Machine Scoring Fails the Test

> [A] computer could not measure accuracy, reasoning, adequacy of evidence, good sense, ethical stance, convincing argument, meaningful organization, clarity, and veracity in your essay. If this is true I don't believe a computer would be able to measure my full capabilities and grade me fairly.—Akash, student
>
> [H]ow can the feedback a computer gives match the carefully considered comments a teacher leaves in the margins or at the end of your paper?—Pinar, student
>
> (Responses to *New York Times* The Learning Network blog post, "How Would You Feel about a Computer Grading Your Essays?", 5 April 2013)

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.

As the Common Core State Standards (CCSS) sweep into individual classrooms, they bring with them a renewed sense of the importance of writing to students' education. Writing teachers have found many aspects of the CCSS to applaud; however, we must be diligent in developing assessment systems that do not threaten the possibilities for the rich, multifaceted approach to writing instruction advocated in the CCSS. 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.

Research[1] on the assessment of student writing consistently shows that high-stakes writing tests alter the normal conditions of writing by denying students the opportunity to think, read, talk with others, address real audiences, develop ideas, and revise their emerging texts over time. Often, the results of such tests can affect the livelihoods of teachers, the fate of schools, or the educational opportunities for students. In such conditions, the narrowly conceived, artificial form of the tests begins to subvert attention to other purposes and varieties of writing development in the classroom. Eventually, the tests erode the foundations of excellence in writing instruction, resulting in students who are less prepared to meet the demands of their continued education and future occupations. Especially in the transition from high school to college, students are ill served when their writing experience has been dictated by tests that ignore the ever-more complex and varied types and uses of writing found in higher education.

These concerns—increasingly voiced by parents, teachers, school administrators, students, and members of the general public—are intensified by the use of machine-scoring systems to read and evaluate students' writing. 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.

[1]All references to research are supported by the extensive work documented in the annotated bibliography attached to this report. The bibliography is drawn from a body of independent and industry research that supports other critiques of machine scoring, such as the Professionals Against Machine Scoring Of Student Essays In High-Stakes Assessment Petition Initiative.

Herrington, Anne & Moran, Charles. (2001). What happens when machines read our students' writing? *College English, 63*(4), 480-499.*

Provides a short history of the field of composition's response to machine scoring and examines two programs now heavily marketed nationwide: *Intellimetric*, the platform of WritePlacer Plus, and *Intelligent Essay Assessor*. Herrington and Moran each submit work to both scoring programs and discuss the different outcomes. Argues that machine scoring does not treat writing as a rhetorical interaction between writers and readers. Calls into question the efficiency and reliability claims companies make as the primary basis for marketing their programs. Argues that machine scoring may send the message to students that human readings are unreliable, irrelevant, and replaceable, and that the surface features of language matter more than the content and the interactions between reader and text—a message that sabotages compositions' pedagogical goals.

Powers, Donald E., Burstein, Jill C., Chodorow, Martin, Fowles, Mary E. & Kukich, Karen. (2001). Stumping *e-rater*: Challenging the validity of automated essay scoring (GRE Report, No. 98-08bP). www.ets.org/Media/Research/pdf/RR-01-03-Powers.pdf *

Reports on a study in which writing specialists, linguists, language testing experts, and computer software experts were encouraged to write and submit essays they believed would trick *e-rater* into giving higher or lower scores than the essays deserved. Human readers scored the essays, as did *e-rater.* Study found that readers agreed with one another within one point of the scoring scale 92% of the time, while *e-rater* and readers agreed within one point of each other 65% of the time. Further, *e-rater* was more likely to give inflated scores than to give lower than warranted scores. Some of the essays given the highest score (6) by *e-rater* but very low scores by human readers were those that repeated whole paragraphs or that used key phrases from the question but that merely agreed with the writing prompt instead of analyzing it, as directed. Essays earning lower than warranted scores were those that included subtle transitions between ideas or frequent literary allusions. Concludes that *e-rater* should not be used without human scorers and that more could be done to train human scorers in the aspects of writing that *e-rater* overlooks. This is a technical report by ETS, so not a peer-reviewed publication, but it offers useful insight into the AES.

Jones, Brett D. (1999). Computer-rated essays in the English composition classroom. *Journal of Educational Computing Research, 20*(2), 169-186.*

Reports on a study designed to determine how middle and high school teachers would use computer-generated ratings of student writing if they were available. Discusses the potential for computer-generated rated essays to help teachers give feedback to student essays. Reviews the types of feedback students find most helpful, suggests that teachers do not have enough time to provide this type of feedback, and argues that *Project Essay Grade (PEG)* is capable of rating the overall quality of an essay, thus leaving more time for teachers to provide more specific and content-based feedback on student papers. Stresses that *PEG* ratings do not give information on why an area of writing is weak (for instance, content, organization, style, mechanics, creativity), but alerts teachers to areas that need attention.

