To: Chair Paz and Members of the State Board of Education

From: The Oregon Writing and English Advisory Committee (OWEAC)

CC: Karen Marrongelle, Lisa Reynolds, and Doug Kosty

Subject: OWEAC's Official Position Against Machine Scoring of Writing

Date: 26 April 2013

The Oregon Writing and English Advisory Committee is a long-standing group of college and university writing faculty from throughout the state with many years of experience, both as writing teachers and as scholars in the field of composition and rhetoric. OWEAC met recently, and we are unanimous in our stance against machine scoring in the assessment of writing. We understand the appeal of an automated system of assessment, especially when faced with the magnitude of assessing all K-12 students in the state of Oregon for achievement of the Common Core State Standards (CCSS). However, a wealth of research by composition and rhetoric experts has demonstrated that machine scoring falls far short of authentic assessment of writing. For example, machine scoring cannot assess how accurately students have worked with and utilized sources, nor can it evaluate reading comprehension. In addition, the pressure to teach to such tests seriously harms the learning environment of students by narrowing the focus of the classroom to the simplistic features of writing that can be detected by a machine. Please refer to the recent position statement by the National Council of Teachers of English, and to the statement that prefaces the Human Readers petition, for detailed descriptions of the many problems with computerized assessments. Excerpts from these statements, along with links to the complete texts, can be found at the end of this document; they include extensive references to the research that supports the position of experts against machine scoring.[i] For the multitude of reasons listed in these statements, which are endorsed by professionals across the nation, OWEAC is strongly opposed to the adoption of a statewide system of computerized writing assessment for the CCSS.

In a recent memo to all superintendents and principals, Deputy Superintendent Rob Saxton noted that "no official action has yet been taken by the State Board of Education regarding adoption of a particular assessment system" and that "there is time and flexibility to adopt a different assessment system if that is deemed superior for our students. We want to make sure that we are picking the assessment system that is the highest quality and that will best serve our students and our schools." While official action to adopt an assessment system may not yet have been taken, it seems clear from the state-sponsored conversations on the CCSS that the state has already decided to adopt the Smarter Balanced system. We hope that this is not true and that there is still time to find an approach that will serve student learning and provide authentic assessment. Composition and rhetoric experts are in agreement about best practices in assessment, including, for example, the evaluation of portfolios of student writing by local teams of teachers. Such assessment practices not only effectively recognize the writing competencies of students, but also function as part of the learning process, along with fostering professional development of involved faculty. An excerpt from the NCTE position statement that explains evidence-based practices for assessment is included at the end of this document.[ii] Many college and university faculty in the state of Oregon are ready and willing to offer our expertise and assistance with the development of an assessment system founded on best practices that have emerged from years of research and experience.

In the same memo from Rob Saxton, he notes that "Whatever assessment is chosen, it will cost our state significantly more than our current system. This is because the new systems will provide a more authentic assessment.... We believe this will be a worthwhile investment, but we also want to make sure we are getting the best possible system for our money." If machine scoring of writing is implemented, it most assuredly will not "provide a more authentic assessment," and, rather than being a worthwhile investment, it will divert much needed funding away from the classroom. Although we do not have exact figures for the Smarter Balanced assessment tests, one estimate put the cost at twenty dollars per test, per child, which works out, conservatively, to ten million dollars a year.

This money would be much better spent on hiring sufficient numbers of teachers to reduce class size (we are currently in the top five for largest class size in the U.S.), providing more professional development for teachers, and implementing an assessment system that relies on actual readers trained to teach, and therefore detect, the elements of writing and critical thinking expected in K-12 and, ultimately, college.

Oregon education leaders frequently suggest that higher education faculty have played a large role in the development of the CCSS and in the adoption and development of the Smarter Balanced assessment system, but in our view, this is not the case. Our experience is that Oregon signed on to the CCSS before involving higher-education faculty and that faculty input into Smarter Balanced assessments has been invited only at very narrowly proscribed moments. Two Past Chairs of OWEAC and a Past President of the Oregon Council of Teachers of English applied to be involved in Smarter Balanced assessment development and were not included. The one OWEAC member involved was assigned to a middle-school group where her expertise was minimally useful, and the most recent call for feedback on Achievement Level Descriptors (ALDs) was another empty gesture since the overall assessment design is not acceptable. Writing assessment is complicated; corporate testing companies promising computerized simplicity in writing assessment are misleading their clients, and education leaders investing in those promises are wasting taxpayer dollars. We urge the Board to make decisions for the K-12 students in our state that are shaped by the research and expertise of higher-education faculty, not only in Oregon, but throughout the nation. It would be wonderful to see Oregon take a leading role in writing assessment.

