J·T·L·A

The Journal of Technology, Learning, and Assessment

Volume 8, Number 6 · March 2010

Utility in a Fallible Tool: A Multi-Site Case Study of Automated Writing Evaluation

Douglas Grimes & Mark Warschauer

www.jtla.org

A publication of the Technology and Assessment Study Collaborative
Caroline A. & Peter S. Lynch School of Education, Boston College

J·T·L·A

Volume 8, Number 6

Utility in a Fallible Tool: A Multi-Site Case Study of Automated Writing Evaluation

Douglas Grimes & Mark Warschauer

Editor: Michael Russell
russelmh@bc.edu
Technology and Assessment Study Collaborative
Lynch School of Education, Boston College
Chestnut Hill, MA 02467

Copy Editor: Jennifer Higgins
Design: Thomas Hoffmann
Layout: Aimee Levy

JTLA is a free online journal, published by the Technology and Assessment Study Collaborative, Caroline A. & Peter S. Lynch School of Education, Boston College.

Preferred citation:

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 [date] from http://www.jtla.org.

J·T·L·A

Abstract:

Automated writing evaluation (AWE) software uses artificial intelligence (AI) to score student essays and support revision. We studied how an AWE program called MY Access!® was used in eight middle schools in Southern California over a three-year period. Although many teachers and students considered automated scoring unreliable, and teachers' use of AWE was limited by the desire to use conventional writing methods, use of the software still brought important benefits. Observations, interviews, and a survey indicated that using AWE simplified classroom management and increased students' motivation to write and revise.

J·T·L·A

Utility in a Fallible Tool: A Multi-Site Case Study of Automated Writing Evaluation

Douglas Grimes
Mark Warschauer
University of California, Irvine

Introduction

> "Every time he stands up in class and assigns a paper he sees in his mind's eye that stack of papers on the corner of his desk waiting for him to grade … he feels guilty because he knows [writing every few weeks] doesn't give his students enough practice and it means that his comment and advice on a student's paper this time will probably have no useful effect at all on what the student writes next time."
> (Elbow, 1981, p. 218)

The demand for writing skills intensifies with our increasing dependence on global information systems. Although there has long been widespread agreement on the need for development of writing skills (see, for example, National Commission on Writing, 2003, p. 3), students' writing practice has been limited by a bottleneck in the flow of communication between teacher and students: teachers simply do not have the time to respond quickly and thoughtfully to student writing assignments. As a result, English language arts (ELA) teachers either assign little writing practice or take on a heavier grading load than their peers in other subject areas (Breland, 1996, pp. 12–13). Teachers we interviewed said that without some form of essay grading assistance, grading delays of a week or more were commonplace, by which time most students had lost interest in the teachers' comments.

Automated writing evaluation (AWE) software has been promoted as a way to remove the bottleneck, primarily because students receive scores and formative feedback very quickly (often within a few seconds of essay submission). The distinguishing features of an AWE system are automated essay scoring and formative feedback to assist the writing process. The two most established AWE systems in the K–12 market are *Criterion*®, from a subsidiary of the Educational Testing Service, and MY Access! ("MA" or "My Access" below), from Vantage Learning, Inc.

Proponents claim that AWE facilitates more writing practice and increases students' motivation to write and revise (Burstein, Chodorow, & Leacock, 2004; Elliot & Mikulas, 2004; Foltz, Laham, & Landauer, 1999). Critics claim that it diminishes the role of teachers and warps students' notions of good writing (Baron, 1998; CCCC Executive Committee, 2004; Cheville, 2004; Ericsson & Haswell, 2006). In order to examine these claims in the context of extended classroom use of AWE, we employed a mixed-method embedded case study (Yin, 1989, p. 49), in eight middle schools, four in each of two Southern California school districts.

Prototypical classroom use of AWE follows an old and familiar cycle for essay assignments in general: the teacher assigns a topic; students write and submit their essays; the teacher grades and comments on essays, and returns them to students. (Students sometimes submit multiple drafts for the teacher's review.) A similar sequence applies with AWE, with one new step and alteration of others:

1. The teacher assigns a topic from a list provided by the vendor. Occasionally teachers write their own essay topics, in which case scoring is less accurate.
2. Students write and submit their essays via the Web, at which point they are visible to the teacher.
3. (New step) The software scores the essay and provides feedback for revision. The previous step and this step may be repeated multiple times, sometimes with oral or written comments from the teacher in addition to automated feedback.
4. Teacher optionally grades the essays and adds comments, which students can see via the Web.

AWE Technology

AWE systems are commonly described as consisting of their two most critical components, a scoring engine and a feedback engine based on a branch of artificial intelligence called natural language processing. MyAccess uses the *IntelliMetric*® scoring engine and feedback from tools called "MY Tutor" and "MY Editor." *Criterion* uses the *e-rater*® scoring engine to score essays and provide diagnostic feedback. We distinguish AWE from its scoring technology, usually called "automated essay scoring" (AES) or "automated essay grading" (AEG) because they serve different purposes. AES is best known in the context of high-stakes testing; AWE is used for lower-stakes writing instruction. AWE systems also contain two other sets of software tools that generally do not use artificial intelligence: a limited form of a learning management system (LMS) and a limited form of an online writing lab (OWL). With the LMS features teachers can manage writing assignments, students can review their writing portfolios, and district administrators can track progress by reports on writing by teacher, student, grade, school, or other criteria. OWL features are ancillary writing aids such as an online dictionary, graphic organizers, and writing rubrics with sample essays.