Whittington, Dave & Hunt, Helen. Approaches to the computerized assessment of free text responses. (1999). *Proceedings of the Third Annual Computer Assisted Assessment Conference* (pp. 207-219). Loughborough, England: Loughborough University.*

Provides clear, brief descriptions of how a number of machine scoring software programs operate, including *Project Essay Grade* (*PEG*), *Latent Semantic Analysis* (*LSA*), Microsoft's *Natural Language Processing Tool,* and Educational Testing Service's *e-rater.* Also describes two other, potentially beneficial, software initiatives: Panlingua, which is based on the assumption that there is a universal language that reflects understanding and knowledge and on several levels would map onto a software program the way the brain understands language/ideas, and *Lexical Conceptual Structure* (*LCS*), which is based on the idea that a machine "must be capable of capturing language-independent information—such as meaning, and relationships between subjects and objects in sentences—whilst still processing many types of language-specific details, such as syntax and divergence" (p. 10). Points out that there are many important limitations of all of these software initiatives but that they hold promise and, together, represent the dominant ways of thinking about how to build software to address the scoring of complex writing tasks.

Breland, Hunter M. (1996). Computer-assisted writing assessment: The politics of science versus the humanities. In Edward M. White, William D. Lutz & Sandra Kamusikiri (Eds.) *Assessment of Writing: Politics, Policies, Practices* (pp. 249-256). New York: Modern Language Association.*

Briefly reviews the development of computer-based evaluation of writing by "scientists" and the resistance to this approach by those in the "humanities." Addresses programs such as Bell Labs *Writer's Workbench* as well as author's own research into Educational Testing Service's *WordMAP* program. Concludes that although many writing teachers still oppose the focus on error and mechanics that characterize the computer-based approach, a "certain amount of standardization, particularly in writing mechanics, is an essential part of writing and writing assessment," and to deny this fact "is not good for writing instruction" (p. 256).

Huot, Brian A. (1996). Computers and assessment: Understanding two technologies. *Computers and Composition, 13*(2), 231-243.*

Examines the problems and possibilities of using assessment technologies, and argues that we must base decisions for using any technology on sound theory and research. Includes a literature review on computer scoring. Considers theoretical assumptions of assessment practices and computer practices with respect to teaching and communicating, paying special attention to the debate about computers as value-free versus value-laden tools. Examines validity and reliability arguments of machine scoring and the theoretical implications of using computers for assessment of and response to student writing.

Brock, Mark N. (1995). Computerized text analysis: Roots and research. *Computer Assisted Language Learning 8*(2-3), 227-258.*

Focuses on computerized text analysis programs, such as *Writer's Workbench, Edit,* and *Critique,* that provide feedback to writers to prompt revision. Explains the way these programs function, summarizes how they were developed, and reviews research about their efficacy. Identifies the "exclusive focus on surface-level features of a text" as the "most severe limitation" of computerized text analysis because it directs students away from meaning making (p. 236). Concludes that the beneficial claims about these programs as writing aids are "at best controversial and at worst simply untrue" (p. 254). Describes how the programs are used to give feedback to writers and contrasts this use with how the programs grade writing.

Prepared by the NCTE Task Force on Writing Assessment

Chris Anson, North Carolina State University (chair)

Scott Filkins, Champaign Unit 4 School District, Illinois

Troy Hicks, Central Michigan University

Peggy O'Neill, Loyola University Maryland

Kathryn Mitchell Pierce, Clayton School District, Missouri

Maisha Winn, University of Wisconsin

NCTE
National Council of Teachers of English
1111 W. Kenyon Road, Urbana, IL 61801-1096

NCTE Position Statement

A statement on an education issue approved by the NCTE Board of Directors

NCTE Position Statement on Machine Scoring

Approved by the NCTE Executive Committee

April 15, 2013

NCTE
National Council of Teachers of English

Shermis, Mark D., & Burstein, Jill (Eds.). (2003). *Automated essay scoring: A cross-disciplinary perspective*. Mahwah, NJ: Lawrence Erlbaum. *

Thirteen original essay-chapters on the development of computer programs to analyze and score "free" or essay-like pieces of discourse. The bulk of the book documents and promotes current computerized methods of text analysis, scoring software, or methods to validate them: Ellis Batten Page on *Project Essay Grade (PEG*); Scott Elliot on *IntelliMetric*; Thomas K. Landauer, Darrell Laham, & Peter W. Foltz on *Intelligent Essay Assessor*; Jill Burstein on *e-rater*; Leah S. Larkey & W. Bruce Croft on binary classifiers as a statistical method for text analysis; Gregory J. Cizek & Bethany A. Page on statistical methods to calculate human-machine rater reliability and consistency; Timothy Z. Keith on studies validating several programs by correlating human rates and machines rates; Mark D. Shermis & Kathryn E. Daniels on use of scales and rubrics when comparing human and machine scores; Claudia Leacock & Martin Chodorow on the accuracy of an error-detection program called *ALEK (Assessment of Lexical Knowledge);* Jill Burstein & Daniel Marcu on the accuracy of a computer algorithm in identifying a "thesis statement" in an open essay. Although chapters are highly informative—data-based and well documented--conspicuously absent are studies of the use and impact of machine scoring or feedback in actual classrooms. The introduction argues that "Writing teachers are critical to the development of the technology because they inform us as to how automated essay evaluations can be most beneficial to students" (xv), but no new information along those lines is presented.