The following are active members of OWEAC who support this statement. OWEAC members represent faculty at their respective institutions who also support OWEAC's position against machine grading. The names below are only a fraction of higher education faculty across the state who oppose machine scoring of writing.

Jillanne Michell, Chair of OWEAC, Umpqua Community College
Vicki Tolar Burton, Oregon State University
Sara Jameson, Oregon State University
Siskanna Naynaha, Lane Community College
Kate Sullivan, Lane Community College
Eva Payne, Chemeketa Community College
Michele Burke, Chemeketa Community College
Ryan Davis, Clackamas Community College
Verne Underwood, Rogue Community College
Christopher Syrnyk, Oregon Institute of Technology
Nancy Knowles, Eastern Oregon University
Donna Evans, Eastern Oregon University
Cori Brewster, Eastern Oregon University
Caroline Le Guin, Portland Community College
Matt Usner, Linn-Benton Community College
Jo Cochran, Klamath Community College
Carolyn Bergquist, University of Oregon
Margaret Artman, Western Oregon University
Chris Rubio, Central Oregon Community College

Thank you for the opportunity to provide input on this important process. We are eager to help develop an assessment system based on best practices. If you have any questions, please feel free to contact Jillanne Michell, Chair of OWEAC at jillanne.michell@umpqua.edu / 1-541-440-4646.

[i] The following excerpt is from the "NCTE Position Statement on Machine Scoring":

> To meet the outcomes of the Common Core State Standards, various consortia, private corporations, and testing agencies propose to use computerized assessments of student writing. The attraction is obvious: once programmed, machines might reduce the costs otherwise associated with the human labor of reading, interpreting, and evaluating the writing of our students. Yet when we consider what is lost because of machine scoring, the presumed savings turn into significant new costs -- to students, to our educational institutions, and to society. Here's why:
>
> - Computers are unable to recognize or judge those elements that we most associate with good writing (logic, clarity, accuracy, ideas relevant to a specific topic, innovative style, effective appeals to audience, different forms of organization, types of persuasion, quality of evidence, humor or irony, and effective uses of repetition, to name just a few). Using computers to "read" and evaluate students' writing (1) denies students the chance to have anything but limited features recognized in their writing; and (2) compels teachers to ignore what is most important in writing instruction in order to teach what is least important.
> - Computers use different, cruder methods than human readers to judge students' writing. For example, some systems gauge the sophistication of vocabulary by measuring the average length of words and how often the words are used in a corpus of texts; or they gauge the development of ideas by counting the length and number of sentences per paragraph.
> - Computers are programmed to score papers written to very specific prompts, reducing the incentive for teachers to develop innovative and creative occasions for writing, even for assessment.
> - Computers get progressively worse at scoring as the length of the writing increases, compelling test makers to design shorter writing tasks that don't represent the range and variety of writing assignments needed to prepare students for the more complex writing they will encounter in college.
> - Computer scoring favors the most objective, "surface" features of writing (grammar, spelling, punctuation), but problems in these areas are often created by the testing conditions and are the most easily rectified in normal writing conditions when there is time to revise and edit. Privileging surface features disproportionately penalizes nonnative speakers of English who may be on a developmental path that machine scoring fails to recognize.
> - Conclusions that computers can score as well as humans are the result of humans being trained to score like the computers (for example, being told not to make judgments on the accuracy of information).
> - Computer scoring systems can be "gamed" because they are poor at working with human language, further weakening the validity of their assessments and separating students not on the basis of writing ability but on whether they know and can use machine-tricking strategies.
> - Computer scoring discriminates against students who are less familiar with using technology to write or complete tests. Further, machine scoring disadvantages school districts that lack funds to provide technology tools for every student and skews technology acquisition toward devices needed to meet testing requirements.
> - 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.

