Pedagogies: An International Journal, 3: 22–36, 2008

ISSN 1554-480X print / 1554-4818 online
DOI: 10.1080/15544800701771580

Routledge
Taylor & Francis Group

Automated Writing Assessment in the Classroom

Mark Warschauer and Douglas Grimes
University of California, Irvine

Automated essay scoring (AWE) software, which uses artificial intelligence to evaluate essays and generate feedback, has been seen as both a boon and a bane in the struggle to improve writing instruction. We used interviews, surveys, and classroom observations to study teachers and students using AWE software in 4 secondary schools. We found AWE to be a modest addition to the arsenal of teaching tools and techniques at the teacher's disposal, roughly midway between the fears of some and the hopes of others. The program saved teachers' time and encouraged more revision but did not appear to result in substantially more writing or greater attention to content and organization. Teachers' use of the software varied from school to school, based partly on students' socioeconomic status, but more notably on teachers' prior beliefs about writing pedagogy.

There is widespread agreement that students need more writing practice (see, e.g., National Commission on Writing in America's Schools and Colleges, 2003, p. 3). However, overburdened teachers have insufficient time to mark students' papers. Proponents of *automated writing evaluation* (AWE; also called *automated essay scoring* or *computerized essay scoring*), which uses artificial intelligence (AI) to score and respond to essays, claim that it can dramatically ease this burden on teachers, thus allowing more writing practice and faster improvement. Because AWE is numb to aesthetics and does not understand meaning in any ordinary sense of the word (Ericsson, 2006), critics contend that it is an Orwellian technology that merely feigns assessment and threatens to replace teachers with machines (Baron, 1998; Cheville, 2004; Conference on College Composition and Communication, 2004). To date, little research exists that might

Correspondence should be sent to Mark Warschauer, UCI Department of Education, 2001 Berkeley Place, Irvine, CA 92697–5500, USA. E-mail: markw@uci.edu

help resolve these competing claims. In this article, we provide background on the development and use of AWE in standardized testing contexts, discuss the development of AWE products for classroom use, and present the findings of an exploratory study investigating the use of AWE in four California schools.

AWE PROGRAMS AND STANDARDIZED TESTING

Automated writing evaluation emerged in the 1960s with Page Essay Grade (PEG), a program that used multiple regression analysis of measurable features of text, such as essay length and average sentence length, to build a scoring model based on a corpus of essays previously graded by hand (Shermis, Mzumara, Olson, & Harrington, 2001). AWE software remained of interest to small groups of specialists until the 1990s, when an increased global emphasis on writing instruction, advances in AI, and more widespread availability of computers and the Internet all combined to create greater developmental and marketing possibilities (for more in-depth histories and overviews of AWE, see Ericsson & Haswell, 2006; Shermis & Burstein, 2003; Warschauer & Ware, 2006).

In the 1990s, Educational Testing Service and Vantage Learning developed competing automated essay scoring engines called e-rater and Intellimetric, respectively (Burstein, 2003; Elliot & Mikulas, 2004). Like PEG, both employed regression models based on a corpus of human-graded essays, but the range of lexical, syntactic, and discourse elements taken into account became much broader and the analysis more sophisticated. For example, e-rater analyzes the rate of errors in grammar, usage, mechanics, and style; the number of required discourse elements (such as thesis statement, main idea, or supporting idea); the lexical complexity (determined by the number of unique words divided by the number of total words); the relationship of vocabulary used to that found in top-scoring essays on the same prompt; and the essay length (Attali & Burstein, 2004; Chodorow & Burstein, 2004). A third scoring engine called Intelligent Essay Assessor, developed by a group of academics and later purchased by Pearson Knowledge Technologies, uses an alternate technique called latent semantic analysis to score essays. The semantic meaning of a given piece of writing is compared with a broader corpus of textual information on a similar topic, thus requiring a smaller corpus of human-scored essays (Landauer, Laham, & Foltz, 2003).