Williamson, Michael M. (2003). Validity of automated scoring: Prologue for a continuing discussion of machine scoring student writing. *Journal of Writing Assessment, 1*(2), 85-104.*

Reviews the history of writing-assessment theory and research, with particular attention to evolving definitions of validity. Argues that researchers and theorists in English studies should read and understand the discourse of the educational measurement community. When theorists and researchers critique automated scoring, they must consider the audiences they address, that they must understand the discourse of the measurement community rather than write only in terms of English Studies theory. Argues that while common ground exists between the two communities, writing teachers need to acknowledge the complex nature of validity theory and consider both the possibilities and problems of automated scoring rather than focus exclusively on what they may see as threatening in this newer technology. Points out that there is a divide in the way writing assessment is discussed among professionals, with the American Psychological Association and the American Educational Research Association discussing assessment in a decidedly technical fashion and the National Council of Teachers of English and Conference on College Composition and Communication groups discussing writing assessment as one aspect of teaching and learning about assessment. Williamson points out that the APA and AERA memberships are much larger than those of NCTE and CCCC, and that writing studies professionals would do well to learn more about the assessment discussions happening in APA and AERA circles.

Powers, Donald E., Burstein, Jill, Chodorow, Martin S., Fowles, Mary E. & Kukich, Karen. (2002). Comparing the validity of automated and human scoring of essays. *Journal of Educational Computing Research 26*(4), 407-425.*

The authors compared *e-rater* scores with students' self-reports of writing ability, writing accomplishment, grades in writing-intensive courses, and other "non-test" variables, and found that expert human ratings of essays correlated better than did *e-rater* ratings, although both were low. They conclude that *e-rater* scores are "less valid than are those assigned by trained readers" (p. 421), but only assuming that the "non-test" variables are valid measures of writing skill.

Shermis, Mark D. & Barrera, Felicia. (2002). Automated essay scoring for electronic portfolios. *Assessment Update, 14*(4), 1-11.*

Provides an update on a grant from the Fund for the Improvement of Postsecondary Education (FIPSE) that explores the use of automated essay scoring (AES) for electronic portfolios. Argues that large numbers of e-portfolios necessitate the use of AES evaluative systems. Presents data showing the validity of three AES systems: *Project Essay Grade (PEG), IntelliMetric,* and *Intelligent Essay Assessor (IEA).* Reports that project researchers were creating national norms for documents; norms will be available through automated software on-line for a period of five years.

Shermis, Mark D., Mzumara, Howard R., Olson, Jennifer & Harrington, Susanmarie. (2001). On-line grading of student essays: PEG goes on the world wide web. *Assessment and Evaluation in Higher Education 26*(3), 247-260.*

Describes two studies in using *Project Essay Grade (PEG)* software for placement of students into college-level writing courses. In the first study, students' papers were used to create a scoring schemata for the software; in the second, scores provided by *PEG* and human readers were compared. Argues that *PEG* works because the computer scores and raters' scores had high correlations; in addition, *PEG* is an efficient and low-cost way to do low-stakes writing assessment like placement. Although the authors note that a good writer could fool the system by submitting a nonsensical essay, the article does not address other potential problems with machine scoring of student essays. In fact, it ends by pointing out how *PEG's* use could be expanded beyond placement assessment into the grading of essays in programs like Write 2000, which promotes more writing in grades 6-12.

- 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.

What Are the Alternatives?

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.

What's Next

In 2010, the federal government awarded $330 million to two consortia of states "to provide ongoing feedback to teachers during the course of the school year, measure annual school growth, and move beyond narrowly focused bubble tests" (United States Department of Education). Further, these assessments will need to align to the new standards for learning in English and mathematics. This has proven to be a formidable task, but it is achievable. By combining the already existing National Assessment of Educational Progress (NAEP) assessment structures for evaluating school system performance with ongoing portfolio assessment of student learning by educators, we can cost-effectively assess writing without relying on flawed machine-scoring methods. By doing so, we can simultaneously deepen student and educator learning while promoting grass-roots innovation at the classroom level. For a fraction of the cost in time and money of building a new generation of machine assessments, we can invest in rigorous assessment and teaching processes that enrich, rather than interrupt, high-quality instruction. Our students and their families deserve it, the research base supports it, and literacy educators and administrators will welcome it.

Works Cited

United States Department of Education. "U.S. Secretary of Education Duncan Announces Winners of Competition to Improve Student Assessments." (2 Sept 2010; retrieved 11 April 2013). http://www.ed.gov/news/press-releases/us-secretary-education-duncan-announces-winners-competition-improve-student-asse

Annotated Bibliography

The following annotated bibliography on machine scoring and evaluation of essay-length writing is based on the 2012 published bibliography in the *Journal of Writing Assessment* 5 (compiled by Richard Haswell, Whitney Donnelly, Vicki Hester, Peggy O'Neill, and Ellen Schendel), and available at: http://www.journalofwritingassessment.org/article.php?article=58.

The bibliography was compiled by reviewing recent scholarship on machine scoring of essays, also referred to as automated essay scoring (AES), using databases such as ERIC and CompPile. Entries were selected for their attention to machine scoring of essays and publication in peer-reviewed venues (with exceptions noted). We also endeavored to cover the breadth of the issues addressed in the research without being overly redundant. We avoided publications that were very narrowly focused on highly technical aspects of assessment. The earliest research—such as Ellis Page's 1966 piece in *Phi Delta Kappan*, "The Imminence of Essay Grading by Computer"—is not included because many more recent entries provide a review of the early development of machine scoring.

The bibliography is organized by publication date, with the most recent entries appearing first. Entries that have been excerpted from the published *JWA* bibliography are indicated by an asterisk.

