the Author

Nancy Patterson (patterna@gvsu.edu) is an assistant professor at Grand Valley State University and chair of the Reading/Language Arts Program in the College of Education. She taught secondary English for almost 30 years.