The first AES system, Page Essay Grade (PEG), was developed by a former high school English teacher, Ellis Page, who used multiple regression to associate surface text features in a target essay to those in a corpus of essays on the same topic that had been scored by English teachers (Page, 1967). Heavily weighted variables included number of words, average sentence length, standard deviation of word length, average word length, and number of commas. As Page recognized, such variables, taken separately, are at best crude approximations to features used by human raters in assessing writing quality. He and other AES developers then introduced estimates of other factors meaningful to human raters, such as grammatical correctness, and word choice (Attali & Burstein, 2006; Page, 1967). Techniques included semantic models to compare vocabularies between the target essay and a corpus of essays scored by trained human graders, a major topic of technical articles on automated scoring (Attali & Burstein, 2006; Ben-Simon & Bennett, 2007; Elliot, 2003; Shermis & Burstein, 2003).

AES developers also grouped the variables into categories (domains or features) which provide a conceptual framework for analytic scores and formative feedback. In MA these are focus, organization, development, language use, and mechanics and conventions. *Criterion* gives scores, feedback, and comments on four analytic categories related to errors (grammar, usage, mechanics, and style) and one higher-level category, organization and development (Attali & Burstein, 2006). All feedback in *Criterion* is visually keyed to specific sections of text (Attali & Burstein, 2006; Burstein et al., 2004, p. 9). In MA the lower-level feedback is keyed to specific words or phrases, but higher-level feedback is generic within each genre and score point; every eighth-grade persuasive genre essay that scores a 3 in organization gets the same generic advice. Some of the marketing material on MA strongly implies human-like intelligence in the IntelliMetric scoring engine, e.g.:

> IntelliMetric® is an intelligent scoring system that emulates the process carried out by human scorers … . The system must be "trained" with a set of previously scored responses containing "known score" marker papers for each score point. These papers are used as a basis for the system to infer the rubric and the pooled judgments of the human scorers. Relying on Vantage Learning's proprietary CogniSearch™ and Quantum Reasoning™ Technologies,

> the IntelliMetric® system internalizes the characteristics of the responses associated with each score point and applies this intelligence in subsequent scoring. … IntelliMetric® is trained to score essays much the same way as expert human raters are trained. (Vantage Learning, 2007)

Prior Research on Automated Essay Scoring

An extensive body of research compares the reliability of human and automated essay scoring (Cohen, Ben-Simon, & Hovav, 2003; Elliot, 2003; Landauer, Laham, & Foltz, 2003; Page, 1994; Shermis & Burstein, 2003). Most of the extant literature, much of it vendor-sponsored, has found the major scoring engines roughly equivalent to human graders in reliability. However, a recent study of placement essays written by mostly Hispanic college-age developmental writers in South Texas found that the mean scores from *IntelliMetric*, the scoring engine in My Access, were significantly higher than those from two faculty members (Wang & Brown, 2007). *Intellimetric* assigned failing scores to only 2.8% of the students, compared with 27.1% by the human graders, potentially leading to many students being assigned to courses for which they were unqualified. The authors conclude that more studies are needed on the generalizability of *Intellimetric* scores across student populations (Wang & Brown, 2007).

A number of authors have pointed to the need for other criteria besides reliability, including validity and the distorted social message of machine scoring (Chen & Cheng, 2008; Chung & Baker, 2003; Keith, 2003; Kelly, 2001; Phillips, 2007). A position statement of the Conference on College Composition and Communication (CCCC) in the U.S. states:

> Writing-to-a-machine violates the essentially social nature of writing: we write to others for social purposes. If a student's first writing-experience at an institution is writing to a machine, for instance, this sends a message: writing at this institution is not valued as human communication—and this in turn reduces the validity of the assessment. (CCCC Executive Committee, 2004)

Prior Research on Classroom Use of AWE

Online AWE programs have been used in classrooms for a decade, and the vendor of My Access claims over a million users (Vantage Learning, 2006). In spite of the apparent popularity of AWE and the recent entry of several new players, few in-depth studies have been conducted on classroom use of AWE. Two quantitative studies on *Criterion* are pertinent here.

The first addressed *Criterion* use by sixth to twelfth graders throughout the United States during the 2002–2003 school year, based on an analysis of 33,171 student essay submissions of 50 or more words (Attali, 2004). A large majority of the essays (71%) were submitted only one time without revision, suggesting that the program's potential to encourage revising was underused. Moreover, the revisions were overwhelmingly mechanical, primarily 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). Over a thousand urban high school students were randomly assigned to either a treatment group using *Criterion*, or a control group in the same classes, which 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. However, the group using *Criterion* showed a significant increase in average essay length and a reduction in the number of errors, especially errors in writing mechanics (spelling, capitalization, punctuation, and grammar).

Several qualitative studies also merit discussion. In three previous reports based on middle school and high school use of My Access and *Criterion*, including early data from Farrington District (described below), the current authors found that students were more motivated to write and revise, and classroom management was easier for teachers (Grimes, 2008; Grimes & Warschauer, 2008b; Warschauer & Grimes, 2008). For example, in one article they noted three paradoxes in using AWE. First, teachers and students recommended AWE, even though they saw the scoring and feedback as flawed. Second, in spite of those favorable opinions towards AWE, there was little increase in the amount of writing practice. Third, although teachers reported that AWE encouraged revision, they scheduled little time for revision.

A study of writing in a regional university in the southeastern United States found mostly positive student attitudes towards *Criterion* (Schroeder, Grohe, & Pogue, 2008). Forty-nine criminal justice majors, mostly African Americans, completed an exit survey on *Criterion*. Overall, they reported that *Criterion* helped them understand their writing errors and made it easier to improve papers before turning them in. Forty-one said they would recommend requiring *Criterion* in the writing class in their major, compared to only three who said they would recommend not using it.