The main commercial use of these engines has been in the grading of standardized tests. For example, the Graduate Management Admissions Test was scored from 1999 to 2005 by e-rater and since January 2006 by Intellimetric. Typically, standardized essay tests are graded by two human scorers, with a third scorer brought in if the first two scores diverge by two or more points. Automated essay-scoring engines are used in a similar fashion, replacing one of

the two original human scorers with the final human scorer again enlisted when the first two scores diverged by two points or more.

The reliability of AWE scoring has been investigated extensively by comparing the correlations between computer-generated and human-rater scores with the correlations attained from two human raters. Based on this measure, *e-rater*, Intellimetric, and Intelligent Essay Assessor all fare well (see summaries in Cohen, Ben-Simon, & Hovav, 2003; Keith, 2003), with correlations with a single human scorer usually in the .80 to .85 range, approximately the same range as correlations between two human scorers. This means that a computer-generated score will either agree with or come within a point of a human-rated score more than 95% of the time, about the same rate of agreement as that between two human scorers (Chodorow & Burstein, 2004; Elliot & Mikulas, 2004). These studies have for the most part studied large-scale standardized tests. The human-computer interrater reliability is expected to be lower in classroom contexts, where the content of student writing is likely to be more important than for the standardized tests (see discussion in Keith, 2003).

Another important psychometric issue is whether AWE software can be tricked. One study has shown that expert writers can fool AWE software programs and get relatively high scores on polished nonsensical essays (Powers, Burstein, Chodorow, Fowles, & Kukich, 2002). However, Shermis and Burstein (2003) convincingly argue that while a bad essay can get a good score, it takes a good writer to produce the bad essay to get the good score.

AWE PROGRAMS FOR THE CLASSROOM

A more recent development in AWE software is as a classroom instructional tool. Each of the main scoring engines discussed earlier has been incorporated into one or more programs directed at classroom use. ETS Technologies (a for-profit subsidiary of Educational Testing Service) has developed *Criterion*, Vantage Learning has created *My Access*, and Pearson Knowledge Technologies has launched *WriteToLearn*. In each case, the programs combine the scoring engine; a separate editing tool providing grammar, spelling, and mechanical feedback; and a suite of support resources, such as graphic organizers, model essays, dictionaries, thesauruses, and rubrics. The editing tools provide feedback similar to that offered by Microsoft Word's spelling and grammar checker but more extensively, for example, by indicating that a word may be too colloquial for an academic essay.

Teachers use these programs by assigning a writing prompt. They can develop their own prompt, but only prompts that come with the program can be scored by the software. Students either type essays on the screen or cut and paste their essays from a word processor, drawing on the editing tools or support resources

as needed. Upon submitting essays online, they instantaneously receive both a numerical score and narrative feedback, either generic from some programs or more particularized from others. For example, My Access provides standardized templates of narrative feedback based on the grade level, score, and genre with all seventh-grade students who receive a score of 3 on a persuasive essay receiving the same recommendations for improvement. Criterion provides some specific, albeit limited, feedback based on discourse analysis of each essay that has been scored, raising questions or comments about the presence or absence of elements such as thesis statements, supporting ideas, or conclusions.

Few studies have been conducted on classroom use of AWE programs. One interesting study gives a detailed account of how Criterion was used by 6th to 12th graders throughout the United States during the 2002–2003 school year, based on analysis of 33,171 student essay submissions of 50 or more words (Attali, 2004). The study found that a strong majority of the student essays (71%) had been submitted only one time without revision, suggesting that the program is not being used in classrooms in ways that it is touted (i.e., as a motivator and guide for more revision of writing by students). For essays submitted more than once, computerized scores rose gradually from first to last submission (from 3.7 to 4.2 on a 6-point scale), but revisions conducted were almost always in spelling and grammar rather than in organization.

A second study attempted to investigate the impact of using Criterion on student's writing development (Shermis, Burstein, & Bliss, 2004). In this study, 1,072 urban high school students were randomly assigned to either a treatment group, which wrote on up to seven Criterion writing prompts, or a control group, which participated in the same classes but completed alternate writing assignments without using Criterion. No significant differences were noted between the two groups on a state writing exam at the end of the training. The authors attributed this at least in part to poor implementation and high attrition, with only 112 of the 537 treatment students completing all seven essays. The researchers calculated that if students had written five more writing assignments each, differences in performance would have been significant. However, such predictions are moot if the reasons the software was underused are not understood or easy to address.

