International Journal of TESOL and Learning
March 2015, Volume 4, Issue 1, pp. 20-32

http://www.untestedideas.com/journals.php?journal=IJTL
http://untestedideas.net/journal.php?journal=ijt
ISSN: 2168-9474 (Print); ISSN: 2168-9482 (Online)

REVIEW ARTICLE

And then a Miracle Occurs: The Use of Computers to Assess Student Writing

Deborah Crusan*
Wright State University

The machine scoring of student writing stands as one of *the* hot topics in writing assessment. Companies promote these products as time- and money-saving. However, the salient question remains: Is this technology appropriate for use in the English as a Second Language writing (SLW) classroom? Administrators and second language writing professionals often seem be at odds when it comes to the use of such programs. Proponents typically express that electronic grading is of great benefit, mainly because it facilitates scoring large numbers of student essays in a short time. Scoring efficiency appeals mainly to administrators searching for cost effective ways to provide classroom writing instruction. Equally appealing to administrators is the notion that class size can be increased as the burden of grading is removed from the teacher. However, many second language writing professionals are dismayed by the notion of a computer scoring or responding to student writing. Although it is important that practitioners not rely solely on their initial response, it is natural that they express concern. However, as researchers, we recognize the need to thoroughly examine the topic, weighing both positive and negative outcomes of the use of such platforms.This issue needs to be studied from multiple perspectives so that teachers are informed about using computers to assess student writing. In this paper, the views of educators, administrators, and developers of artificial intelligence are examined with respect to the use of machines to score student writing. These programs are then situated in the context of writing assessment theory and their use critiqued in terms of pedagogical value. The paper concludes with an exploration of both the consequences and potential benefits of using these systems in second language writing classrooms as well as suggestions to help second language writing professionals work with administrators pushing for this type of assessment for instructional purposes.

*Correspondence should be sent to Deborah Crusan, rofessor of TESOL/Applied Linguistics at Wright State University, Dayton, Ohio, USA. E-mail: *deborah.crusan@wright.edu*

INTRODUCTION

Across the globe, increased attention is being paid to understanding the processes and components of writing and its implications for the teaching of writing. This relatively recent focus on the teaching of writing has been facilitated in part by "linguistic demands of a globalized informational society" (Warschauer & Ware, 2006, p. 158). As the number of writing classes explodes, so does the work of the writing teacher, often reading and responding both formatively and summatively to the writing of 100 or more students each term. This work, critical to the growth of the writer, has added hours to teachers' days and increased their workload significantly. How to deal with this workload while still providing effective feedback to students has been the topic of countless conference presentations, doctoral dissertations, academic books, and graduate classes. One answer is the use of computers to score student writing. Research into the mechanisms, development, reliability and validity, include Automated Essay Evaluation (AEE), Automated Essay Scoring (AES), Automated Writing Evaluation (AWE), and the machine scoring of essays (Ericsson & Haswell, 2006). The term *machine scoring of essays* will be used in this paper to refer to those aliases.

Testing has always been a controversial concept, and writing assessment ranks as possibly the most contentious category. Discipline specific factions often assume rigid positions regarding which testing methods work most effectively. Cost and ease of administration are often component parts of decisions about the forms of writing assessment that will be used. The machine scoring of essays sits squarely in the midst of this controversy. Advocates and opponents argue passionately about both the benefits and drawbacks of this contentious issue in writing assessment. According to Herrington and Moran,

> Firms that are marketing the machine scoring of student writing all explicitly or implicitly define the task of reading, evaluating, and responding to student writing not as a complex, demanding, and rewarding aspect of our teaching, but as a "burden" that should be lifted from our shoulders (2001, p. 480).

Other scholars (Crusan, 2010a; Haswell, 2006) refer to the burden of grading and infer that teachers might be getting what they have long asked for – relief from the staggering amount of grading they do. On the other hand, Shermis and Burstein (2003) admit that while machine scoring is far from perfect, it holds the "promise of addressing most of the writing of the classroom" (p. xv).