Three other qualitative studies deserve mention not because they illustrate the potential of AWE, but because, when taken together with successful AWE implementations, they reveal the lackluster results that can be expected with either an inadequate AWE system or an inadequate implementation. Two of these are overlapping case studies which addressed the use of an online learning environment called ETIPS in a large Midwestern university (Riedel, Dexter, Scharber, & Doering, 2005). The same three teachers were involved in both studies, conducted with post-baccalaureate education students. Unlike Criterion and MA, both of which allow teachers to choose between a four-point or six-point scoring scale, the ETIPS AWE system uses a three-point scale. It also uses a different sort of semantic model (Bayesian Analysis) to compare target essays to a corpus of human-scored essays. Students in the first study felt that the three-point ETIPS scale was inadequate to provide useful guidance for revision. Students in the second study felt that the automated scores were only slightly helpful in composing their essays, and they were only a little confident of the accuracy of the scores. However, they displayed strong emotions about their experiences with automated scoring (Scharber, Dexter, & Riedel, 2008). The authors concluded from a review of writing research that for automated feedback to be useful, it "must provide recommendations for improvement and explain why the essay needs improvement" (Riedel et al., 2005, p. 270).

In a study of three university-level writing classes on English as a Foreign Language in Taiwan, one of the three teachers abandoned the program after two assignments due to perceived flaws in the scoring, and the other two teachers used the program throughout the semester (Chen & Cheng, 2008). None of the 53 students who responded to a survey agreed with the statement "The scores in My Access are adequate." Only 12 students (23%) agreed with the statement "The written feedback given by My Access is helpful for revision." They regarded the program most favorably when the teacher required that they achieve a minimum score of 4 out of 6 before submitting to her, a practice that was seen as a confidence-builder. A number of students suggested that the program would be more helpful for students in the early stages of learning English. It is important to note that the teachers in this study received only one hour of instruction in My Access. In contrast, most teachers in our current study received at least a day of training in My Access.

The current study seeks to add to this literature by presenting one of the first in-depth, multi-site case studies of AWE use in classrooms. The study focuses on a previously overlooked student population, middle schools (grades seven and eight), a prime market for AWE vendors. Whereas prior qualitative studies have addressed only new implementations, this is the first field study to follow an AWE implementation from its introduction to the point where it is an accepted part of the status quo for writing instruction in a district.

This study also takes a different theoretical perspective toward AWE, that of social informatics, which erases analytical separations between technologies, people, and organizations, and considers them as a "heterogeneous socio-technical network" in which none of the three can be understood without the other two (Kling, 1999). Given the variety and mutability of social relationship and technologies, this view predicts a more complex and locally situated process of technology diffusion than predicted by a "tools" view that focuses only on the technology per se. Kling argued that questions such as "What will be the impact of technology X?" imply a technological determinism that consistently underestimates the importance of people and organizations in shaping the use of technology (Kling, 2000). "The analytical failure of technological determinism is one of the interesting and durable findings from social informatics research" (Kling, 1999). Social informatics beckons us to inquire into local non-technological factors to explain why each of major the AWE systems has been well received in some contexts and rejected in others.

Research Sites

This study primarily covers the first three years of large-scale AWE use in the middle schools of two school districts, "Farrington" and "Sunrise," between Los Angeles and San Diego, California. (Names have been changed for anonymity.) Both districts had four middle schools that used My Access as part of larger efforts to improve students' writing skills; both financed their AWE projects primarily through large, highly competitive grants; and both began large-scale implementation of AWE in the 2004–2005 school year. Our data collection took place in year 1 (2004–2005 school year) and year 3 (2006–2007 school year).

Farrington had approximately 3,100 middle school students. In two of their four middle schools all students participated in a one-to-one laptop program, which gave them easier access to Internet-connected computers than other students in the district. Sunrise District had over 5,600 middle school students.

Farrington District had a higher percentage of underserved minorities and a lower percentage of students living in poverty (based on qualification for the Free or Reduced Price Lunch program) than Sunrise District (Table 1). One mostly Hispanic school in Farrington ranked in the fourth decile on state tests. All the other middle schools in both districts ranked in the seventh decile or above, including two in Sunrise in the top decile.

Table 1: Research Sites: Middle School Students in Two Districts, 2005–2006

District	Students	Free and Reduced Lunch %	Underserved Minority %
Farrington	3,175	31%	45%
Sunrise	5,655	15%	27%

Research Questions

This study is based on broader research (Grimes, 2008), which addressed a wide range of issues, including educational technology diffusion and student test score results. In this paper we focus on two research questions:

1. Teachers: What were teachers' attitudes toward and instructional practices with AWE?
2. Students: What were students' attitudes toward and writing practices with AWE?

Methodology

Data collection included classroom observations, interviews of teachers and district administrators, surveys of teachers and students, collection of sample essays, and reports of MA use generated by the MA administrator tools.

Interviews and Classroom Observations

We selected two focus schools in each district for onsite data collection (classroom observations and interviews with principals and teachers). Selection of schools was based on socio-economic status, which coincided roughly with academic ranking based on state test scores. We also interviewed district administrators responsible for computer systems, ELA instruction, MA training, and grant administration. A summary of field research data collection methods appears in Table 2 (next page).

Table 2: Field Research Data Collection Summary

Year	District	School	Administrator Interviews	Teacher Interviews	Classroom Observations
2005	Farrington	Subtotal	2	6	19
2007	Farrington	Subtotal	7	6	26
2007	Sunrise	Subtotal	6	4	20
		Totals	**15**	**16**	**65**

Note. In Farrington District we also observed two half-day MA training classes in 2004, two full-day training classes in 2006. We conducted two interviews with the trainer, an experienced former teacher serving as Vantage Learning's national director of MA training.