Neither of these studies conducted any observations or interviews to analyze in situ how AWE software is being used in the classroom. Our study thus sought to make a contribution in that area, by examining firsthand the ways that teachers and students make use of AWE programs.

Our theoretical framework could be called "socio-technical constructivism." It is social-constructivist in that it sees knowledge as constructed by learners in terms of their own experiences and thinking patterns, which are shaped by personal social settings. It is socio-technical in holding that technologies must be understood in the context of users' prior beliefs, practices, and social-institutional

settings (Kling, 1999; Orlikowski & Robey, 1991). Research indicating that humans apply social rules to computers illuminates the psycho-social mechanisms through which computers become embedded in the socio-technical fabric of our daily lives (Nass & Moone, 2000; Reeves & Nass, 2003).

From this theoretical vantage point, we predicted that teachers would vary widely in their adoption of the new technology, and even those who embrace it would adapt its use to their prior instructional practices and current institutional pressures, especially the need to prepare for standardized tests. Following Nass and Moone (2000), we predicted that students would treat automated scores and automated feedback with much of the deference they give teachers, albeit with less fear of humiliation and with more skepticism about the authoritativeness of the automated responses.

METHOD

In 2004–2005, we conducted a mixed-methods exploratory case study to learn how AWE is used in classrooms and how that usage varies by school and social context. We studied AWE use in a convenience sample of a middle school (Grades 6–8), two junior high schools (Grades 7–8), and one high school (Grades 9–12) in Southern California that were deploying AWE software programs (see Table 1). Two of the four schools used Criterion and two used My Access (the third program referred to earlier, WriteToLearn, was not released until September 2006, after data collection for this study was completed). The student populations in the four schools varied widely in academic achievement, socioeconomic status (SES), ethnic makeup, and access to computers. The two junior high schools were part of a larger one-to-one laptop study reported elsewhere in this special

TABLE 1
Schools in the Study

School	*Software Used*	*Computer Configuration*	*SES*	*Predominate Ethnic Group*	*Academic Performance Index*	*Length of Time Using AWE*
Flower Junior High	My Access	One-to-one laptops	High	Asian	High	1st year
Nancy Junior High	My Access	One-to-one laptops	Low	Latino	Low	1st year
Timmons Middle	Criterion	Computer lab	High	White	High	At least 3 years
Walker High	Criterion	Computer lab	High	White	High	3 years

issue (hence the greater amount of data from these two schools), where all students in certain grades had personal laptops, and My Access was available to language arts teachers. The middle school and high school used Criterion in computer labs. One of the one-to-one schools was a new, high-SES technology-oriented school with mostly White and Asian students. The other was an older, low-SES school with two thirds Latino students.

At Flower and Nancy Junior High Schools, all teachers who were available were included in the study; this included all but one of the junior high language arts teachers in the two schools. At each of the other two schools, one language arts teacher who regularly used the AWE program was recommended by a senior administrator to participate.

Sources of data included transcribed semistructured interviews with three principals, eight language arts teachers, and two focus groups of students; observations of 30 language arts classes; a survey completed by seven teachers and 485 students using My Access at Flower and Nancy schools; reports of 2,400 essays written with My Access; and in-depth examination of ten essays submitted to My Access and revised after the students received feedback from the program.

Interview data and observations notes were analyzed using standardized qualitative coding and pattern identification techniques, assisted by the qualitative data analysis software HyperResearch. Survey data was analyzed using descriptive statistics.

In analyzing the findings, we first review some overall patterns of use across the four schools. Then, by examining similarities and differences between the three middle schools/junior highs, we consider how program usage related to social context.

Contradictory Patterns of Use

In examining how AWE was used across the four schools, we noticed two dominant paradoxical findings. First, teachers and students valued the AWE programs, yet they were seldom used in the classrooms. Second, the programs apparently did contribute to greater student revision, yet almost all the revisions were superficial.