Klobucar, Andrew, Deane, Paul, Elliot, Norbert, Raminie, Chaitanya, Deess, Perry & Rudniy, Alex. (2012). Automated essay scoring and the search for valid writing assessment. In Charles Bazerman et al. (Eds.) *International Advances in Writing Research: Cultures, Places, Measures* (pp. 103-119). Fort Collins, CO: WAC Clearinghouse & Parlor Press.

This chapter reports on an ETS and New Jersey Institute of Technology research collaboration that used Criterion, an integrated instruction and assessment system that includes automated essay scoring. The purpose of the research was "to explore ways in which automated essay scoring

Sandene, Brent, Horkay, Nancy, Bennet, Randy Elliot, Allen, Nancy, Braswell, James, Kaplan, Bruce, & Oranje, Andreas. (2005). Part II: Online writing assessment. *Online assessment in mathematics and writing: Reports From the NAEP Technology-Based Assessment Project, Research and Development Series*. NCES 2005–457). U.S. Department of Education, National Center for Education Statistics. Washington, DC: U.S. Government Printing Office.

While not a traditional peer-reviewed publication, the NAEP research report is considered a high quality scholarly source; it describes the results of the 2002 Writing Online study of a national sample of eighth-graders writing online and compared the results of those students taking the traditional pencil-and-paper format of the test. The report is a comprehensive comparison, which includes the machine scoring of essays using erater 2.0, with one subsection on the AES (pp. 37-44). Results of the study "showed that the automated scoring of essay responses did not agree with the scores awarded by human readers." Moreover, AES "produced mean scores that were significantly higher" than those awarded by human readers and that the human readers "agreed with each other" at higher rates than the agreement between the AES scores and those produced by the human readers. In rank ordering essay, again human readers and AES did not agree at the same rates as human readers did with each other.

Penrod, Diane. (2005). *Composition in convergence: The impact of new media on writing assessment*. Mahwah, New Jersey: Lawrence Erlbaum.*

Argues that since writing and writing assessment are intertwined, and since writing and writing standards are rapidly changing under the impact of digital technology, machine scoring cannot keep up: "The current push for traditional assessment standards melding with computer technology in forms like the Intelligent Essay Assessor, E-rater, and other software programs provides a false sense of establishing objective standards that appear to be endlessly repeated across time and space" (p. 164).

Shermis, Mark D., Burstein, Jill, & Leacock, Claudia. (2005). Applications of computers in assessment and analysis of writing. In Charles A. MacArthur, Steve Graham, & Jill Fitzgerald (Eds.), *Handbook of writing research* (pp. 403-416). New York: Guilford Press.*

A review of what the authors call "automated essay scoring" (AES) from the perspective of the testing industry. There is a brief history of the development of the most successful software, a very informed discussion of reliability and validity studies of AES (although validity is restricted to correlations with other assessments of student essays), a useful explanation of the different approaches of Ellis Page's *Project Essay Grade* (PEG), ETS's *e-rater*, Vantage's *IntelliMetric*, and Thomas Landauer and Peter Foltz's *Intelligent Essay Assessor*, and a shorter discussion of computerized critical feedback programs such as *Criterion* and *c-rater.* The authors conclude that teachers need to understand how the technology works, since "the future of AES is guaranteed, in part, by the increased emphasis on testing for U. S. schoolchildren" (p. 414).

Whithaus, Carl. (2005). *Teaching and evaluating writing in the age of computers and high-stakes testing*. Mahwah, NJ: Lawrence Erlbaum.*

The larger argument of this book is that digital technology changes everything about the way writing is or should be taught. That includes evaluating writing. Whithaus critiques high-stakes writing assessment as encouraging students to "shape whatever material is placed in front of [them] into a predetermined form" (p. 11) rather than encouraging thinking through how to communicate to different audiences for different purposes and through different modalities. He argues that if the task is to reproduce known facts, then systems such as *Project Essay Grade* (*PEG*) or *Intelligent Essay Assessor* (*IEA*) may be appropriate; but if the task is to present something new, then the construction of electronic portfolios makes a better match. Suggests that using e-portfolios creates strong links between teaching and assessment in an era when students are being taught to use multimodal forms of communication. Argues that scoring packages such as *e-Write* or *e-rater*, and the algorithms that drive them such as latent semantic analysis or multiple regression on countable traits may serve to evaluate reproducible knowledge or "dead" text formats such as the 5-paragraph essay (p. 121), but cannot fairly assess qualities inherent in multimedia and multimodal writing of blogs, instant messaging, or e-portfolios, where the production is epistemic and contextual and where the evaluation should be situated and distributed (judged by multiple readers). Making this book particularly useful is its extended analysis of contemporary student texts.

Cheville, Julie. (2004). Automated scoring technologies and the rising influence of error. *English Journal 93*(4), 47-52.*

Examines the theoretical foundations and practical consequences of *Criterion*, the automated scoring program that the Educational Testing Service is still developing. Bases her critique on information provided by ETS as part of an invitation to participate in a pilot study. Contrasts the computational linguistic framework of *Criterion* with a position rooted in the social construction of language and language development. Links the development of the program with the high-stakes large-scale assessment movement and the "power of private interests to threaten fundamental beliefs and practices underlying process instruction" so that the real problem—"troubled structures of schooling" (p. 51)—will remain.