We asked principals in the four focus schools to recommend two ELA teachers to participate in interviews and classroom observations, one teacher with average or above-average students, and one with below-average or ELL students. It is likely that principles chose from among their better teachers for this purpose, however, the answers to survey questions did not differ substantially between teachers we interviewed and those we did not interview. All the teachers we asked to interview and observe agreed to participate.

Interviews of teachers were semi-structured and lasted from 30 to 60 minutes each, except for two teachers who responded to questions in writing because they were unable to schedule an interview at school. Interviews of administrators were also semi-structured and lasted from 15 to 60 minutes, except for one that was only 5 minutes. All interviews were transcribed, and the transcriptions were coded with qualitative data analysis software (atlas.ti) using a grounded theory coding scheme (Strauss & Corbin, 1998). Field notes from classroom observations were similarly coded for analysis. The majority (83) of the codes related to use of AWE or attitudes towards AWE. Frequently used subtopics included classroom management, scaffolding, and instructional method (e.g., process-oriented or formulaic).

Surveys

In year 1 we conducted two online surveys in Farrington as part of a larger study of their one-to-one laptop program. Those surveys, one of teachers and one of students, included a number of questions on MA. In year 3 we conducted a new online survey on MA in both districts, and invited all ELA teachers who had used it in class for more than 15 minutes. It was completed by 42 teachers (19 in Farrington and 23 in Sunrise), over two-thirds of all full-time ELL and ELA teachers in the two districts.

Survey questions addressed teachers' and students' uses of and attitudes toward MA. Responses to many questions were based on a five-point Likert scale, e.g., "Very Negative," "Slightly Negative," "Neutral," "Slightly Positive," and "Very Positive." We converted those ordinal response categories to the integers from one to five. (Although we report mean ranks as a rough guide to central tendency, we caution that they are not true means because there is no way of assuring that the underlying intervals are equidistant.)

We used three non-parametric statistical tests to analyze survey data. The Wilcoxon signed rank test was used when sampling from a single group, such as all students or all teachers who used MA. This test showed the probability that the median response was not neutral. In order to compare inter-district differences in teachers' attitudes we used the Mann-Whitney rank sum test (AKA "Wilcoxon Mann-Whitney" rank sum test). We used a Pearson chi-squared test to determine whether teachers preferred specific writing tools for teaching specific genres and six traits of writing (Tables 9 and 10, page 22). The signed rank and rank sum tests were run with exact algorithms, and checked with asymptotic approximations.

Quantitative Analysis of My Access Use

We knew from our early interviews with teachers that most of them reported that students wrote more and revised more when using AWE. We also knew that the administrators in both districts strongly supported MA, and we wondered if administrative pressure might bias teachers' reported use. A key methodological question therefore was how to assess the trustworthiness of teacher reports. Our solution was to generate MA administrative reports showing all essays for all students in each of the eight schools, and to compare the "hard data" from those reports with teachers' claims of MA use. We collected three full years of data on MA for Farrington, and data for year 3 in Sunrise. For each of the available years we then calculated the average number of essays per student per year and the average number of revisions per essay. We used the resulting numbers to triangulate teacher's reports of MA usage.

Findings

AWE Dissemination

In order to understand the local contexts of AWE use, we begin with a description of MA dissemination in the two districts. Both Farrington and Sunrise districts obtained grants to fund most or all of the costs for expanding their technology infrastructures (computers and network hardware). However, the two districts differed markedly in their technology infrastructures, professional training of teachers, and district policies toward AWE.

In year 1, after a small pilot program with MA the prior year, Farrington District introduced it to students in its four middle schools. In two of those schools, where MA was implemented as part of a one-to-one laptop program (Grimes & Warschauer, 2008a; Suhr, Hernandez, Grimes, & Warschauer, 2010; Warschauer, 2006a, 2006b), access to MA was easier than in the other two schools, where access was through computer labs or laptop carts. Sunrise implemented MA for all seventh graders in year 1 and all eighth graders in year 2, primarily via new laptop carts dedicated to MA use, and secondarily via older computer labs.

In year 1 Farrington ELA teachers received a half a day of MA training from an expert trainer provided by Vantage Learning, and a full day of training from her in years 2 and 3. In contrast, Sunrise sought self-sufficiency in teacher support, and implemented what one administrator called a “mother hen” approach. The expert trainer from Vantage trained a core group of teachers including the writing project director (an experienced ELA teacher) and the technical support leader. The project director then led MA training for other ELA teachers in the district. First she led one-day MA training classes for all ELA teachers. Then she spent approximately one day in each teacher’s classroom, first teaching a few classes with MA, then watching and assisting the classroom teacher in her first few classes teaching with MA. In addition, one ELA teacher in each school in Sunrise was appointed Site Technology Coordinator to support other teachers with MA at her school.

Training in both districts went far beyond technical features of the program; it emphasized how to integrate MA into teachers’ instructional plans. Both districts provided ELA teachers with additional non-AWE training in writing instruction, especially in Sunrise.

Both districts initially experienced some network instability, which was less severe and was more quickly remedied in Farrington than in Sunrise. Although ELA teachers in Sunrise had access to new laptop carts reserved exclusively for MA use, many preferred to use computer labs because of the greater speed and reliability of wired network connections.

Administrators that we interviewed in both districts expressed satisfaction with their MA programs, and in year 3 both districts paid in advance for a three year extension to their MA licenses. Farrington used district funds for the extension and reduced the number of MA licenses to approximately 65% of middle school students. Sunrise administrators used surplus funds from their grant to purchase licenses for almost all middle school students.

AWE Usage

Usage levels were modest and approximately equal in both districts, in spite of substantial differences in computer access, network reliability, professional training, and peer support networks for teachers. As Table 3 shows, students using MA in Sunrise District wrote a mean of 3.6 essays in year 3, and those in Farrington wrote a mean of 4.0 essays.