The position statement of the NCTE, including an annotated bibliography of the research that supports the above points, can be found in its entirety at http://www.ncte.org/positions/statements/machine_scoring

The following excerpt is from the "Research Findings" section of the Human Readers petition against machine scoring:

> Research findings show that no one—students, parents, teachers, employers, administrators, legislators—can rely on machine scoring of essays:
>
> 1. computer algorithms cannot recognize the most important qualities of good writing, such as truthfulness, tone, complex organization, logical thinking, or ideas new and germane to the topic (Byrne, Tang, Truduc, & Tang, 2010)
> 2. to measure important writing skills, machines use algorithms that are so reductive as to be absurd: *sophistication of vocabulary* is reduced to the average length or relative infrequency of words, or *development of ideas* is reduced to average sentences per paragraph (Perelman, 2012b; Quinlan, Higgins, & Wolff, 2009)
> 3. machines over-emphasize grammatical and stylistic errors (Cheville, 2004) yet miss or misidentify such errors at intolerable rates (Herrington & Moran, 2012)
> 4. machines cannot score writing tasks long and complex enough to represent levels of writing proficiency or performance acceptable in school, college, or the workplace (Bennett, 2006; Condon, 2013; McCurry, 2010; Perelman, 2012a)
> 5. machines require artificial essays finished within very short time frames (20-45 minutes) on topics of which student writers have no prior knowledge (Bridgeman, Trapani, & Yigal, 2012; Cindy, 2007; Jones, 2006; Perelman, 2012b; Streeter, Psotka, Laham, & MacCuish, 2002; Wang, & Brown, 2008; Wohlpart, Lindsey, & Rademacher, 2008)
> 6. in these short trivial essays, mere length becomes a major determinant of score by both human and machine graders (Chodorow & Burstein, 2004; Perelman, 2012b)
> 7. machines are not able to approximate human scores for essays that do fit real-world writing conditions; instead, machines fail badly in rating essays written in these situations (Bridgeman, Trapani, & Yigal, 2012; Cindy, 2007; Condon, 2013; Elliot, Deess, Rudniy, & Joshi, 2012; Jones, 2006; Perelman, 2012b; Powers, Burstein, Chodorow, Fowles, & Kukich, 2002; Streeter, Psotka, Laham, & MacCuish, 2002; Wang & Brown, 2008; Wohlpart, Lindsey, & Rademacher, 2008)
> 8. high correlations between human scores and machine scores reported by testing firms are achieved, in part, when the testing firms train the humans to read like the machine, for instance, by directing the humans to disregard the truth or accuracy of assertions (Perelman, 2012b), and by requiring both machines and humans to use scoring scales of extreme simplicity
> 9. machine scoring shows a bias against second-language writers (Chen & Cheng, 2008) and minority writers such as Hispanics and African Americans (Elliot, Deess, Rudniy, & Joshi., 2012]
> 10. for all these reasons, machine scores predict future academic success abysmally (Mattern & Packman, 2009; Matzen & Hoyt, 2004; Ramineni & Williamson, 2013)
>
> And that machine scoring does not measure, and therefore does not promote, authentic acts of writing:
>
> 1. students are subjected to a high-stakes response to their writing by a device that, in fact, cannot read, as even testing firms admit (Elliott, 2011)
> 2. in machine-scored testing, often students falsely assume that their writing samples will be read by humans with a human's insightful understanding (Herrington & Moran, 2006)
> 3. conversely, students who knowingly write for a machine are placed in a bind since they cannot know what qualities of writing the machine will react to positively or negatively, the specific algorithms being closely guarded secrets of the testing firms (Frank, 1992; Rubin & O'Looney, 1990)—a bind made worse when their essay will be rated by both a human and a machine

4. students who know that they are writing only for a machine may be tempted to turn their writing into a game, trying to fool the machine into producing a higher score, which is easily done (McGee, 2006; Powers, Burstein, Chodorow, Fowles, & Kukich, 2001; see item 6, above)
5. teachers are coerced into teaching the writing traits that they know the machine will count–surface traits such as essay length, sentence length, trivial grammatical mistakes, mechanics, and topic-related vocabulary—and into *not* teaching the major traits of successful writing—elements such as accuracy, reasoning, organization, critical and creative thinking, and engagement with current knowledge (Council, 2012; Deane, 2013; Herrington & Moran, 2001; National, 2010)
6. machines also cannot measure authentic audience awareness, a skill essential at all stages of the composing process and correlative with writing competence of students both in the schools (Wolmann-Bonilla, 2000) and in college (Rafoth, 1985)
7. as a result, the machine grading of high-stakes writing assessments seriously degrades instruction in writing (Perelman, 2012a), since teachers have strong incentives to train students in the writing of long verbose prose, the memorization of lists of lengthy and rarely used words, the fabrication rather than the researching of supporting information, in short, to dumb down student writing.

