An Annotated Bibliography of Writing Assessment: Machine Scoring and Evaluation of Essay-length Writing

by Richard Haswell, **Whitney Donnelly**, **Vicki Hester**, **Peggy O'Neill**, and **Ellen Schendel**

This installment of the *JWA* annotated bibliography focuses on the phenomenon of machine scoring of whole essays composed by students and others. "Machine scoring" is defined as the rating of extended or essay writing by means of automated, computerized technology. Excluded is scoring of paragraph-sized free responses of the sort that occur in academic course examinations. Also excluded is software that checks only grammar, style, and spelling. Included, however, is software that provides other kinds of evaluative or diagnostic feedback along with a holistic score. While some entries in this bibliography describe, validate, and critique the ways computers "read" texts and generate scores and feedback, other sources critically examine how these results are used. The topic is timely, since the use of machine scoring of essays is rapidly growing in standardized testing, sorting of job and college applicants, admission to college, placement into and exit out of writing courses, content tests in academic courses, and value-added study of learning outcomes.

This installment of the *JWA* annotated bibliography focuses on the phenomenon of machine scoring of whole essays composed by students and others. This bibliography extends Rich Haswell's bibliography on the topic, "A Bibliography of Machine Scoring of Student Writing, 1962-2005" in *Machine Scoring of Student Essays: Truth and Consequences* (Logan, UT: Utah State University Press, pp. 234-243). Here we define "machine scoring" as the rating of extended or essay writing by means of automated, computerized technology. We exclude scoring of paragraph-sized free responses of the sort that occur in academic course examinations. We also exclude software that checks only grammar, style, and spelling. But we include software that provides other kinds of evaluative or diagnostic feedback along with a holistic score.

While some entries in this bibliography describe, validate, and critique the ways computers "read" texts and generate scores and feedback, other sources critically examine how these results are used. The topic is timely, since the use of machine scoring of essays is rapidly growing in standardized testing, sorting of job and college applicants, admission to college, placement into and exit out of writing courses, content tests in academic courses, and value-added study of learning outcomes.

With this abridged collection of sources, our goal is to provide readers of *JWA* with a primer on the topic, or at least an introduction to the methods, jargon, and implications associated with computer-based writing evaluation. We have omitted pieces of historical interest, focusing mainly on developments in the last twenty years, although some pieces recount the forty-five years of machine scoring history. The scholarship should provide teachers of writing, writing program administrators, and writing assessment specialists a broad base of knowledge about how machine scoring is used and to what ends, as well as insights into some of the main theoretical and pedagogical issues current among those who advocate for or caution against machine scoring. At the least, we hope this collection of resources will provide readers with a sense of the conversation about machine scoring by experts in the enterprise of making and marketing the software and scholars in the field of writing studies and composition instruction. We all need to be better prepared to articulate the benefits, limitations, and problems of using machine scoring as a method of writing assessment and response.

Attali, Yagli. (2004). Exploring the feedback and revision features of *Criterion*. Paper presented at the National Council of Measurement in Education, April 12-16, San Diego, CA.
http://www.ets.org/Media/Research/pdf/erater_NCME_2004_Attali_B.pdf
Reports on a large-scale, statistically based study of the changes in student essays, from grades 6-12, from the first to last submission to Educational Testing Services's *Criterion*. Evaluates "the effectiveness of the automated feedback" feature to decide whether "students understand the feedback provided to them and have the ability to attend to the comments" (p. 17). Concludes that statistical analysis demonstrates students were able to significantly lower the rate of specific errors and significantly increase the occurrence of certain desirable discourse elements (e.g., introduction, conclusion); therefore, students are able to revise and improve their essays using the automated feedback system. Although the study's research questions are framed in terms of how students understand the feedback and why they choose to revise, the comparison of first and last essay submissions is purely text based. The study doesn't present any data about how feedback was used by the students, if there was any teacher intervention, or if other factors external to the feedback could have influenced the students' revisions and final texts.