Positive Opinions Versus Limited Use

All of the teachers and administrators we interviewed expressed favorable overall views of their AWE programs. Several talked glowingly about students' increased motivation to write. All seven teachers who responded to our survey on My Access said they would recommend the program to other teachers, and

six of seven said they thought the program helped students develop insightful, creative writing. Students also indicated positive assessments of the program in both surveys and focus-group interviews.

A major advantage of automated writing evaluation, reported by teachers and confirmed by our observations, was that it engaged students in autonomous activity while freeing up teacher time. Instead of sitting idly at the end of a writing session, faster writers were engaged in revising and resubmitting for higher scores while slower writers continued to work on their first draft. Teachers still graded essays, but they were able to be more selective about the essays and parts of essays they chose to grade. In many cases, teachers allowed students to submit early drafts for automated computer scoring and a final draft for teacher evaluation and feedback. One teacher compared My Access to "a second pair of eyes" to watch over a classroom.

In spite of teachers' positive attitude toward My Access, they used the program infrequently. Seventh-grade students in the two one-to-one laptop schools each averaged only 2.3 essays using AWE software between November 2004 and May 2005. Limited usage seemed to be due to two main factors. First, teachers at the schools felt a great deal of pressure to cover as much curriculum as possible in order to prepare students for state examinations. Much of this curriculum was in reading or language arts rather than in composition, limiting the time available for writing instruction.

Second, the programs could only support essays written to specific prompts that come with the program. When teachers wanted students to engage in other types of writing, such as newspaper articles, brochures, or business letters, or wanted students to write essays on topics for which there were no presupplied prompts, they did not use the program.

At the two schools using Criterion, usage was greater. However, at each of those two schools, we included only one teacher in the study who had been recommended due to extensive usage of the program. It appears that other teachers at the same two schools used the program much less frequently, if at all.

The limited usage of AWE software in these four schools reconfirms a prior study showing similar results (Shermis et al., 2004, discussed previously) and raises questions about the current viability of AWE software. Interestingly, another type of software that has been touted as a magic bullet for improving test scores, programmed reading instruction, has suffered from similar patterns of limited implementations and correspondingly null effects on student performance (Kulik, 2003; Slayton & Llosa, 2002). No matter how much teachers claim that they like a type of software (responding, perhaps, to the expectation from administrators and the public that they *should* like it), if they find various reasons not to use the software, it cannot be expected to have much impact.

Emphasis on Revision Versus Limited Revision

The teachers and administrators we interviewed were unanimous in recommending the AWE for promoting student revision. As one teacher told us, "I feel that [the program] puts the emphasis on revision. It is so wonderful to be able to have students revise and immediately find out if they improved." An administrator anecdotally spoke of a student revising a paper 17 times in order to improve the score.

Yet, our data suggest that students usually submitted their papers for scores only one time, not 17, and almost all of the revisions that students made were narrow in scope. Supporting Attali's (2004) findings discussed earlier, 72% of the student essays in our sample were submitted for a score only one time, and the majority of the rest were resubmitted just once. Of course this in itself does not fully indicate how much revision was done, as students could revise their papers, making use of editorial tools prior to submitting for a score the first time. And indeed, teachers and administrators we interviewed, including the high school principal who had monitored use of Criterion over several years, told us that students did revise their papers more in anticipation of getting a score. This limited resubmission does seem to indicate, though, that although students paid close attention to scores (sometimes shouting with glee when they got high ones), they either weren't especially concerned with working extra hard to raise them or were not provided the time to do so.

More important, almost all the revisions addressed mechanics, not content or style. In our observations, virtually all the revisions we saw students making were of spelling, word choice or grammar, not content or organization. To confirm this, we reviewed 10 randomly chosen essays that were submitted two or more times to observe the changes between first and last draft. None had been revised for content or organization. Except for one essay in which a sentence was added (repeating what had already been said), all of the revisions maintained the previous content and sentence structure. Changes were limited to single words and simple phrases, and the original meaning remained intact. Most changes appeared to be in response to the automated error feedback.

This limited revision is consistent with more general practices in U.S. public schools, in which student rewriting invariably focuses on a quick correction of errors pointed out by the teacher or peer. (In contrast, when students write for authentic audiences, evidence suggests they more readily revise for content; see Butler-Nalin, 1984). Using AWE programs, students recognized that the easiest way to raise their scores was through a series of minor corrections; few took the time to even read through the more general narrative feedback provided regarding ways to improve content and organization, and those who did either failed to understand it or failed to act upon it.