Certainly though, many writing teachers have a visceral response to even the idea of machine scoring, a quite understandable reaction. The fear of replacing teacher-as-reader is real; however, this reaction cannot allow us to hide our heads in the sand and pretend that the phenomenon of machine scoring will somehow disappear if ignored. Instead, it is important to "thoroughly examine implications, advantages, and disadvantages prior to making a decision about using (or not) any of the many platforms available to score the writing our students produce" (Crusan, 2010a, p. 159) to control the pedagogy of their classrooms and maintain autonomy. Simply put, teachers need to be informed about current topics in writing and writing assessment. Particularly important, at least for the purposes of

this paper, are the effects of machine scoring on second language writers – a topic which, until recently, has not enjoyed much attention; however, investigations into second language writers and writing, their attitudes toward machine scoring, and how they might be affected by machine scoring have begun to appear more frequently in the literature (Dikli & Bleyle, 2014; Xi, 2010).

Machine Scoring Defined

Computerized essay scoring is defined as computer technology that evaluates and scores written prose (Shermis & Burstein, 2003). It is also known by several aliases mentioned above although McAllister and White (2006) argue that the operation is far from autonomous and, instead, requires human interaction. Rather than using any of the above aliases currently employed, they insist on the term *computer-assisted writing assessment* as they see the process as only partially automated.

Technologies

To date, there are three basic technologies used to score writing using a computer. The first is *Artificial Intelligence* (AI): defined as the science of making intelligent machines, its applications include game playing and speech recognition (Dikli, 2006). The next is *Latent Semantic Analysis* (LSA): "a statistical model of word usage that permits comparisons of the semantic similarity between pieces of textual information" (Foltz, 1996, p. 197). LSA is considered highly correlated with human meaning (Landauer, Laham, & Foltz, 2003a).

The third technology is *Natural Language Processing* (NLP), a sub-field of artificial intelligence. NLP is complex to understand. It has several categories of language tasks: speech recognition, syntactic analysis, discourse analysis, information extraction, and machine translation (Dikli, 2006). NLP captures syntactic variety, identifies cue words, terms, and syntactic structures, and analyzes text for word frequencies and word weights.

These three techniques for processing information have been adapted for use by testing companies, led by Educational Testing Services (ETS), Pearson, and Vantage Learning. Table 1 displays company products and use of technology.

Table 1
Machine Scoring Software

Company	Scoring Engine	Technique	Instructional Program	Scoring
ETS	e-rater ®	Natural Language Processing (NLP)	CRITERION sm	Single holistic score
Pearson	Intelligent Essay Assessor (IEA)	Latent Semantic Analysis (LSA)	WriteToLearn™	Holistic and component scoring
Vantage Learning	IntelliMetric ™	Artificial Intelligence (AI)	My Access!® (and others)	Holistic and component scoring

Table adapted from Dikli (2006) and Warschauer & Ware (2006).

These three companies have employed the powerful scoring engines (e-rater ®, Intelligent Essay Assessor, and IntelliMetric ™) and the techniques of NLP, LSA, and AI to create programs for in-class uses.

Table 2
Instructional Applications of Machine Scoring Software

Program	Company	Technology	Claims	Uses
CRITERION sm	ETS	e-rater ® /Natural Language Processing	Evaluates writing skills and provides score reporting and diagnostic feedback to instructors and students. Students draft and submit essays and receive immediate feedback in the form of a holistic score and diagnostic annotations within each essay that guides instruction.	In class feedback, score reporting, and diagnostic feedback to instructors and students.
MY Access!®	Vantage Learning	IntelliMetric ™ /artificial intelligence	Provides students with the practice they need to improve their writing skills. The program's powerful scoring engine grades students' essays instantly and provides targeted feedback, freeing teachers from grading thousands of papers by hand and giving them more time to conduct differentiated instruction and curriculum planning.	In class writing
WriteToLearn	Pearson	Intelligent Essay Assessor (IEA) /latent semantic analysis	Provides easy to read reports to enable teachers to monitor class and individual progress; can help students improve reading comprehension and writing skills	In class writing skills and reading comprehensio n development tool

Table adapted from Crusan (2010a)

A BRIEF HISTORY OF THE MACHINE SCORING OF ESSAYS

Huot (2006) noted that historically, writing assessment has been controlled by measurement experts who "developed, constructed, and privatized" (p. 549) it, creating a "technological apparatus whose inner workings are known only to those with specialized knowledge" (p. 549). Nowhere is this truer than with machine scoring. Ellis Page developed the first recognized essay scoring machine in 1966 (Page, 2003) and called it Project Essay

Grader (PEG). PEG employed computer automation to assess student writing, calculating variables such as prepositions and articles, which he called *proxes* (Elliot, 2005, p. 221). In 1968, he reported that grades gleaned from machine scored proxes were highly correlated with scores of human raters. This early experiment was the first correlational analysis and demonstrated that a computer seemed to assess essays as well as human raters.