The Human Readers petition can be found in its entirety at
http://humanreaders.org/petition/works_cited.htm

[ii] The following excerpt on best practices for assessment systems is from the NCTE position on machine scoring:

> What Are the Alternatives [to Machine Scoring]?
> Together with other professional organizations, the National Council of Teachers of English has established research-based guidelines for effective teaching and assessment of writing, such as the *Standards for the Assessment of Reading and Writing* (rev. ed., 2009), the *Framework for Success in Postsecondary Writing* (2011), the *NCTE Beliefs about the Teaching of Writing* (2004), and the *Framework for 21st Century Curriculum and Assessment* (2008, 2013). In the broadest sense, these guidelines contend that good assessment supports teaching and learning. Specifically, high-quality assessment practices will
>
> - 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;
> - yield high-quality, useful information to inform teachers about curriculum, instruction, and the assessment process itself;
> - balance the need to assess summatively (make final judgments about the quality of student work) with the need to assess formatively (engage in ongoing, in-process judgments about what students know and can do, and what to teach next);
> - recognize the complexity of literacy in today's society and reflect that richness through holistic, authentic, and varied writing instruction;
> - at their core, involve professionals who are experienced in teaching writing, knowledgeable about students' literacy development, and familiar with current research in literacy education.
>
> A number of effective practices enact these research-based principles, including portfolio assessment; teacher assessment teams; balanced assessment plans that involve more localized (classroom- and district-based) assessments designed and administered by classroom teachers; and "audit" teams of teachers, teacher educators, and writing specialists who visit districts to review samples of student work and the curriculum that

has yielded them. We focus briefly here on portfolios because of the extensive scholarship that supports them and the positive experience that many educators, schools, and school districts have had with them.

Engaging teams of teachers in evaluating portfolios at the building, district, or state level has the potential to honor the challenging expectations of the CCSS while also reflecting what we know about effective assessment practices. Portfolios offer the opportunity to

- look at student writing across multiple events, capturing growth over time while avoiding the limitations of "one test on one day";
- look at the range of writing across a group of students while preserving the individual character of each student's writing;
- review student writing through multiple lenses, including content accuracy and use of resources;
- assess student writing in the context of local values and goals as well as national standards.

Just as portfolios provide multiple types of data for assessment, they also allow students to *learn* as a result of engaging in the assessment process, something seldom associated with more traditional one-time assessments. Students gain insight about their own writing, about ways to identify and describe its growth, and about how others -- human readers -- interpret their work. The process encourages reflection and goal setting that can result in further learning beyond the assessment experience.

Similarly, teachers grow as a result of administering and scoring the portfolio assessments, something seldom associated with more traditional one-time assessments. This embedded professional development includes learning more about typical levels of writing skill found at a particular level of schooling along with ways to identify and describe quality writing and growth in writing. The discussions about collections of writing samples and criteria for assessing the writing contribute to a shared investment among all participating teachers in the writing growth of all students. Further, when the portfolios include a wide range of artifacts from learning and writing experiences, teachers assessing the portfolios learn new ideas for classroom instruction as well as ways to design more sophisticated methods of assessing student work on a daily basis.

Several states such as Kentucky, Nebraska, Vermont, and California have experimented with the development of large-scale portfolio assessment projects that make use of teams of teachers working collaboratively to assess samples of student work. Rather than investing heavily in assessment plans that cannot meet the goals of the CCSS, various legislative groups, private companies, and educational institutions could direct those funds into refining these nascent portfolio assessment systems. This investment would also support teacher professional development and enhance the quality of instruction in classrooms -- something that machine-scored writing prompts cannot offer.

The "NCTE Position on Machine Scoring" can be found in its entirety at http://www.ncte.org/positions/statements/machine_scoring