Differences Among Schools

This analysis considers overall patterns at the four schools. We also compared usage between the different schools to understand how teacher belief and social context affected use of the program. We found major differences, which we illustrate through portrayals of a seventh-grade teacher at each of three middle school/junior highs. The teachers at the first two schools had attended the same or similar 1-day training sessions in My Access. The teacher at the third school had more limited formal training in Criterion.

Nancy junior high. Nancy Junior High was two thirds Latino and had slightly below average Academic Performance Index scores compared with similar schools in California. Almost 60% of the students were on free or reduced lunch programs. Nancy had started a one-to-one laptop program that year and had purchased AWE software as part of the program.

Ms. Patterson, the teacher we observed most frequently at Nancy and whose use of AWE seemed consistent with that of most other teachers at the school, taught both English-as-a-Second-Language (ESL) classrooms and regular classrooms. However, even in Ms. Patterson's regular classes, students were performing below grade level in reading, writing, and language arts. According to Ms. Patterson, many of her students had never written a whole paragraph before they entered seventh grade. Most also had limited typing skills.

Ms. Patterson attempted to integrate My Access in a process-oriented writing program. However, students in her class worked very slowly due to limited reading and writing ability and typing skills. In addition, Ms. Patterson explained that there was no strong tradition of doing homework at the school and that she thus kept written homework tot a minimum to avoid failing too many students. As a result of these challenges, Ms. Patterson was not able to use AWE much during the year. She constantly felt the need to focus on broad coverage of the curriculum and thus had insufficient time for writing instruction. When students did make use of AWE, their ability to understand the program's feedback, other than its most basic aspects, was weak, and Ms. Patterson didn't bother to try explaining it as students would have little time to revise in any case. Students appeared motivated by the scores but showed no indication of using either the scores or the feedback to improve their writing, other than for correction of spelling errors.

Like many teachers at her school, Ms. Patterson began the year with relatively little experience with computers and thus approached the school's new laptop program with some trepidation. As the year wore on, though, she became enthusiastic about aspects of laptop use in her classroom, especially the authentic writing and production that students carried out. Ms. Patterson beamed when speaking about her students' use of laptops to produce a literary newspaper or

to produce a movie trailer about a book they had read. Her students also showed great excitement when working on those assignments. We witnessed a much lower level of enthusiasm among both students and teacher with regard to the use of AWE.

Flower junior high. Ms. Samuels was the only junior high English language arts teacher at Flower, a high-SES school of mostly Asian and White students across town from Nancy. Flower, like Nancy, had started a laptop program that year and had begun use of AWE software in the context of that program. However, other aspects of the context were quite different. Students at the school were tech-savvy. Teachers, including Ms. Samuels, had been recruited for Flower based on previous teaching success and enthusiasm for technology.

Like Ms. Patterson, Ms. Samuels sought to integrate My Access into a holistic process-oriented approach, but her conditions were much better for doing so. Her students' higher level of language, literacy, and computer skills allowed Ms. Samuels to devote more attention to teaching them to understand the program's feedback. After some experience with the program she also began to encourage students to work on their essays at home, and it appeared that a number of them did so. She also organized substantially more peer collaboration than the other teachers.

In spite of Ms. Samuels' desire to make more extensive use of AWE, the results achieved were not noticeably different from those of Ms. Patterson. Students in her classes submitted their papers no more frequently than students at Nancy; nor did they carry out more revisions. Perhaps the overall climate of instruction in the district and state, which emphasized a rapid gain in measurable test scores, accounted for this. Or perhaps the limited time of 6 months that the students had used the program did not allow them to fully master it.

Contrary to our expectations, Ms. Samuels' enthusiasm for AWE seemed to wane over time, and she used the program slightly less the following year, due in part to the fact that her students had already completed a number of the relevant prompts. In addition, like Ms. Patterson, Ms. Samuels was much more excited about other ways of using laptops that involved more meaningful, authentic communication, such as multimedia interpretations of literature in a literary pageant.