Burstein, Jill, & Marcus, Daniel. (2003). A machine learning approach for identification of thesis and conclusion statements in student essays. *Computers and the Humanities 37*, 455-467.*

Explains how a machine may be able to evaluate a criterion of good writing (organization) that many teachers think cannot be empirically measured. Argues that essay-based discourse-analysis systems can reliably identify thesis and conclusion statements in student writing. Explores how systems generalize across genre and grade level and to previously unseen responses on which the system has not been trained. Concludes that research should continue in this vein because a machine-learning approach to identifying thesis and conclusion statements outperforms a positional baseline algorithm.

Neal, Michael R. (2011). *Writing assessment and the revolution in digital texts and technologies*. New York: Teachers College Press.*

After a thorough review of the response to machine scoring from composition scholars (pp. 65-74), argues that the mechanization represented by machine scoring is a "misdirection" in which we, the teaching community, are partly complicit: "somewhere along the way we have lost the idea of how and why people read and write within meaningful rhetorical situations," noting, however, that machine scoring is "a cheap, mechanized solution to a problem that we have not had opportunity to help define" (p. 74).

Elliott, Scott. (2011). Computer-graded essays full of flaws. *Dayton Daily News* (May 24). http://www.daytondailynews.com/project/content/project/tests/0524testautoscore.html*

Describes how the reporter tested Educational Testing Service's *e-rater* by submitting two essays, one his best effort and one designed to meet the computer program's preference for "long paragraphs, transitional words, and a vocabulary a bureaucrat would love" but also filled with such nonsense as "king Richard Simmons, a shoe-eating television interloper, alien beings and green swamp toads." *E-rater* gave the first essay a score of 5 (on a scale of 1 up to 6) and the nonsense essay a score of 6. An English teacher gave the first essay 6+ and the second 1 on the same scale. Rich Swartz of Educational Testing Service explained that "we're a long way from computers actually reading stuff." This isn't a peer reviewed publication but provides a useful perspective on the limitations of AES.

Dikli, Semire. (2010). The Nature of Automated Essay Scoring Feedback. *CALICO Journal* 28(1), 99-134.

A study of the feedback on their writing received twelve adult English Language Learners from My Access! an Automated Essay Scoring (AES) program that uses the Intellimetric system and teacher's feedback. The program was not scoring essays but providing students with feedback. The study used case study methodology including observation, interviews with the students, and examination of the texts. Students were divided into two groups: one group of six received feedback from the computer system and one the teacher. The feedback from AES and the teacher differed extensively in terms of length, usability, redundancy, and consistency. The researcher reported that My Access provided substantially more feedback than the teacher but that it wasn't as useable, it was highly redundant, generic, but consistent. The AES system did not use positive reinforcement and didn't connect on a personal level to the student. The researcher concluded that the AES program did not meet the needs of nonnative speakers.

Byrne, Roxanne, Tang, Michael, Truduc, John & Tang, Matthew. (2010). eGrader, a software application that automatically scores student essays: with a postscript on the ethical complexities. *Journal of Systemics, Cybernetics & Informatics 8* (6), 30-35.*

Provides a very brief overview of three commercially available automatic essay scoring services (Project Essay Grade, Intellimetric, and eRater) as well as eGrader. eGrader differs from others because it operates on a client PC; requires little human training; is cost effective; and does not require a huge database. While it shares some processes as these other AES applications, differences include key word searching of web pages for benchmark data. Authors used 33 essays to compare the eGrader results with human judges. Correlations between the scores were comparable with other AES applications. In classroom use, however, the instructor "found a disturbing pattern": "The machine algorithm could not detect ideas that were not contained in the readings or Web benchmark documents although the ideas expressed were germane to the essay question." Ultimately, the authors decided not to use machine readers because they "could not detect other subtleties of writing such as irony, metaphor, puns, connotation and other rhetorical devices" and "appears to penalize those students we want to nurture, those who think and write in original or different ways."

Crusan, Deborah. (2010). *Assessment in the second language classroom*. Ann Arbor, MI: University of Michigan Press.*

With an interest in second-language instruction, the author tested out Pearson Educational's *Intelligent Essay Assessor* and found the diagnosis "vague and unhelpful" (p. 165). For instance, *IEA* said that the introduction was "missing, undeveloped, or predictable. Which was it?" (p. 166). Her chapter on machine scoring (pp. 156-179 compares all the major writing-analysis software, with an especially intense look at Vantage Learning's *MY Access!* (based on *IntelliMetric*), and finds the feedback problematical, in part because it can be wrongly used by administrators and it can lead to "de-skilling" of teachers (p. 170). Cautions that the programs, "if used at all, ought to be used with care and constant teacher supervision and intervention" (p. 178).

McCurry, Doug. (2010). Can machine scoring deal with broad and open writing tests as well as human readers? *Assessing Writing 15*(2), 118-129.*

Investigates the claim that machine scoring of essays agrees with human scorers. Argues that the research supporting this claim is based on limited, constrained writing tasks such as those used for the GMAT, but a 2005 study reported by NAEP shows automated essay scoring (AES) is not reliable for more open tasks. McCurry reports on a study that compares the results of two machine scoring applications to the results of human readers for the writing portion of the Australian Scaling Test (AST), which has been designed specifically to encourage test takers to identify an issue and drafting and revising to present a point of view. It does not prescribe a form or genre, or even the issue. It has been designed to reflect classroom practice, not to facilitate grading and inter-rater agreement, according to McCurry. Scoring procedures, which are also different than those typically used in large-scale testing in the USA, involve four readers scoring essays on a 10-point scale. After comparing and analyzing the results between the human scores and the scores given by the AES applications, McCurry concludes that machine scoring cannot score open, broad writing tasks more reliably than human readers.