Page believed machine scoring would impact the world of assessment, and it has. There were setbacks though as interest in his technology subsided in the 1960s because of the instability of computers; however a more sophisticated version made a comeback in the 1980s when computers and technology became more stable (Warschauer & Ware, 2006). Page continued this work until his death in 2005, his zeal never diminishing. It is much to his credit that the machine scoring of essays remains an interest in writing assessment today, an interest so great that "most established test development firms in the United States now have at least one platform for the machine scoring of essays" (Crusan, 2010a, p. 161).

The work focusing on machine scoring is an ongoing process, even sparking demonstrations and public competitions (Shermis, 2014). Shermis detailed the results of one such competition – the Automated Student Assessment Prize – intended to stimulate innovations in machine scoring. Shermis (2014) explains, "The goal of the public competition was to encourage and make available to the commercial vendors new software technology or programming approaches that improve the overall agreement of the machine scoring algorithms with human scores" (p. 72). Prizes were awarded to the system that best mimics the scores of experienced human raters. A second prize focused on identifying new programs or algorithms, which improve the correlation between machine and human raters. In combination, these two prizes are intended to push the state-of-the art of machine scoring of essays, making them comparable to the scoring strategies used by experienced teachers/raters.

VARIOUS PERSPECTIVES ON MACHINE SCORING

Advocates: The definitions of writing and writing assessment are best understood through the lens of machine scoring developers and their advocates to understand their reasoning and their inclination to quantify the various processes of writing. To them, writing should be defined as production of text, divorced from any context. Essentially, the rationale for developing machines to score writing centers on two issues.

The first rationale for support of machine scoring the assessment of writing has always been that human raters are subjective and that human scoring is unreliable (i.e. human scoring suffers from flaws). Humans are prone to foibles – they get tired, distracted, and hungry. These issues, proponents argue, interfere with objectivity, as humans tend to overgeneralize; that is, they tend to judge one characteristic based on another observed trait. For example, if a man is viewed as handsome, it is assumed that he is also intelligent. This halo effect, Bereiter (2003) claims, might be overcome by the use of machine scoring technology. However, it is interesting to note that calibration depends on the correlation of machine scores with human raters' scores.

The second rationale for support of machine scoring is the perception that reading and responding to student writing is a supposed burden on teachers. When writing teachers are asked about the most difficult task in their work, almost to a person, the answer is grading. Grading takes time and energy. Providing feedback on students' drafts is cognitively taxing and often a thankless and underappreciated task. So administrators, seeking to relieve faculty of this tremendous task now have access to tools, which can provide a fast and easy brand of assessment.

Tied to this ease in assessment is the omnipresent managerial quest to save money. Administrators appreciate that fact that large numbers of essays can be graded quickly and inexpensively. They welcome machine scoring for its ability to process both testing and instruction for second language writers and writing. Further, administrators could very well figure that a writing teacher relieved of the burden of grading (Haswell, 2006) might therefore be able to manage larger classes, another money saver in terms of salaries, benefits, and other administrative worries.

Proponents of machine scoring (Shermis & Burstein, 2003; Shermis & Burstein, 2013; Landauer, Laham, & Foltz, 2003b) maintain that machine scoring strongly correlates with normed scores by human raters. They also assert that machine scoring retains construct validity. To understand this concept, it is important to comprehend the meaning of construct. Briefly, writing ability is a construct – it is a skill or ability that happens in the brain and is defined by composition theory. For something to have construct validity, it needs to be measuring the construct. Therefore, proponents of machine scoring claim that it can measure writing ability. Further, Weigle (2010) reminds us, "There is no single ideal testing format or scoring procedure, but *e-rater* certainly holds promise as an additional tool in the language tester's toolkit" (p. 350).

Adversaries: Most who are critical of machine scoring of essays define writing as a social act, written to a specific audience for a specific purpose, revised several times to clarify meaning through intercession of others, particularly peers and teachers. To them, machine scoring software is not a valid substitute for what happens in the writing classroom nor the individualized feedback a writing teacher gives in interactions with her students (Warschauer & Ware, 2006). Teachers worry that when disinterested machines score essays, the very definition of writing is altered. They worry about how students will think about writing when their audience is a machine and not a human being.