Table 3: My Access Use by District, Year 3 (2006–07)

District	# Students	# Essays/Student	% Revised	# Revisions
Sunrise	5,227	3.6	61%	3.1
Farrington	1,778	4.0	53%	3.2
Combined	7,005	3.7	59%	3.1

Note. Columns:

- #Students indicates the number of students who used MA.
- #Essays/Student is the mean number of essays submitted to MA by each student. The overall averages are likely to be low because they include groups (English Language Learners, special education students, and transfer students) who were likely to have limited use of MA. Mainstream native English speakers therefore probably used it more than the numbers here indicate. Usage in Farrington District was also depressed in the 2006–2007 school year because of administrative delays at Vantage Learning.
- % Revised is the percentage of essays that were revised at least once.
- # Revisions is the mean number of revisions for essays that were revised. Most of the revisions were limited to correction of spelling, punctuation, grammar, or word usage highlighted by the program, and some of the resubmissions contained no changes at all.
- The percentage of essays revised and the average number of revisions per essay are probably lower than what they would be under optimal conditions. Reasons include: (a) teachers used MA to simulate timed writing exams, (b) districts used MA for district assessments (year 2 Farrington and all years in Sunrise), (c) lack of class time for writing, especially for slow writers.

The number of essays per student in Farrington District approximately doubled between years 1 and 3, possibly in part because teachers' confidence in the program increased with additional training and experience with MA. The percentage of essays that were revised also increased, from 12% to 53% in the same period.

Usage levels for years 1 and 2 were not available in Sunrise. Several teachers suggested that it may have decreased in year 3 because the writing project coordinator had almost no time to support teachers. Although our formal data collection ceased at the end of year 3, we learned halfway through year 4 that some teachers in Sunrise had given up in their attempts to use MA due to decreased support and continued problems with wireless networks.

Teachers

We analyze teachers' attitudes and instructional practices in seven sub-sections: *Changes in Amount of Writing, Approaches to Composition Instruction, Teachers' Attitudes toward Automated Holistic Scoring, Teachers' Attitudes toward Automated Domain Scoring, Classroom Management with AWE, Different Types of Writing, Different Types of Students,* and *Inter-District Differences.*

Changes in Amount of Writing

Teachers varied widely in their estimates of how much using MA increased or decreased the amount of writing their students composed; the mean estimated change was 33% increase (first question in Table 4). Teachers reported that 39% of their students' writing was done with AWE, but this percentage also varied widely among teachers (second question in Table 4). Interviews confirmed that non-AWE writing remained an important part of most teachers' instructional plans. The fact that more writing was done without AWE than with it was due largely to teachers' desire to balance AWE with other writing, a practice we found in even the most enthusiastically pro-AWE teachers.

Table 4 : My Access Teacher Survey: Questions on Amount of Writing with My Access

"How much has using My Access increased or decreased the amount of writing your students do overall?"

Pct.	60–90% decrease	30–60% decrease	10–30% decrease	No change	10–30% increase	30–60% increase	60–90% increase	90–120% increase	120–400% increase	400+% increase	Mean %
Count	0	0	0	9	12	10	6	1	0	0	33%

"Roughly what percent of your students' writing for your class is done with My Access?"

Pct.	0	10%	20%	30%	40%	50%	60%	70%	80%	90%	100%	Mean %
Count	1	8	6	6	2	3	3	5	2	2	0	39%

Note.

1. The "Pct." rows are percentage change. The "Count" rows show response counts. For example, in response to the second question, 1 teacher answered 0%, 8 answered 10%, etc.
2. The "Mean %" columns are weighted means of all responses. The weighted mean for the first question is based on the midpoint of each allowed answer except the last. For example, the midpoint of "10–30% decrease" is –20%; for "10–30 increase" it is +20%, etc. Three ranges that had zero responses and have been compressed into the one column labeled "120- 400% increase."
3. 99% confidence intervals, based on a student's t distribution:
   a. First question: (20%, 45%), n = 38
   b. Second question: (27%, 51%), n = 38

[4] It was clear that most teachers were not accustomed to quantifying how much their students wrote. We therefore take this table as only a rough collective guesstimate. In order to create a reliable quantitative comparison of AWE and non-AWE writing, it would be necessary for all writing to be submitted in digital form to accurately count the number of words in each essay, each revision, etc.

Teachers indicated in the survey and interviews that time for writing practice was limited by the need for other instruction, especially in writing mechanics. As one teacher said, "I like it (MA) a lot and I would use it more often if we did not have test scores to worry about."

Approaches to Composition Instruction

We noted two non-technical influences on teachers' instructional practices when teaching with MA: pressures to use a formulaic model for essay structure, and teachers' individual instructional practices, which they adapted to MA. A formulaic model dictates a standardized essay structure, most commonly a five paragraph formula consisting of a thesis statement, three supporting paragraphs, and a conclusion. A spokesman for Vantage Learning said that MA rewards essays structured like the high-scoring essays used to train the scoring engine on essays for a particular prompt, so if the human graders of the training essays rewarded a particular formula, the software does likewise (personal communication, B. Maguire, Oct., 2009). Whatever the actual automated scoring mechanism, many teachers and students in this study believed following a five-paragraph formula would boost an essay's score. A formulaic approach also dominated the non-AWE in-service training in writing instruction that both districts provided their ELA teachers.