Herrington, Anne, & Moran, Charles. (2009). Writing, assessment, and new technologies. In Marie C. Paretti & Katrina Powell (Eds.), *Assessment in writing* (Assessment in the disciplines, Vol. 4) (pp. 159-177). Tallahassee, TN: Association of Institutional Researchers.*

Herrington and Moran argue against educators and assessors "relying principally or exclusively on standardized assessment programs or using automated, externally developed writing assessment programs" (p. 177). They submitted an essay written by Moran to Educational Testing Service's *Criterion*, and found that the program was "vague, generally misleading, and often dead wrong" (p. 163). For instance, of the eight problems *Criterion* found in grammar, usage, and mechanics, all eight were false flags. The authors also critique Edward Brent's *SAGrader*, finding the software's analysis of free responses written for content courses generally helpful if used in pedagogically sound ways; but they severely question Collegiate Learning Assessment's ability to identify meaningful learning outcomes, especially now that CLA has resorted to Educational Testing Service's *e-rater* to score essays composed for CLA's "more reductive" task-based prompts (p. 171).

Scharber, Cassandra, Dexter, Sara, & Riedel, Eric. (2008). Students' experiences with an automated essay scorer. *Journal of Technology, Learning and Assessment* 7(1). Retrieved 4/1/2013 from http://www.jtla.org.

The study explored preservice English teachers' experience with automated essay scoring (AES) for formative feedback in an online, cased-based course. Data collected included post-assignment surveys, a user log of students' actions within the cases, instructor-assigned scores on final essays, and interviews with four selected students. The course used ETIPS, a comprehensive, online system that includes an AES option for formative feedback. The cases "are multimedia, network-based, online instructional resources that provide learning opportunities. . . to practice instructional decision-making skills related to technology integration and implementation" (p. 6). Twenty-five of the thirty-four students agreed to participate in the study, and thirteen of the twenty-five agreed to be interviewed with four being selected through a purposive sampling matrix. Survey results showed that "most students did not assign strong positive ratings to any aspect of the scorer" but did find the AES helpful in composing their own response yet they did not have much confidence in its evaluation. In response to an open-ended question about their use of the AES, the authors reported the two most frequent responses from the students were "that they tried to 'please and then beat the scorer'" (n=16) and they "used the scorer then gave up" (n=9) (p. 14). From the four case studies, the authors concluded that "the nature of the formative feedback given to these students by the ETIPS scorer was not sophisticated enough for them to know what specific sort of revision to make to their answers" (p. 28).

Shermis, Mark D., Shneyderman, Aleksandr, Attali, Yigal. (2008). How important is content in the ratings of essay assessments? *Assessment in Education: Principles, Policy & Practice,* 15 (1), 91-105.

EBSCO ABSTRACT: This study was designed to examine the extent to which "content" accounts for variance in scores assigned in automated essay scoring protocols. Specifically it was hypothesised that certain writing genre would emphasise content more than others. Data were drawn from 1668 essays calibrated at two grade levels (6 and 8) using "e-rater[TM]", an automated essay scoring engine with established validity and reliability. "E-rater" v 2.0's scoring algorithm divides 12 variables into "content" (scores assigned to essays with similar vocabulary; similarity of vocabulary to essays with the highest scores) and "non-content" (grammar, usage, mechanics, style, and discourse structure) related components. The essays were classified by genre: persuasive, expository, and descriptive. The analysis showed that there were significant main effects due to grade, $F(1,1653) = 58.71$, p less than 0.001, and genre $F(2, 1653) = 20.57$, p less than 0.001. The interaction of grade and genre was not significant. Eighth grade students had significantly higher mean scores than sixth grade students and descriptive essays were rated significantly higher than those classified as persuasive or expository. Prompts elicited "content" according to expectations with lowest proportion of content variance in persuasive essays, followed by expository and then descriptive. Content accounted for approximately 0-6% of the overall variance when all predictor variables were used. It accounted for approximately 35-58% of the overall variance when "content" variables alone were used in the prediction equation. (Contains 9 tables, 2 figures and 2 notes.)

Chen, Chi-Fen Emily, & Cheng, Wei-Yuan Eugene. (2008). Beyond the design of automated writing evaluation: Pedagogical practices and perceived learning effectiveness in EFL writing classes. *Language Learning and Technology 12*(2), 94-112.*

Used naturalistic classroom investigation to see how effectively *MY Access!* worked for ESL students in Taiwan. Found that the computer feedback was most useful during drafting and revising but only when it was followed with human feedback from peer students and from teachers. When students tried to use *MY Access!* on their own, they were often frustrated and their learning was limited. Generally, both teachers and students perceived the software and its feedback negatively.

Wang, Jinhao, & Brown, Michelle Stallone. (2008). Automated essay scoring versus human scoring: A correlational study. *Contemporary Issues in Technology and Teacher Education 8*(4) (http://www.citejournal.org/vol8/iss4/languagearts/article1.cfm)*

In one of the few empirical comparisons of machine with human scoring conducted outside the testing companies themselves, Wang and Brown had trained human raters independently score student essays that had been scored by *IntelliMetric* in *WritePlace Plus*. The students were enrolled in an advanced basic-writing course in a Hispanic-serving college in south Texas. On the global or holistic level, the correlation between human and machine scores was only .11. On the five dimensions of focus, development, organization, mechanics, and sentence structure, the correlations ranged from .06 to .21. These dismal machine-human correlations question the generalizability of industry findings, which, as Wang and Brown point out, emerge from the same population of writers on which both machines and raters are trained. *IntelliMetric* scores also had no correlation (.01) with scores that students later achieved on the human-scored essay in a state-mandated exam, whereas the two human ratings correlated significantly (.35).