Standardized writing assessment is detached from teaching and often reduces writing to the least common denominator – that which can be measured and counted. It often leads to the loss of teacher autonomy. Even though researchers (Landauer, Laham, & Foltz, 2003b) claim that machine scoring is a stronger measure of expression and knowledge retrieval than multiple choice, teachers are concerned that machine scoring constructs writing as an individual activity rather than an activity that is shared, disregarding the social interaction of writing. In machine scoring, essays seem to be based on norms and become displays of technical prowess rather than communication for real purposes.

Machine scoring also seems to emphasize grammar, usage, and organization, which intentionally or unintentionally displaces teacher authority. With the current political milieu with its focus on standardized testing, it is important that we protect the historical role of teachers and in doing so the human element in writing and the higher order feedback with is

currently far beyond the abilities of computerized grading or feedback programs. The important question – why do students need to know how to write – needs better answers than because they'll need to write in college. Why do students need to know how to write? They need to know how to write in order to discover who they are, and where they fit in the world (Patterson, 2007). In sum, the ability to write coherent well-researched documents is strongly correlated with financial and career related success.

The act of scoring is another issue. While proponents of machine scoring cite its ability to score essays within seconds, Perelman (2014) refutes the assertion made by Shermis (2014) that machines can dependably replicate and at times surpass human raters' performance in high-stakes assessments. Perelman argues that machines heavily favor essay length and reward it in the computation of scores.

BENEFITS AND CONSEQUENCES OF MACHINE SCORING

One of the touted benefits of using machine scoring is its use in high stakes assessment of writing. High stakes writing tests typically require students to write either an informative essay or an argumentative essay. These essay modalities limit the focus of student writing. However, Madaus, Russell, and Higgins (2009) argue, "Focusing students on certain topics is designed to produce objective scores" (p. 179). It follows then that this focus on objectivity creates an opening for the use of machine scoring since machine scoring can generate results as impartial as normed human raters while shrinking essay scoring time and cost. So the first benefit of using machines to score essays is efficiency.

Another possible benefit of machine scoring is authenticity. Madaus, Russell, and Higgins (2009) argue that the relative ease of submission, subsequent immediate feedback, and interim scoring can lead students to revise and resubmit, concepts largely absent from high stakes writing assessment when the student writes with pen and paper. Arguing for the prospect of authenticity in machine scoring, Madaus, Russell, & Higgins (2009) posit, " . . . allowing students to receive preliminary feedback and then revise their writing provides a better representation of the writing process and may result in more valid measures of student writing abilities" (pp. 180-181). For high stakes assessment, which does not appear to be going away anytime soon, the creation of an even remotely more authentic environment in which to write is certainly beneficial.

Additionally, students can greatly benefit when the results of machine scoring are used for instructional purposes to improve their writing. Chen and Cheng (2008) claim that the diagnostic feedback from machine scoring programs such as MY Access!® "seems pedagogically appealing for formative learning" (p. 97). In a related vein, the use of programs like MY Access!® might certainly promote learner autonomy because it is meant to encourage multiple revision and self evaluation. Therefore, it might be useful in the classroom as supplemental to or as an extension of teacher feedback.

Another possible advantage to machine scoring concerns the unlimited amount of source material at the disposal of the machine. Consider Watson, the computer who once played the television quiz show *Jeopardy* and beat his human competitors, winning $77,147. Although Watson was far from perfect, the machine had access to countless sources not so

readily available to the human brain. IBM, Watson's creator, offered "proof that the company has taken a big step toward a world in which intelligent machines will understand and respond to humans, and perhaps inevitably, replace some of them" (Markoff, 2011, p. 1). The great depth of source material available to a machine and the speed with which that data can be accessed is beyond the scope of the human brain; an essay scoring program has access to thousands of essays that have been uploaded into its memory, from which the machine pulls to score and provide feedback. As an aside, while the notion of potential replacement of humans with machines might cause consternation, decades of research and development lie ahead before that eventuality (Williamson, 2014).

This notion that computers can replace humans strikes fear in the hearts of teachers and others, who see themselves in a futuristic setting, as nothing more than technical room monitors who troubleshoot when the occasional technical glitch occurs and have nothing to do with helping to create knowledge and understanding for their students. Quite possibly, this is the biggest disadvantage to machine scoring – its probable promotion of the deskilling of teachers. Rather than empowering students, teachers are becoming, more and more, workers in the classroom, distributing materials and reading poorly written scripts. Not trusted to create materials they've been taught to construct in their university classes, teachers are instead forced to use mandated materials concocted by a textbook company far removed from their contest and unacquainted with the needs of their students.

