REVIEW

John Henry vs. the Machine: *Machine Scoring of Student Essays: Truth and Consequences.*

JACOB S BLUMNER, UNIVERSITY OF MICHIGAN, FLINT

Patricia Freitag Ericsson and Richard H. Haswell, eds.
Machine Scoring of Student Essays: Truth and Consequences.
Logan, UT: Utah State University Press, 2006. 274 pages. $24.95.

I THOUGHT IT WAS serendipitous that I was listening to Johnny Cash's "Legend of John Henry's Hammer" when I received a copy of *Machine Scoring of Student Essays: Truth and Consequences,* edited by Patricia Freitag Ericsson and Richard H. Haswell. As I skimmed the book's table of contents and glanced at random pages, I thought about the parallels between Henry's story and that of writing teachers. Certainly the comparison breaks down on many levels: for example, I don't imagine English teachers' hearts giving out from working so hard. But, the obvious comparison of "man vs. machine" proves accurate.

Though the book is primarily directed toward writing teachers, for a WAC audience the book serves an excellent purpose as well. WAC coordinators can use it to inform faculty who have questions about the value and process of machine scoring, particularly when faculty are understandably trying to find ways to ease a heavy workload. In fact, many faculty ask about machine scoring because they have heard about it in the news, have received promotional materials from a company offering machine scoring, or

DOI: 10.37514/WAC-J.2006.17.1.06

from others in academia. For faculty across the disciplines, the book provides valuable insight into what writing teachers value in writing, how one goes about valuing it, and what machines currently can and cannot do when assessing it. There is much to learn in this book for all teachers, administrators, and community members.

In the introduction of *Machine Scoring,* Ericsson and Haswell approach the topic of machine scoring with a true spirit of inquiry, listing questions they will address in the book and claiming this is the first significant volume that provides a voice to teachers and students—two constituencies that have not had a voice in the debate of this "emerging technology." Chapters in the book cover topics ranging from Ken S. McAllister and Edward M. White's book-opening history of machine scoring in "Interested Complicities: The Dialectic of Computer-Assisted Writing Assessment," to practical, application-focused chapters such as Edmund Jones' "ACCUPLACER's Essay-Scoring Technology," and finally to Bob Broad's examination of the future implication of machine scoring in "More Work for Teacher? Possible Futures of Teaching Writing in the Age of Computerized Writing Assessment." It ends with a thorough bibliography by Richard Haswell that is helpful for readers who want to better understand the development of machine scoring and begin to track what the future might hold. In many ways, *Machine Scoring* is a response to Mark Shermis and Jill Burstein's edited collection, *Automated Essay Scoring: A Cross-Disciplinary Perspective,* a book in which nearly all the contributors are involved with the machine scoring industry. *Machine Scoring* lays a parallel track to Shermis and Burstein's book, providing writing teacher/scholar perspectives on the role of computers in writing assessment.

McAllister and White acknowledge writing teachers' apathy in addressing the growing demand and fiscal appeal of machine scoring. They note that writing teacher's voices have been nearly silent, allowing a commercial industry to grow up and tap into legislative, administrative, and public demands for "objective" accountability with a small price tag. Neglecting to respond, they note, has let the industry build a head of steam that appeals to the aforementioned constituencies. Much of the chapter, though, is dedicated to explaining the foundation upon which electronic scoring is based: formalism and natural language processing, a linguistic process used to gather content information from texts. Natural language processing was not developed to assess writing, but according to McAllister and White, many commercial programs designed to assess text use it. Then, after introducing the general process used for machine scoring, McAllister and White describe why formalism and natural language processing fail to accomplish what evaluators

of college writing want. The chapter is a well-crafted key to understanding and appreciating the rest of the book.

A pattern in each chapter emerges in the middle of the book. A scholar examines an automated scoring program, explains that he or she can only guess the criteria by which a program evaluates text because the company won't reveal proprietary secrets of their programming, and submits essays to the program to try and finesse an understanding of how the program works. I must admit I wasn't surprised by many of the results. Most of us have heard stories of teachers testing these programs, creating nonsense essays, submitting them, and receiving excellent results. You can do a small test for your own amusement to see how this works. Create a Microsoft Word document in which instead of typing words, use the letter "x" repeatedly. So your text might look like this, "Xxxx, xxxx xxxxxxx xx xxxxxx, xxxxx xxxxx xxxx, xxx." Run the spelling and grammar feature, simply ignoring all of the misspellings and grammar suggestions. What you will see at the end is a box that includes "Counts," "Averages," and "Readability," an evaluation using the Flesch-Kincaid system of evaluation. The nonsense quote above had 0% passive sentences, a Flesch reading level of 28.5 and a Flesch-Kincaid Grade Level of 11.5 (out of 12). Not bad for a bunch of x's.