Timmons middle school. Timmons Middle School, in a nearby district of Nancy and Flower, served predominately high-SES students. Ms. Tierney was nominated by her administrator as a teacher who had extensive experience using AWE. Whereas the two previously mentioned teachers attempted to integrate AWE into a process-oriented writing program, with students drafting and sometimes revising a paper over 1 to 2 weeks, Ms. Tierney used the

program in an entirely different fashion—as an explicit form of test preparation. Ms. Tierney scheduled 1 day per week in the school's computer lab and had students compose in Microsoft Word and Criterion on alternate weeks. Each week, she gave them 10 minutes to choose from a variety of pencil-and-paper prewriting techniques, then 30 minutes for a timed writing exam. Ms. Tierney explained that she had been giving weekly timed writing tests even before she used the AWE, and she now continued those tests as before. The only difference was that now, by using AWE every other week, she could save several hours with a cursory grading of papers that week, grading more thoroughly for papers on the alternate weeks when she didn't use AWE.

Although familiar with the process-writing approach, Ms. Tierney used a very structured, traditional approach for the writing of a formal five-paragraph essay on the days her class was at the computer lab. And in accord with Hillocks's (1986) finding that "teacher comment has little impact on student writing" (p. 165), she discounted the value of teacher comments for writing instruction. As she explained,

> I have 157 students and I need a break … I cannot get through these essays and give the kids feedback. One of the things I've learned over the years is it doesn't matter if they get a lot of feedback from me, they just need to write. The more they write, the better they do.

As Ms. Tierney's students did not have personal laptops, she could not use AWE in the same integrative way as Ms. Patterson and Ms. Samuels even if she had wanted to. Yet, we had the sense that the differences in instructional approach were only partly due to the differential access to technology. Rather, each of the three teachers moulded the use of AWE to their own particular belief systems, with Ms. Patterson and Ms. Samuels favoring a process approach and Ms. Timmons favoring teaching to the standardized writing tests.

DISCUSSION

The companies that produce and market AWE programs for classroom use make two principal claims: first, that the programs will save teachers grading time, thus allowing them to assign more writing; and second, that the scores and feedback will motivate students to revise their papers more, thus encouraging a more iterative writing process. In the schools we investigated, it appears that each of both claims was partially true. All the teachers we interviewed and observed indicated that the program helps to save them time, whether outside of class (when they let the AWE program handle part of their grading) or

inside of class (when students work more independently with the AWE program, allowing the teacher to provide more attention to individual students). Yet we saw little evidence that students used AWE to write substantially more than they had previously. In most cases, the main factor limiting how much writing teachers assigned was not *their* time available to *grade* papers but rather *students'* time available to *write* papers, and that was not increased by the use of AWE. Insufficient number of relevant prompts also limited how much teachers could use AWE for graded writing practice.

As for the second claim, we observed most students revising their papers in response to editorial feedback from the programs, and nearly half the students we surveyed agreed that they edited their papers more when using an AWE program. Yet, almost all the revisions made were at the word or sentence level, and we witnessed none of the broadly iterative process through which writers hone their content, sharpen their organization, and thus learn to transition from writer-based to reader-based prose (see Flower, 1984). In addition, nearly three quarters of the time, students submitted their essays for scores only once rather than revising and resubmitting for an improved score.

At the same time, the negative effects that critics have pointed to were not noted. The software did not replace teachers but rather freed up teachers' time for other activities. The software did not distort the way that teachers taught writing. As exemplified in the aforementioned three cases, all the teachers in our study continued to teach writing very similarly to how they had previously and integrated the AWE software into that approach. Finally, we did not see much evidence that the AWE software promoted stilted writing. There were a few minor examples of students turning away from a more appropriate colloquial expression for a more bland standard form because the error feedback discouraged colloquialisms. However, for the most part the feedback provided by the programs was either beneficial or benign. The main cause of stilted writing was not the AWE software programs themselves but rather the broader standards and high-stakes testing regime in public schools that encourage teachers to focus narrowly on five-paragraph essays.