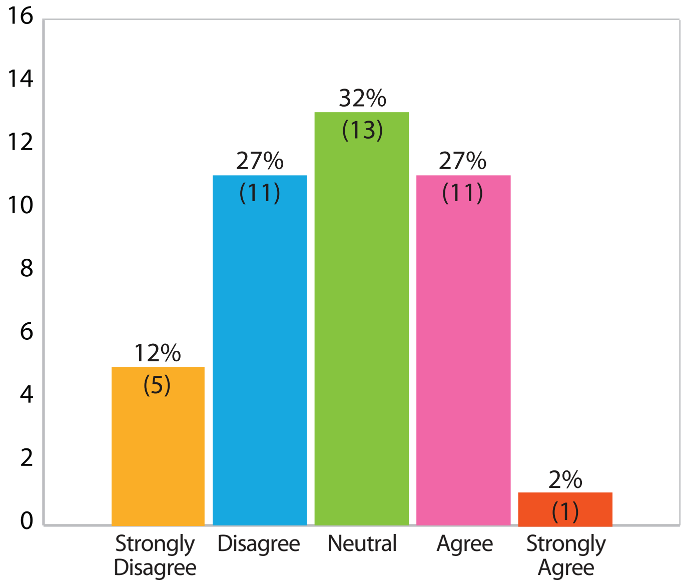
Ironically, several teachers who approached writing instruction in very formulaic, procedural ways also used MA in creative ways and advised advanced students to dispense with writing formulas. For example, one assigned nine short first-person narratives based on a character in a book that students read, a practice she had established before introducing MA. Students submitted each chapter in their narratives to MA for feedback and scoring, using a prompt that was unrelated to their stories. (A scoring engine called "Approximator" that does not require a corpus of human-graded essays for each prompt had recently been added to MA, but this teacher was not aware of it.) She used MA as a motivator to encourage writing and revising, while continuing with other instructional practices she had used for years.

Teachers using MA also continued their prior practice of commenting on individual essays, either orally or in writing. (Both *Criterion* and MA allow teachers to insert comments in students' papers.)

Teachers' Attitudes toward Automated Holistic Scoring

Most teachers in this study treated automated scoring as useful, even when they put little confidence in it. Overall opinions regarding the fairness and accuracy of automated scores were only slightly less than neutral (Figure 1), and the difference from neutral was not statistically significant.

Figure 1: Responses to "My Access gives fair and accurate scores"



One expert teacher who had generally positive attitudes toward MA wrote a comment that illustrates the need for a sensible teacher or mentor to remind students of the limitations of automated scoring:

> Sometimes while helping edit a student's paper, the grade drops after we fix major mistakes. Students get discouraged. However, I tell them the grade from the computer is not accurate. Usually, after sitting down and helping the students, we can always get their grades up and over what they were earlier. Last week, I had one honors student who accidentally had the word "So" as a sentence at the bottom of his essay (apparently he had forgotten to delete the word). When he deleted the word, his grade dropped from all 4's to 3's. He was frustrated. He told me he wanted to leave the word "So" at the bottom of the page so that he would get the higher grade. I had to tell him that it didn't make any sense to leave the word there and that he must delete it. I encouraged him to add something to another part of his essay. He did and his score went back up. It is these types of errors that can be frustrating to the students.

Teachers varied widely in how much they said they relied on automated scoring to assign students' grades. The mean percentage of students' grades that was reportedly determined by My Accesss scores was only 18% (Table 5).

Table 5: My Access Teacher Survey: Responses to "Roughly what percent of your students' grades are determined by the scores My Access gives?"

Grade %	0	10%	20%	30%	40%	50%	60%	70%	80%	90%	100%	Sum
Count	9	14	8	2	0	2	0	2	0	1	0	40
Teacher %	24%	37%	21%	5%	0%	5%	0%	5%	0%	3%	0%	100%

Note.

1. The weighted mean was 18%.
2. "Grade %" is the response category, i.e., the percentage of essay grade determined by automated scores from My Access.
3. "Count" is the number of teachers in each response category.
4. "Teacher %" is the percentage of teachers in each response category.
5. 99% confidence interval, based on a student's t distribution: (9%, 28%), n = 38

Teachers' Attitudes toward Automated Domain Scoring

The user interfaces in MA and *Criterion* are organized largely by domains, implying that the technology for domain scoring is sufficiently valid and reliable to use as a guide for revising an essay. However domain scores were so closely correlated with holistic scores and each other that there was little room for meaningful distinction among domains. In a sample of over 3,300 MA essays, we found that correlations between domain scores and holistic scores ranged from .89 to .98 (Table 6). Scores in at least one domain, focus and meaning, were so closely correlated with holistic scores that it was meaningless to report them separately. Although lower correlations would certainly not indicate valid measures, they would at least allow room for meaningful differences among domain scores.

Table 6: Pearson Correlation Coefficient of Domain Scores with Holistic Scores (N = 3,317)

Focus & Meaning	Organization	Content & Development	Language Use & Style	Mechanics & Conventions
0.98	0.90	0.89	0.93	0.91

A number of studies authored primarily or entirely by researchers at ETS have investigated feature scoring in e-rater (Attali & Burstein, 2004, 2006; Burstein et al., 2004; Burstein & Marcu, 2003; Davey, 2009; Lee, Gentile, & Kantor, 2008; Quinlan, Higgins, & Wolff, 2009), and have established internal standards for acceptable levels of association between human raters and e-rater (Davey, 2009; Williamson, 2009). In spite of what we felt were obvious reasons to doubt the domain scores in MA (high error rates in low-level domains, lack of specificity in high-level domains, and high correlation between holistic and domain scores), we observed several teachers strongly encourage students to use them to guide revision. However, at least one teacher considered the domain scores a misleading distraction and turned them off.

Classroom Management with AWE

When all the technologies worked as planned, AWE simplified classroom management. Teachers we observed appeared more relaxed when students wrote with AWE instead of pencil and paper, and students often became noticeably more focused and engaged the minute the teacher let them starting writing with AWE. Table 7 presents the statistically significant responses of teachers to a multi-part survey question on their attitudes toward MA.