Of course ACCUPLACER, WriterPlacer, and E-Write are more sophisticated than the evaluation Microsoft Word performs, but the results relayed in this book are not that different. Still, the results are enlightening about work that needs to be done so interested parties can better understand what the programs are and are not capable of accomplishing. Despite my lack of surprise, the chapters in which teachers test the programs are my favorites because they begin to pry open the black box of computer scoring. In Jones's chapter, his tests of ACCUPLACER reveal ridged, narrow assessment techniques. Jones tested a paragraph that demonstrated that ACCUPLACER looks for grammar errors but cannot adequately judge syntax or usage problems. Jones digs deeper to find that ACCUPLACER is only good at certain kinds of sentence-level problems, but not mechanical ones. Another discovery is that ACCUPLACER values text length—400 words or more. These results lead readers to question what is valuable in writing and what is lost by using this and other machine scoring programs.

Richard N. Matzen Jr. and Colleen Sorensen describe Utah Valley State College's research into placement tests in "E-Write as a Means for Placement into Three Composition Courses." Their experience is disturbing and hilarious; they experienced technical difficulty after technical difficulty, from essays receiving no

scores to e-Write's server crashing. Ultimately, they concluded, "the validity of e-Write scores is questionable. If the e-Write scores had been used for placement purposes, for example, apparently only 4 of 298 students would have enrolled in the lower-level basic writing course, an outcome that the experienced basic writing teachers at UVSC believe is inaccurate" (137). Of course these results assume that the server is working. Though not explicitly stated, Matzen and Sorensen's results remind readers that technology needs to work for it to be useful.

One of my favorite chapters that tested software is Tim McGee's "Taking a Spin on the Intelligent Essay Assessor." McGee takes aim at the program's claim that it "is the only essay evaluation system in which meaning is dominant" (80), and tests Intelligent Essay Assessor's (IEA) definition of meaning to his own. IEA provides sample essays, and McGee used them to test the program's understanding of meaning. In one test he looked at a sample essay's test score, and then he reversed the sentences of the essay, noting, "the effect is more like that of the movie *Memento*" (86). The scores were identical. Next, McGee tested the program's ability to measure factual information. He simply changed facts in the model history essay to completely contradict known facts. The opening sentence reads, "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" (88). Again the machine awarded the text a high score. I found myself laughing out loud.

The problem is this isn't simply a laughing matter. Not taking machine scoring seriously led to the apathy described by McAllister and White. With a better understanding of how machine scoring works, readers approach the last section of the book ready to examine the implications for teachers and students. Beth Ann Rothermel, in "Automated Writing Instruction: Computer-Assisted or Computer-Driven Pedagogies?," asserts that machine scoring companies, MY Access! in particular, "show … disdain for classroom teachers working at the primary and secondary levels" (199) and that they have an ideology "that defines not just writing, but also teaching and learning, as formulaic and asocial endeavors" (200). MY Access! "constricts and narrows the learning environment" (204), and it won't "say back to the student in its own words what it thinks the student 'means'" (205). William Condon continues the critique in "Why Less is Not More: What We Lose by Letting a Computer Score Writing Samples" by listing the costs of using machine scoring, including the loss of local control; loss of the human element that contributes to professional development and writing program cohesion; and the limitations of a short, timed-writing exam.

And finally, in the last chapter of the book, "More Work for Teachers? Possible Futures of Teaching Writing in the Age of Computerized Writing Assessment," Bob Broad accomplishes two things. He provides a fascinating insight into the misnomer that technology saves us time, and he passionately argues that assessment is an integral part of teaching and that we should value it more and fight to keep it. For the former point, he uses Ruth Schwartz Cowan's book, *More Work for Mother: The Ironies of Household Technology from the Open Hearth to the Microwave,* to show that new technology can unexpectedly create more work. He points to the wood-burning stove and the vacuum as two tools that increased the demands on women in the home rather than alleviating the workload. Broad warns that this can happen to teachers as well. When I consider the increased demands placed upon faculty because of the advent of email, Broad's fears ring true. Broad's latter argument, though, leaves a greater impression. He asks us to consider what our courses are about, to define rhetoric for ourselves, and to ask if machines can measure those things we value in writing. The answer for me is a resounding "no."

As many authors in this collection note, there is a place for technology in the teaching of writing, and we have much to learn about machine scoring. But what is more important is that teachers and students be more involved in the conversation about machine scoring. Pry open the black box and see what makes it work. Try to influence changes in the programs so that they better serve our pedagogical needs. This book is a beginning. It starts to build a body of literature that teachers can use to influence policy decisions. The bigger issue it addresses, though, is what makes us human, and how do we value that in writing. John Henry knew who he was and what he stood for. Teachers need to continue to lay the track that will define us as human and clearly articulate how that manifests in writing.