Our findings on the whole confirmed our predictions for both students and teachers. For example, differences were noted from site to site, with high-SES students more readily able to fully use the program due to better keyboard skills, better computer and Internet access at home, and a stronger language and literacy background. However, the most important difference across sites was not due to student SES but rather teachers' habits and beliefs, with instructors using the program very differently, depending on whether or not they adopted a process approach to writing. Interestingly, the teacher who cared least about student revision actually was able to use an AWE program most; because she was teaching for general test-taking ability rather than focusing on particular academic content, it was of little concern to her whether the prompts she chose

matched her current curriculum, and thus a wider number of prompts were available.

Although most teachers expressed positive attitudes toward AWE in the survey and interviews, the wide range of usage levels among them attests to their widely varying levels of adoption, as predicted.

In summary, AWE, like other technologies, is neither a miracle nor monster; rather, it is a tool whose influence is mediated by complex relationships among social and educational contexts, teacher and student beliefs, other technologies, and prior instructional practices. In particular, the implementation of AWE in California is shaped by a strong emphasis on raising test scores and teaching the five-paragraph essay to meet standards; highly diverse student populations by language, literacy, SES, computer experience and social capital; differing teacher goals, beliefs and backgrounds; and the present state of hardware (e.g., whether a school had one-to-one laptops) and software (e.g., vendors' current repertoires of included prompts).

CONCLUSION

Although decades of research and millions of dollars have been invested into developing automated scoring technologies, current available programs remain relatively error-prone and insensitive to individual learners' skills and needs. The possible utility of machine scoring, in spite of its flaws, can be understood in light of Brian Huot's (1996) insight that assessment is never context-free; the purpose of an assessment is as essential as the text to be assessed. Although scoring engines' biases and flaws loom large in high-stakes placement exams, they appear to have little or no negative impact if used as preprocessors for teachers in low-stakes settings that also engage students in ample nonformulaic writing for real human audiences.

Automated assessment will neither destroy nor rescue writing instruction. Currently, it seems most helpful for motivating and assisting young writers to reduce the number of mechanical errors, thereby freeing the teacher to focus on content and style. The potential benefits may expand as AWE software becomes more capable, computer hardware becomes more prevalent in schools, and teachers and students become more comfortable with technology. If the 50-year history of AI-based text processing is any indication of its future, the need for sensitive human readers will not disappear, no matter how closely automated scores approximate scores by expert human graders. We expect that the software will continue to have substantial blind spots, including insensitivity to connotation and context, and that, like many technologies, it will be used in different ways in different settings. Further research is needed on the factors affecting teachers' use of AWE and on the learning processes and outcomes that

result when it is used in particular ways with particular groups of students (see Warschauer & Ware, 2006, for a more detailed proposed research agenda).

As has been the case with other educational technologies, both the techno-optimists and the techno-pessimists have overstated their cases, each side taking a deterministic view of technology's "impact" and failing to appreciate the complex interactions among social-institutional-technological contexts, individuals' goals and backgrounds, and prior instructional practices that shape the role of new educational technologies. Teachers need not stand in "awe" of automated writing evaluation's alleged benefits or shortcomings; rather, they can critically evaluate whether and how to deploy it to best meet their and their students' needs.

REFERENCES

Attali, Y. (2004, April). *Exploring the feedback and revision features of Criterion.* Paper presented at the meeting of the National Council on Measurement in Education (NCME), San Diego, CA.

Attali, Y., & Burstein, J. (2004, June). *Automated essay scoring with e-rater V.2.0.* Paper presented at the Conference of the International Association for Educational Assessment, Philadelphia, PA.

Baron, D. (1998, November 20). When professors get A's and the machines get F's. *Chronicle of Higher Education, A56*.

Burstein, J. (2003). The e-rater scoring engine: Automated essay scoring with natural language processing. In M. D. Shermis & J. Burstein (Eds.), *Automated essay scoring: A cross-disciplinary perspective* (pp. 113–121). Mahwah, NJ: Erlbaum.

Butler-Nalin, K. (1984). Revising patterns in students' writing. In A. N. Applebee (Ed.), *Contexts for learning to write* (pp. 121–133): Ablex.