Baron, Dennis. (1998). When professors get A's and machines get F's. *The Chronicle of Higher Education* (November 29), p. A56.
Examines Intelligent Essay Assessor (IEA), arguing that while this assessment program claims to scan student essays consistently and objectively in seconds using what its developers call "latent semantic analysis," consistent, objective readings would not necessarily improve classroom assessments or teaching--if such readings could be achieved. Baron posits that consistency, whether in the grading or reading of student essays, is not a human trait.

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 the 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).

Bridgeman, Brent, Trapani, Catherine, and 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" (p. 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" (p. 38). Essays that are slightly off topic tend to get higher scores by the e-rater. They also explain, "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" (p. 39).

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.

Burstein, Jill, Leacock, Claudia, & Swartz, Richard. (2001). *Automated evaluation of essays and short answers*. Princeton, NJ: Educational Testing Service. http://www.etstechnologies.com
Focuses on *e-rater*, Educational Testing Services's main software platform for evaluating essays, first used commercially in 1999 to score the Graduate Management Admission Test (GMAT). Discusses history of the development of the product; connection with ETS holistic scoring; natural-language processing features; statistical modules that analyze syntactic variety, arrangement of ideas, and vocabulary usage; "training" of the program with human-scored essays; and feedback for writers as embedded in *Criterion*, ETS's web-based essay evaluation service. Good introduction to an essay-scoring program.

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.

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 *e-rater*) 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" (p. 33). 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" (p. 35).

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.
Uses naturalistic classroom investigation to see how effectively *MY Access!* worked for ESL students in Taiwan. Finds 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.

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.

Conference on College Composition and Communication Executive Committee. (2004). *CCCC position statement on teaching, learning, and assessing writing in digital environments*.
http://www.ncte.org/cccc/resources/positions/digitalenvironments.
The authors of this policy statement (Kathleen Yancey as chair, Andrea Lunsford, James McDonald, Charles Moran, Michael Neal, Chet Pryor, Duane Roen, Cindy Selfe) assert, "We oppose the use of machine-scored writing in the assessment of writing." Although automated scoring has the advantage of speed, it violates the social nature and purposes of writing, hides the criteria by which writing is valued, and sends a message to students, high schools, and the public that post-secondary institutions do not value writing as human communication

Chodorow, Martin, & Burstein, Jill. (2004). *Beyond essay length: Evaluating e-rater's performance on TOEFL essays* (TOEFL research report, No. RR-04-73). Princeton, NJ: Educational Testing Service.
Using TOEFL essays, analyzes Educational Testing Services *e-rater* scores, human holistic scores, and essay length, and finds that the newer version of *e-rater* (*e-rater01*) is less reliant on length, with more of the score explained by topic and vocabulary measures. When essay length is removed in a regression model, however, even *e-rater01*'s other measures account for only .14 of the human-reader variance in scores. Notes that *e-rater* has lower levels of exact agreement with human raters (.53) than human raters have with themselves (.56).

Crusan, Deborah. (2010). *Assessment in the second language classroom*. Ann Arbor, MI: University of Michigan Press.
With an interest in second-language instruction, tests Pearson Educational's *Intelligent Essay Assessor* and finds the diagnosis "vague and unhelpful" (p. 165). For instance, *IEA* said that the introduction was "missing, undeveloped, or predictable. Which was it?" (p. 166). Crusan's 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).

Drechsel, Joanne. (1999). Writing into silence: Losing voice with writing assessment technology. *Teaching English in the Two-Year College, 26*(4), 380-387.
Examines *Intelligent Essay Assessor* and argues that machine scoring is inconsistent with composition theories. Argues that meaning resides in the negotiation between readers and texts, and because writing assessment is contextual, assessments should be rooted in classroom theories and practices. Further, because writers (and texts) need active readers to interpret and help to construct textual meaning, machines cannot replace human readers. Argues that students who return to college after many years often write about what they know--about their personal hopes and dreams--with the anticipation of human reactions to their essays. Posits that by listening, readers empower writers; computerized placements dehumanize both readers and writers as they silence students and disregard the expertise of faculty readers.