Another issue in the automated scoring of essays is the idea of writing to a machine. One of the principal tenets of composition theory is the idea of audience. Consideration of the potential reader or readers is a central component of any piece of writing. For most writers, teachers, or reviewers, this is of paramount importance in planning, writing, and reviewing a document. What people say and how they say it hinges on the background of their audience and whether it is expert or possesses a more general knowledge of their topic.

Weigle (2010) points out several issues concerning machine scoring particularly for English language learners. One question concerns the mechanical aspects of language – grammar, usage, and vocabulary – and whether the features used are those that present problems for nonnative speakers. She also raises the issue of nonnative speaker errors and their identification. In my own classes, I saw evidence that the machine could not adequately describe students' linguistic errors. In a study of two machine scoring systems' ability to determine and revise article and preposition errors, Chodorow, Gamon, and Tetreault (2010) concluded, " . . . the complexity and variability of article and preposition usage make these elements of English syntax . . . difficult for automated error-detection and correction systems" (p. 420). It seems, then, that machine scoring has not yet been perfected, but with technological advances occurring faster and faster, can that day be far off?

Another problem with machine scoring revolves around the issue of textual borrowing. While few, if any, researchers have mentioned this problem, a high incidence of plagiarism is always possible if students are able to access the Internet. However, few studies have investigated the use of machine scoring in the second language writing classroom (Chen & Cheng, 2008; Choi & Lee, 2010; Dikil, 2010; Warschauer & Grimes, 2008; Warschauer & Ware, 2006) and none of these studies consider the notion of textual borrowing to add information or length to student essays.

One of the biggest disadvantages, it seems, is that fact that although the machine grades effectively and provides feedback to the students and to the teacher, the teacher is not really seeing her students' work. She is not intimately involved in the process of writing with her students, guiding them to better essays. The machine removes the human element from the equation, displacing teacher authority. Furthermore, it seems that the machine constructs writing as an individual activity, disregarding the social interaction of writing. When examined closely, it is obvious that the machine equates good writing with the subskills of writing – mechanics, usage, grammar, and spelling - MUGS. Clearly, all of these notions challenge the writing process, defining writing in the narrowest of terms.

However, many writers, particularly second language writers, ask for grammatical feedback on their writing. Ferris (2014) examined teachers' feedback practices and found differences among instructors' feedback practices. She also discovered gaps between what teachers say they do regarding providing feedback to their students and their actual practices. Therefore, the feedback on mechanics and grammar might be a welcome addition for students.

IMPLICATIONS FOR TEACHERS

Although it may be difficult for an admittedly biased teacher to present a balanced perspective regarding machine scoring, it is easy to see that there are distinct advantages and disadvantages to the use of computer essay grading in the classroom. More revision. More writing. However, the benefits of additional writing may exact a heavy price: students' essay scope may be confined by prompts that often encourage writing in the form of the five-paragraph essay. Of course, the five-paragraph essay should not be demonized, especially for writers whose first language is not English. Many of these writers are attempting to learn the specific requirements of English academic writing, particularly the thesis statement, the argument, the need for evidence to support that argument, and the need for citation. They are also attempting to learn argument, comparison and contrast, analysis, and description in ways often very different from the rhetoric of their first languages. For some students, these requirements might be quite foreign and require models and templates. The five-paragraph essay introduces the template of English academic writing. It provides scaffolding to facilitate learners' development. However, it is important that students recognize the need to move from the five-paragraph essay to more complex writing. As these demands for more sophisticated genres emerge, machine scoring and its accompanying classroom applications might become less useful.

Nevertheless, whether machine scoring enriches or mechanizes writing and the writing classroom depends on the teacher and the ways in which the teacher uses and relates to machine scoring (Grimes & Warschauer, 2010). If teachers use machine scoring uncritically, leaving writing instruction to the machine and not interceding in the grading process or are preoccupied with grammatical and mechanical correctness, machine scoring will indeed mechanize the writing class. On the other hand, if teachers use machine scoring for writing *practice* – to help students become familiar with multiple drafting, revision, and

editing and to help with low-level errors – they can then turn their attention to constructs such as content; therefore, students' writing will be enriched.