Table 7: My Access Teacher Survey: Responses to "Please indicate your agreement or disagreement: My Access …

	Strongly Disagree	Disagree	Neutral	Agree	Strongly Agree	n	Mean	p	V
makes writing instruction easier."	1	6	9	18	7	41	3.59	.0013	425
saves me time."	1	3	3	18	16	41	4.10	<.001	678
makes teaching more enjoyable."	1	5	11	13	11	41	3.68	<.001	393
lets me focus on higher concerns of writing instead of mechanics."	2	9	6	14	10	41	3.51	<.05	463

Note.

1. The five columns from "Strongly Disagree" to "Strongly Agree" show the number of responses to each point on the Likert scale.
2. The "Mean" column shows the mean response on a 5-point scale, where "Strongly Disagree" is 1, "Neutral" is 3, and "Strongly Agree" is 5.
3. The "*p*" column is *p*-value.
4. The "V" column is the test statistic for Wilcoxon signed rank Test.
5. Only questions with statistically significant findings are shown.

Overall responses were significantly positive on three questions related to classroom management: "My Access makes writing instruction easier", "My Access saves me time", and "My Access makes teaching more enjoyable." The lower limits of the 95% confidence intervals for all three questions were at least 3.5, indicating that the true medians were significantly positive.

Interviews and classroom observations confirmed that teachers felt MA made teaching easier and more enjoyable. One teacher called it "a second pair of eyes in the classroom."

Table 8 may help assuage fears that teachers relied solely on automated scoring to assign grades on essays. In response to a question on how closely they read essays in MA, none of the teachers who responded said they did not read them at all, and 41% said they read essays just as thoroughly as without MA.

Table 8: My Access Teacher Survey: Responses to "How closely do you usually read your students' essays in My Access?"

Response	n	Pct.
"I hardly read them at all."	0	0%
"I spot check a few sentences, but leave most the review to My Access."	11	28%
"I read the essays, but not as thoroughly as I do without My Access."	12	31%
"I read the essays just as thoroughly as when they don't use My Access."	16	41%
Total	**39**	**100%**

Except when students were unable to complete writing assignments in class, the vast majority of writing for MA was done in class, not as homework. Teachers we interviewed gave three main reasons for not assigning homework in MA: concern that some students might not have access to the Internet at home, concern that many students would not do their homework, and concern that students would get unauthorized assistance on their assignments. To address the first concern some teachers assigned the first draft to be written at home and copied into MA in class. Teachers appeared to feel somewhat powerless about the second concern. One weary teacher, when asked about students' attitudes towards homework, said, "They don't care and their parents don't care." Although her school ranked lowest in socioeconomic status (SES) of schools in this study, the homework completion problem affected the high-SES schools, too. One teacher in the highest-SES school estimated that only 50 or 60 percent of students did assigned homework. The low homework completion rates placed additional demands on classroom time, as teachers used class time for activities they would have assigned as homework in a more favorable academic culture.

Different Types of Writing

In spite of the positive attitudes teachers expressed toward MA as a motivator and a classroom management tool, it was not their preferred writing tool in three out of the four genres we asked about (Table 9). However, preferences for low-tech writing tools (word processor, pen, or pencil) were not statistically significant on a chi-squared test that included "no preference," (*p* = .481), and only marginally significant (*p* = .074) in a binomial test that omitted "no preference".

Table 9: My Access Teacher Survey: Responses to "Which writing tool do you prefer for teaching each of the following genres?"

	My Access	Word processor, pen or pencil	No preference
Informative	14	**15**	12
Narrative	14	**18**	9
Persuasive	**17**	14	10
Literary Analysis	8	**21**	11
Totals	53	**68**	42

Note.

1. Chi-square = 5.491, *p* = .481 with 6 df (4 rows, 3 columns).
2. Binomial probability of totals for first two columns = .0744. Based on 53 or fewer of 121 equally likely choices (53 + 68 = 121)

When asked about their favorite writing tool for teaching each of six traits of writing (Spandel, 2008), teachers preferred conventional tools for four traits (ideas, organization, voice, and word choice), and preferred MA for only two (sentence fluency and conventions, Table 10). This time the probability was significant (chi-squared, *p* = .033).

Table 10: My Access Teacher Survey: Responses to "Which writing tool do you prefer for teaching each of the following skills ('Six Traits')?"

	Ideas	Organization	Voice	Word Choice	Sentence Fluency	Conventions
My Access	6	14	9	13	**17**	**21**
Word processor, pen or pencil	**27**	**18**	**21**	**17**	14	13
No preference	8	8	11	11	10	7

Note.

1. Chi-square = 19.586, *p* = .033 with 10 df.

It is not surprising that MA was the preferred tool for teaching the mechanistic writing skills because those skills are more amenable to automated scoring and feedback than higher-order concerns. At least three of the four traits for which teachers preferred conventional writing tools were traits which appear difficult to assess computationally: ideas, organization, and voice.

Different Types of Students

Classrooms we observed included diverse student groups in terms of SES, achievement levels, English skills, and motivation to learn. Students of all types appeared more focused when writing with MA than when writing with pen and pencil. Teachers in the survey reported that MA assists writing development for all five categories of learners mentioned—English language learners, special education students, gifted students, at-risk students, and general students without special needs (Table 11). The positive attitudes were statistically significant for all five learner categories.

Table 11: My Access Teacher Survey: Responses to "For each category of students, please indicate how much My Access assists or impedes their writing development:"

	Very Negative	Slightly Negative	Neutral	Slightly Positive	Very Positive	n	*p*	V
English Language Learners	1	1	7	10	8	27	< .001	208.5
Special Education	0	2	2	13	9	26	< .001	309
Gifted	1	4	1	7	11	24	< .01	258
At-risk	1	1	2	19	10	33	< .001	491
General students, no special needs	0	1	3	19	10	33	< .001	485.5

Note.

