Machine Scoring of Writing

Because machine scoring cannot validly assess student writing, Congress should not support assessment practices that rely wholly or in part on machine scoring of writing for high stakes outcomes.

A group of professionals concerned about moves toward machine scoring of writing on a large scale summarizes research about machine scoring of writing in the following points:

1. Computer algorithms cannot recognize the most important qualities of good writing, like truthfulness, tone, complex organization, logical thinking, or new ideas germane to the topic.
2. Machine algorithms are reductive: sophistication of vocabulary is reduced to average length or relative frequency of words.
3. Machines over-emphasize or misidentify grammatical and stylistic errors.
4. Machines cannot score writing tasks long and complex enough to represent levels of writing proficiency acceptable in school, college, or the workplace.
5. Machines often require artificial essays finished within very short time frames with length being a major determinant of scores.
6. Machines do not approximate human scores for essays that fit real-world writing conditions.
7. High correlations between human scores and machine scores reported by testing firms are achieved in part through training humans to read like machines or requiring both to use simplistic scoring scales.
8. Machine scoring can be biased against second-language writers and minority writers.

If so much evidence exists to discount machine scoring of writing, why is it used or considered for use? First, cost is indicated as a factor: paying humans to assess writing takes funds that may be hard to find. Yet, if human scoring were regarded as professional development, which research shows it is, the cost factor can be mitigated. Second, time is an issue. Assessments are most helpful if they are done in a timely fashion so that, if necessary, teachers can alter instruction. Yet, insufficient or even inaccurate assessments can lead to poor instruction and less learning. Third, using computers is easier than organizing educators to respond to writing. The potential of computer responses to distort identification of writing quality, however, must override any challenges of organizing for human scoring.

In a research project at North Carolina State University, Professor Chris Anson compared evaluation of writing by computers, paid readers, and classroom teachers. He looked at the way each kind of evaluator encourages deep critical thinking, gives feedback on different analytic features of writing, gives helpful feedback, gives accurate feedback, can respond to irony, can respond to humor, can respond to meaning, judges accuracy of facts/claims, can respond to multiple genres, can judge complex structures, can be gamed, and has high validity. Anson found that accurate assessment of good writing was not accomplished by computers, was sometimes accomplished by paid readers, and was accomplished by classroom teachers.
As the Common Core State Standards assessments are being developed, critical decisions are being made about using human or machine scoring. Because machine scoring is acceptable for a limited number of factors, the writing tasks on a test would need to be written to those limitations, even if the Common Core State Standards point to more complex writing tasks. When states have the opportunity to implement high standards for writing, it will be tragic if machine scoring subverts those standards.

More importantly, however, machine scoring of writing signals to students, parents, and educators that only certain features of writing need to be practiced and taught. The recent *Framework for Success in Postsecondary Writing*, developed by the three main professional organizations in the country centered on writing (National Council of Teachers of English, Council of Writing Program Administrators, and National Writing Project), identifies eight habits of mind essential for success in college writing: curiosity, openness, engagement, creativity, persistence, responsibility, flexibility, and metacognition. To cultivate those habits, students need experience with rhetorical knowledge, critical thinking, writing processes, knowledge of conventions, and ability to compose in multiple environments. Except for conventions, these experiences are not validly scored by machines. When machines replace humans as readers of writing, students are not supported in the kinds of writing that prepare them for their lives as citizens, workers, and learners.

Particularly problematic is that students, teachers, and schools are all judged on the scoring of writing in high stakes ways. Machine scoring relied on for what it can do may be one component among multiple assessments of writing. Because of its limitations, however, it must not be the only or even the primary means by which writing is assessed. Humans only are capable of reading for the most important aspects of writing. Investment in human scoring is warranted to gain validity and to acknowledge all that we know about writing processes and products.

April 12, 2013