Machine scoring can be a double-edged sword. Its value is determined in part by the philosophy of the teacher and its implementation in the classroom. If teachers allow the machine scoring of essays in their classrooms for whatever reason, students' use of these tools should be limited, closely monitored, and controlled by the teachers. Further, if these additional components of writing are introduced into the writing classroom, it would be more congruent with composition theory and writing assessment theory if the criterion for writing, lesson goals and emphasis, and assessment were locally developed and administered (Crusan, 2010b, 2013, 2014; Huot, 2002; White, 2008; Herrington & Moran, 2001) and if administrative software capable of integrating machine feedback coupled with teacher grades is provided to writing teachers free of charge.

The machine scoring of essays will neither save education nor destroy it. Neither can it improve upon poor teaching, but in the hands of effective teachers, it can be an effective tool in the writing classroom provided teachers use it in conjunction with personalized teacher and/or peer feedback and continue to offer this higher level feedback to students throughout the multiple draft process.

My purpose has not been to demonize machine scoring but rather to "protect the kind of writing assessment teachers do in the contexts of their own institutions" (Crusan, 2010b, p. 258). Machine scoring is not the magic bullet some might believe will improve writing instruction and assessment. Nor does its judicious use doom creative or insightful writing. Rather than resisting completely, contemplation of sensible integration of machine scoring into the classroom should come with a mandate to examine it more thoroughly. Teachers and researchers need to carefully investigate machine scoring in their individual contexts. This examination, allowing better understanding of the educational implications and impact of machine scoring, is urgently needed. It is my hope that, in spite of the presence and alleged power of the machine, writing teachers remain advocates for human intervention. Finally, it is important to recognize that machine scoring will not replace teachers but can be a tool in the writing classroom.

ABOUT THE AUTHOR

The work of Deborah Crusan, professor of TESOL/Applied Linguistics at Wright State University, Dayton, OH has appeared in academic publications including *Assessing Writing* and *TESOL Quarterly* (among others) and collections about second language writing. Her research interests include writing placement and machine scoring. Her book, *Assessment in the Second Language Writing Classroom*, was published by University of Michigan Press. E-mail: *deborah.crusan@wright.edu*

REFERENCES

Bereiter, C. (2003). Foreword. In M. D. Shermis & J. C. Burstein (Eds.). Automated essay scoring: A cross-disciplinary perspective (pp. vii–ix). Mahwah, NJ: Lawrence Erlbaum.

Chen, C-F. E., & Cheng, W-Y, E. (2008). Beyond the design of automated writing evaluation: Pedagogical practices and perceived learning effectiveness in EFL writing classes. *Language Learning & Technology, 12* (2), 94-112.

Chodorow, M., Gamon, M., & Tetreault, J. (2010). The utility of grammatical error detection systems for English language learners: Feedback and assessment. *Language Testing, 27*(3), 419–436.

Choi, J., & Lee, Y. (2010). The use of feedback in the ESL writing class integrating automated Essay Scoring (AES). In D. Gibson, & B. Dodge (Eds.), *Proceedings of society for information technology & teacher education international conference* (pp. 3008–3012). Chesapeake, VA: AACE.

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

Crusan, D. (2010b). Assess thyself lest others assess thee. In T. Silva & P. K. Matsuda (Eds.), Practicing theory in second language writing (pp. 245–262). West Lafayette, IN: Parlor Press.

Crusan, D. (2013). Assessing writing. In A. J. Kunnan (Ed.), *The companion to language assessment. Volume 1: Abilities, Contexts, Learners* (pp. 201-215). Hoboken, NJ: Wiley/Blackwell.

Crusan, D. (2014). Placement testing. In C. A. Chapelle (Ed.), *The Encyclopedia of Applied Linguistics* (pp. 17-25). Hoboken, NJ: Wiley/Blackwell.

Dikli, S. (2006). An overview of automated scoring of essays. *Journal of Technology, Learning, and Assessment, 5*(1). Retrieved from *http://www.jtla.org*.

Dikli, S., & Bleyle, S. (2014). Automated essay scoring feedback for second language writers: How does it compare to instructor feedback? *Assessing Writing, 22*, 1-17. http://dx.doi.org/10.1016/j.asw.2014.03.006

Elliot, N. (2005). On a scale: A social history of writing assessment. New York: Peter Lang.

Ericsson, P. F., & Haswell, R. H. (Eds.), (2006). *Machine scoring of essays: Truth and consequences.* Logan: Utah State University Press.