Wohlpart, James, Lindsey, Chuck, & Rademacher, Craig. (2008). The reliability of computer software to score essays: Innovations in a humanities course. *Computers and Composition 25*(2), 203-223.*

Considers Florida Gulf Coast University's general-education course Understanding the Visual and Performing Arts, taught online in two large sections. Used *Intelligent Essay Assessor* to score two short essays that were part of module examinations. On four readings, using a four-point holistic scale, faculty readers achieved exact agreement with two independent readers only 49, 61, 49, and 57 percent of the time. *IEA*'s scores correlated with the final human scores (achieved sometimes after four readings) 64% of the time. When faculty later re-read discrepant essays, their scores almost always moved toward the *IEA* score. With essays where there was still a discrepancy, 78% were scored higher by *IEA*. The faculty were "convinced" that the use of *IEA* was a "success." Note that the authors do not investigate the part that statistical regression toward the mean might have played in these results.

Hutchison, Dougal. (2007). An evaluation of computerised essay marking for national curriculum assessment in the UK for 11-year-olds. *British Journal of Educational Technology* 38 (6), 977-989.

This study examines "how well the computer program can replicate human marking" (p. 980) and the "discrepancies between computer and human marking, and to try to identify the reasons for these" (p. 981). It used erater and a subset of 600 essays collected as part of the National Foundation for Educational Research's work in developing National Curriculum Assessments in English. The comparison of erater scores with human readers showed that "e-rater scores agree nearly as often with human readers as two human readers agree with each other, and more closely with the average of the readers" (p. 981). To determine the reason for the discrepancies, the markers discussed the texts that had received discrepant scores and the researcher identified three reasons for the discrepancies that he termed Human Friendly, Neutral and Computer Friendly. Based on the analysis of the results of the studies, the author concluded that "that even the most sophisticated programs, such as e-rater , which bases its assessment on a number of dimensions, can still miss out on important intrinsic qualities of an essay, such as whether it was lively or pedestrian" (988).

James, Cindy L. (2007). Validating a computerized scoring system for assessing writing and placing students in composition courses. *Assessing Writing 11*(3), 167-178.*

Compares scores given by *ACCUPLACER OnLine WritePlacer Plus* (using *IntelliMetric*) with student essay scores given by "untrained" faculty at Thompson Rivers University, and then compares the success of these two sets in predicting pass or failure in an introductory writing course and a course in literature and composition. ACCUPLACER was administered during the first week of class. Correlations between machine and human scores (ranging from .40 to .61) were lower than those between humans (from .45 to .80). Neither machine nor human scores accounted much for the variation in the composition or literature courses success (machine: 16% and 5%; humans: 26% and 9%). *IntelliMetric* picked only one of the 18 nonsuccessful students, and humans picked only 6 of them.

Ericsson, Patricia Freitag & Haswell, Richard H. (Eds.). (2006). *Machine scoring of student essays: Truth and consequences*. Logan, UT: Utah State University Press.*

A compilation of seventeen original essays by teachers of composition discussing the assessment methodology and educational impact of commercial computer-based essay-rating software such as the College Board's *WritePlacer Plus*, ACT's *e-Write*, ETS's *e-rater*, Measurement, Inc.'s *Project Essay Grade (PEG),* as well as essay feedback software such as Vantage Learning's *MY Access!* and ETS's *Criterion*. Addresses many issues related to the machine scoring of writing: historical understandings of the technology (Ken S. McAllister & Edward M. White; Richard Haswell; Bob Broad); investigation into the capability of the machinery to "read" student writing (Patricia F. Ericsson; Chris M. Anson; Edmund Jones; William Condon); discussions of how students have reacted to machine scoring (Anne Herrington & Charles Moran); analysis of the poor validity in placing students with machine-produced scores (Richard N. Matzen, Jr. & Colleen Sorensen; William W. Ziegler; Teri T. Maddox); a comparison of machine scores on student essays with writing-faculty evaluations (Edmund Jones); a discussion of how writers can compromise assessment by fooling the computer (Tim McGee); the complicity of the composition discipline with the methods and motives of machine scoring (Richard Haswell); writing instructors' positive uses of some kinds of computer analysis, such as word-processing text-checkers and feedback programs (Carl Whithaus); an analysis of the educational and political ramifications of using automated grading software in a WAC content course (Edward Brent & Martha Townsend); and an analysis of commercial promotional material of software packages (Beth Ann Rothermel). Includes a 190-item bibliography of machine scoring of student writing spanning the years 1962-2005 (Richard Haswell), and a glossary of terms and products.