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

Chodorow, M., & Burstein, J. C. (2004). *Beyond essay length: Evaluating e-rater's performance on TOEFL essays*. Princeton, NJ: Educational Testing Service.

Cohen, Y., Ben-Simon, A., & Hovav, M. (2003, October). *The effect of specific language features on the complexity of systems for automated essay scoring.* Paper presented at the 29th Annual Conference of the International Association for Educational Assessment, Manchester, UK.

Conference on College Composition and Communication. (2004). *CCCC position statement on teaching, learning, and assessing writing in digital environments*. Retrieved September 21, 2006, from http://www.ncte.org/cccc/resources/positions/123773.htm

Elliot, S. M., & Mikulas, C. (2004, April). *The impact of MY Access!™ use on student writing performance: A technology overview and four studies.* Paper presented at the Annual Meeting of the American Educational Research Association, San Diego, CA.

Ericsson, P. F. (2006). The meaning of meaning. In P. F. Ericsson & R. Haswell (Eds.), *Machine scoring of human essays: Truth or consequences* (pp. 28–37). Logan: Utah State University Press.

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

Flower, L. (1984). Writer-based prose: A cognitive basis for problems in writing. In S. McKay (Ed.), *Composing in a second language* (pp. 16–42). New York: Newbury House.

Hillocks, G. J. (1986). *Research on written composition*. Urbana, IL:ERIC Clearinghouse on Reading and Communication Skills and NCTE.

Huot, B. (1996). Computers and assessment: Understanding two technologies. *Computers and Composition*, *13*(2), 231–243.

Keith, T. Z. (2003). Validity and automated essay scoring systems. In M. D. Shermis, & J. C. Burstein (Eds.), *Automated essay scoring: A cross-disciplinary perspective. (pp*. 147–167). Mahwah, NJ: Erlbaum.

Kling, R. (1999). What is social informatics and why does it matter? *D-Lib Magazine*, *5*(1). Retrieved June 14, 2007, from http://www.dlib.org/dlib/january99/kling/01kling.html

Kulik, J. A. (2003). *Effects of using instructional technology in elementary and secondary schools: What controlled evaluation studies say*. Arlington, VA: SRI.

Landauer, T. K., Laham, D., & Foltz, P. (2003). Automated scoring and annotation of essays with the Intelligent Essay Assessor. In M. D. Shermis & J. Burstein (Eds.), *Automated essay scoring: A cross-disciplinary perspective* (pp. 87–112). Mahwah, NJ: Erlbaum.

Nass, C., & Moone, Y. (2000). Machines and mindlessness: Social responses to computers. *Journal of Social Issues*, *56*(1), 81–103.

National Commission on Writing in America's Schools and Colleges. (2003). *The neglected "r": The need for a writing revolution*. New York: The College Entrance Examination Board.

Orlikowski, W., & Robey, D. (1991). Information technology and the structuring of organizations. *Information Systems Research*, *2*(2), 143–169.

Powers, D. E., Burstein, J. C., Chodorow, M., Fowles, M. E., & Kukich, K. (2002). Stumping e-rater: Challenging the validity of automated essay scoring. *Computers in Human Behavior*, *18*, 103–134.

Reeves, B., & Nass, C. (2003). *The media equation: How people treat computers, television, and new media like real people and places*. California: Center for the Study of Language and Information (CSLI), Stanford, CA: Stanford University.

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

Shermis, M. D., Burstein, J. C., & Bliss, L. (2004, April). *The impact of automated essay scoring on high stakes writing assessments.* Paper presented at the annual meeting of the National Council on Measurement in Education, San Diego, CA.

Shermis, M. D., Mzumara, H. R., Olson, J., & Harrington S. (2001). On-line grading of student essays: PEG goes on the World Wide Web. *Assessment & Evaluation in Higher Education*, *26*(3), 247–259.

Slayton, J., & Llosa, L. (2002). *Evaluation of the Waterford Early Reading Program 2001–2002: Implementation and student achievement*. Retrieved September 21, 2006, from http://notebook.lausd.net/pls/ptl/url/ITEM/EF8A0388690A90E4E0330A081FB590E4

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