Ferris, D. R. (2014). Responding to student writing: Teachers' philosophies and practices. Assessing Writing, 19, 6-23. http://dx.doi.org/10.1016/j.asw.2013.09.004

Foltz, P. W. (1996). Latent Semantic Analysis for text-based research. *Behavior Research Methods, Instruments and Computers, 28*(2), 197–202. Retrieved from *http://www-psych.nmsu.edu/~pfoltz/*

Grimes, D. & Warschauer, M. (2010). Utility in a Fallible Tool: A Multi-Site Case Study of Automated Writing Evaluation. *Journal of Technology, Learning, and Assessment, 8* (6). Retrieved from *http://www.jtla.org*

Haswell, R. H. (2006). Automatons and automated scoring: Drudges, black boxes, and dei ex machina. In P. F. Ericsson & R. Haswell (Eds.), *Machine scoring of essays: Truth and consequences* (pp. 57–78). Logan: Utah University Press.

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

Huot, B. (2002). *(Re)Articulating writing assessment for teaching and learning*. Logan: Utah State University Press.

Huot, B. (2006). Toward a new theory of writing assessment. *College Composition and Communication, 47* (4), 549-566.

Landauer, T. K., Laham, D., & Foltz, P. (2003a). Automatic scoring and annotation of essays with Intelligent Essay AssessorTM. In M. D. Shermis & J. C. Burstein (Eds.), *Automated essay scoring: A cross-disciplinary perspective* (pp. 82-106). Mahwah, NJ: Lawrence Erlbaum.

Landauer, T. K., Laham, D., & Foltz, P. (2003b). Automatic essay assessment. *Assessment in Education, 10* (3), 295-308. doi: 10.1080/0969594032000148154

Madaus, G., Russell, M., & Higgins, J. (2009). *The paradoxes of high stakes testing: How they affect students, their parents, teachers, principals, schools, and society*. Charlotte, NC: Information Age Publishing.

Markoff, J. (2011, February 16). Computer wins on Jeopardy: Trivial, it's not. *The New York Times*. Retrieved from *http://www.nytimes.com/2011/02/17/science/17jeopardy-watson.html*.

McAllister, K. S. & White, E. M. (2006). Interested complicities: The dialectic of computer-assisted writing assessment. In P. F. Ericsson and R. Haswell (Eds.), *Machine scoring of essays: Truth and consequences* (pp. 8-27). Logan, UT: Utah University Press.

Page, E. B. (2003). Project essay grade: PEG. In M. D. Shermis & J. C. Burstein (Eds.), *Automated essay scoring: A cross-disciplinary perspective* (pp. 43–54). Mahwah, NJ: Lawrence Erlbaum.

Patterson, N. (2007). The devil in the machine: Problems with computerized writing assessment. *Language Arts Journal of Michigan, 23* (1). Retrieved from *http://scholarworks.gvsu.edu/lajm/vol23/iss1/14*.

Shermis, M. D. (2014). State-of-the-art automated essay scoring: Competition, results, and future directions from a United States demonstration. *Assessing Writing, 20*, 53-76.

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

Shermis, M. D., & Burstein, J. (Eds.). (2013). *Handbook of automated essay evaluation: Current applications and new directions*. Englewood, CO: Taylor & Francis.

Vantage Learning. (2011). MY Access! school edition: Because writing matters efficacy report. Retrieved from *http://www.vantagelearning.com/school/products/myaccess/*.

Warschauer, M., & Ware, P. (2006). Automated writing evaluation: defining the classroom research agenda. *Language Teaching Research, 10* (2), 157–180.

Weigle, S. C. (2010). Validation of automated scores of TOEFL iBT tasks against non-test indicators of writing ability. *Language Testing 27*, (3), 335-353 doi: 10.1177/0265532210364406

White, E. M. (2008). Testing in and testing out. *WPA: Writing Program Administration 32* (1) 51-54.

Williamson, D. M. (2013). Probable cause: Developing warrants for automatic scoring of essays. In M. D. Shermis & J. C. Burstein (Eds.), *Handbook of automated essay evaluation: Current applications and new directions* (pp. 153-180). Englewood, CO: Taylor & Francis.

Xi, X. (2010). Automated scoring and feedback systems: Where are we and where are we heading? *Language Testing, 27*(3), 291–300. DOI: 10.1177/0265532210364643