Wilson, Maja. (2006). Apologies to Sandra Cisneros: How ETS's computer-based writing assessment misses the mark. *Rethinking Schools 20*(3).*

Wilson tested Educational Testing Service's *Critique*, the part of *Criterion* that provides "diagnostic feedback," by sending it Sandra Cisneros' chapter "My Name," from *House on Mango Street*. *Critique* found problems in repetition, sentence syntax, sentence length, organization, and development. Wilson then rewrote "My Name" according to *Critique's* recommendations, which required adding an introduction, a thesis statement, a conclusion, and 270 words, turning it into a wordy, humdrum, formulaic five-paragraph essay.

might fit within a larger ecology as one among a family of assessment techniques supporting the development of digitally enhanced literacy" (105). The study used scores from multiple writing measures including the SAT-W, beginning of the semester impromptu essays scored by Criterion, an essay written over an extended time line scored by faculty, end of semester portfolios, and course grades. The researchers compare the scores and conclude that when embedded in a course, AES can be used as "an early warning system for instructors and their students." Authors also noted concerns that over-reliance on AES could result in a fixation on error and surface features such as length.

Perelman, Les. (2012). Construct validity, length, score, and time in holistically graded writing assessments: The case against automated essay scoring (AES). In Charles Bazerman et al. (Eds.) *International Advances in Writing Research: Cultures, Places, Measures* (pp. 121-150). Fort Collins, CO: WAC Clearinghouse & Parlor Press.

An accessible critique of the writing tasks (the timed impromptu) and the automated essay scoring process. The author argues that while "the whole enterprise of automated essay scoring claims various kinds of construct validity, the measures it employs substantially fail to represent any reasonable real-world construct of writing ability" (p. 121). He explains how length affects scoring: for short impromptus, length correlates to scores, but once more time is given to write and subjects are known in advance, the influence of length on scores diminishes. He also explains how AES is different from holistic scoring in spite of a single number being generated because that number is generated by a set of analytical measures. These individual measures (e.g., word length, sentence length, grammar, and mechanics) are not the same construct it purports to measure (writing ability). The AES program discussed is primarily the ETS erater 2.0 system because ETS has been more transparent about it than other AES developers. Perelman draws on his own research into AES, many ETS technical reports and peer-reviewed research in making his argument.

Bridgeman, Brent, Trapani, Catherine & Yigal, Attali. (2012). Comparison of human and machine scoring of essays: Differences by gender, ethnicity, and country. *Applied Measurement in Education* 25(1): 27-40.*

Reports on two studies comparing human and machine scoring in terms of certain sub-populations. The data examines the scoring of writing samples for high stakes exams—the Graduate Record Exam (GRE) and the Test of English as a Foreign Language internet-based form (TOEFL iBT) that are scored with e-rater, an automated essay scoring program. The study uses a large pool of US and international test-takers. The authors, all affiliated with the Educational Testing Service, contend that the studies reported here build on the earlier work of Chodrow and Burstein (2004) in three ways: 1) it uses the a more recent version of e-rater that considers micro-features; 2) the samples include both domestic US subgroups and more comprehensive international test-takers; and 3) it identifies "some of the features that appear to contribute to the discrepancy between human and machine scores"29). The authors conclude that while "differences between human and e-rater scores for various ethnic and language or country subgroups are generally not large, they are substantial enough that they should not be ignored" (38). Essays that are slightly off topic tend to get higher scores by the e-rater. They also explain that "it appears that, for certain groups, essays that are well organized and developed," but are flawed in terms of "grammar, usage, and mechanics, tend to get higher scores from e-rater than human scorers" (39).

Cope, Bill, Kalantzis, Mary, McCarthey, Sarah, Vojak, Colleen & Kline, Sonia. (2011). Technology-mediated writing assessments: Principles and processes. *Computers and Composition* 28, 79–96.

This article is not specifically focused on machine scoring but argues for a more comprehensive approach to the assessment of writing with technology. Meaningful assessment, the authors argue, should "be situated in a knowledge-making practice, draw explicitly on social cognition, measure metacognition, address multimodal texts, be 'for learning,' not just 'of learning,' be ubiquitous" (p. 81-82). Technology is defined more broadly than merely programs for machine scoring although it does include ways that those types of programs may be incorporated in a more comprehensive approach. It is a companion piece to the authors' other article in the same issue of *Computers and Composition* (see below).

Vojak, Colleen, Kline, Sonia, Cope, Bill, McCarthey, Sarah, & Kalantzis, Mary (2011). New Spaces and old places: An analysis of writing assessment software. *Computers and Composition* 28, 97-111.

A systematic review of seventeen computer-based writing assessment programs, both those that score or rate essays as well as those that provide technology–mediated assessment. The programs included, among others, Criterion, MY Access! Essayrater, MyCompLab, Project Essay Grader, and Calibrated Peer Review. The analysis framed writing as "a socially situated activity" that is "functionally and formally diverse" and considers it to be "a meaning-making activity that can be conveyed in multiple modalities" (p. 98). Authors reviewed various components of each program, considering its components such as its primary purpose, the underlying primary algorithm, feedback mechanisms, genres/forms of writing promoted, and opportunities for engaging in writing process. They also identified strengths and weaknesses for each program. Although this review considers more than AES, it includes it as part of a larger movement to incorporate technology in various forms of writing assessment, whether formative or summative. The authors conclude that the programs do help raise test scores but that they "largely neglect the potential technology has" in terms of the three fundamental understandings of writing that they identified promoting instead a "narrow view that conforms to systems requirements in an era of testing and accountability." They "found evidence of formulaic approaches, non-specific feedback, incorrect identification of errors, a strong emphasis on writing mechanics such as grammar and punctuation, and a tendency to value length over content" and that the programs assumed "that successful student writers would reproduce conventional, purely written linguistic generic structures" (108).