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 Services'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 Services explained that "we're a long way from computers actually reading stuff."

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

Compiles 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.

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.

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.
Argues against educators and assessors "relying principally or exclusively on standardized assessment programs or using automated, externally developed writing assessment programs" (p. 177). The authors submitted an essay written by Moran to Educational Testing Services'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).

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.

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 non-successful students, and humans picked only 6 of them.

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.

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 draft and revise 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.

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).

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).

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.
Compares *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. Concludes 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.

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.

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

Lawrence Erlbaum.
Compiles 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 & Kathyrn 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" (p. xv), but no new information along those lines is presented.

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.
Reviews 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).

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 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.

Streeter, Lynn, Psotka, Joseph, Laham, Darrell, & MacCuish, Don. (2002). The credible grading machine: Automated essay scoring in the DoD. In *Proceedings of the Interservice/Industry Training, Simulation and Education Conference, 2002*. Orlando, FL: I/ITSEC
Compares human ratings of essays with ratings generated by *Intelligent Essay Assessor* (*IEA*), the machine-scoring software developed by KAT. The objective of the authors, personnel from the Department of Defense (DoD) and from Knowledge Analysis Technologies (KAT), is to validate the use of *IEA* in scoring of essays written by DoD trainees. Their benchmark--standard in the machine-scoring industry--is software whose scores correlated as well with human scorers as human scorers correlated with themselves. This they achieved. On four different memos, ranging from two to three pages long, correlations of *IEA* scores with DoD instructor scores averaged .50, and correlations of instructor scores with instructor scores also averaged .50. Another trial compared the scoring of take-home essay examinations written by students at the Air Command and Staff College, essays averaging 2,000 words long. Instructor to instructor reliabilities were .33 and .31; *IEA* to instructor reliabilities were .36 and .35. Undeterred by these weak interrater reliability coefficients, the authors conclude that "the automated grading software performed as well as the better instructors in both trials, and well enough to be usefully applied to military instruction."

Valenti, Salvatore, Neri, Francesca, & Cucchiarelli, Alessandro. (2003). An overview of current research on automated essay grading. *Journal of Information Technology Education* 2, 319-330
Describes and analyzes ten current computerized tools for automated essay grading: *Project Essay Grade, Intelligent Essay Assessor, Educational Testing Service I, e-rater, c-rater, BETSY, Intelligent Essay Marking System, SEAR, Paperless School free-text Marking English*, and *Automark*. The piece concludes with a discussion of problems in comparing the performance of these programs, noting that "the most relevant problem in the field of automated essay grading is the difficulty of obtaining a large corpus of essays each with its own grade on which experts agree" (p. 328).

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

Reports one of the few empirical comparisons of machine with human scoring conducted outside testing companies. 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).

Warschauer, Mark, & Ware, Paige D. (2006). Automated writing evaluation: Defining the classroom research agenda. *Language Teaching Research, 10*(2), 157-180.
Observes ESL classroom teachers using automated feedback programs, and finds both good and bad effects. For instance, the technology encouraged students to turn in more than one draft of an assignment, but it "dehumanized" the act of writing by "eliminating the human element" (p. 176). The authors feel that more classroom research is needed before deciding the true worth of machine analysis.

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

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.

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.

Wilson, Maja. (2006). Apologies to Sandra Cisneros: How ETS's computer-based writing assessment misses the mark. *Rethinking Schools, 20*(3), 148-155.

Tests Educational Testing Service' *Critique*, the part of *Criterion* that provides "diagnostic feedback," by sending it Sandra Cisneros' chapter "My Name," from *The 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.

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. Uses *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.

Richard Haswell
Texas A&M University, Corpus Christi (emeritus)

Whitney Donnelly
California State University, Stanislaus

Vicki Hester
St. Mary's University

Peggy O'Neill
Loyola University Maryland

Ellen Schendel
Grand Valley State University