1. *P*-values are from a one-sided signed rank test. They indicate the probability that the true medians are greater than 3 (neutral). A two-sided test also gives *p* < = .001 for all of the five categories except gifted, for which *p* is only slight more than .001.
2. The "V" column is the test statistic for the Wilcoxon signed rank test

One teacher who specialized in English Language Development said only the more advanced ELL students can use MA to advantage because the lower level ELL students are unable to write three paragraphs. Two teachers noted in interviews that it is easier to see benefits with students at the lower end of the grade spectrum (ELL and special education students) than with honors students. One said:

> There have been times when I've turned off the scoring so they couldn't see it so they would focus on what they knew about writing based on what we talked about in class. And then I would let them see after they submitted the final time on the deadline, then I would show them what they got… . Some of them hated it because that's what they want from My Access, the instant gratification, the immediate feedback. But some of my honors kids, especially, if there are prompts that I think grade too easily … their first draft they get a six, so there is no motivation to revise. Whereas if they think they need to make it better, if they think there are still things to work on, then they will keep working on it.

Two middle school teachers also noted in interviews that some of their gifted students learned to spoof the automated scoring engine, for example, by repeating words just to increase the word count. This observation accords with findings in a study we conducted on *Criterion* in another school district, where one teacher and the principal (a *Criterion* expert and former English teacher) reported that honors students quickly learned to spoof the automated scoring engine, especially in grades eleven and twelve (Grimes & Warschauer, 2006).

Inter-District Differences in Teachers' Attitudes and Practices

We compared survey responses of teachers in the two districts. Compared to teachers in Farrington, those in Sunrise reported (a) more positive attitudes toward using MA and greater benefit for students, (b) greater peer collaboration, (c) a higher percentage of students' writing was done in MA, and (d) a higher percentage of grades were based on automated scores. (For more detail, see Grimes, 2008.) We attribute this difference in attitudes to stronger top-down support for AWE in Sunrise, primarily through systematic peer support and the inspired leadership of the district writing project director, who became an effective MA evangelist. Although Farrington also provided teachers with excellent MA training and more reliable computer networks than Sunrise, teachers in Farrington were not pressured to use MA, and had less of a centrally-organized peer support system.

Students

In this section we look at students' attitudes and writing practices in six sub-sections: *Students' Attitudes toward Automated Scores, Amount of Revision, Types of Revision, Use of Feedback*, and *Students' Writing Development*.

Students' Attitudes toward Automated Scores

Classroom observations revealed that even when students realized that automated scores were sometimes ungrounded, the prospect of receiving a quick score motivated them to focus more than if they expected to wait days or weeks for their score. This increased motivation was confirmed in the survey, where 30 out of 40 teachers agreed or strongly agreed that students were more motivated to write with MA than with a word processor (Table 12).

Table 12: My Access Teacher Survey: Responses to "Compared to writing with a WORD PROCESSOR, when using My Access my students …

	Strongly Disagree	Disagree	Neutral	Agree	Strongly Agree	n	*Mean*	*p*
are more motivated to write."	0	1	9	18	12	40	4.03	< .001
revise more."	0	5	5	14	16	40	4.03	< .001
write more mechanically and superficially."	1	18	13	6	1	39	2.69	>.05

Note.

1. As in Table 7, the "Mean" column shows the mean response on a 5-point scale, where "Strongly Disagree" is 1, "Neutral" is 3, and "Strongly Agree" is 5.
2. *P*-values indicate the probability that the sample median is equal to 3 (neutral) based on a two-tailed test.

When we observed students who were accustomed to writing with MA, most of them appeared to realize that the automated scores often did not reflect a human understanding of their essay, as a teacher or peer would. They took their automated scores more lightly than they took grades from their teacher, probably due at least in part to knowing that their teacher was likely to override or discount the automated scores when assigning grades. Teachers and students compared writing with AWE to playing a video game, an attitude that was confirmed by our classroom observations. In two classes we observed students brag on receiving a top score, or moan with mock agony on receiving a low one. It was commonplace for teachers to encourage students to score high, thereby implying confidence in automatic scoring. When students challenged the automated scores, it put teachers in a slightly awkward position of explaining that the software judged essays differently than people, and sometimes made mistakes that would be obvious to an intelligent person.

Teachers also reported in interviews an increase in student autonomy when using AWE, and we repeatedly observed students who sought advice from the built-in writing tools instead of asking the teacher for help. Teachers also reported that faster writers kept constructively engaged after submitting their first draft of an essay; instead of sitting idle until the end of class, they would revise their essays in hopes of a better score.

A student survey we conducted on the one-to-one laptop program in Farrington District in 2005 included many questions on MA. The aggregate responses to eight of these indicated statistically significant positive attitudes toward using MA (*p* < .001 in all cases, Table 13).

Table 13: 2005 Farrington Student Survey: Responses to "Please indicate how much you agree or disagree with each of the following statements:"

	Mean	P	*V*	*n*
I find My Access easy to use.	3.7	< .001	51,854	464
I sometimes have trouble knowing how to use My Access.	2.5	< .001	15,715	470
I like writing with My Access.	3.2	< .001	28,008	469
I revise my writing more when I use My Access.	3.4	< .001	38,507	456
Writing with My Access has increased my confidence in my writing.	3.3	< .001	32,843	459
My Tutor has good suggestions for improving my writing.	3.2	< .001	22,357	389
The essay scores My Access gives are fair.	3.4	< .001	40,655	463
My Access helps improve my writing.	3.6	< .001	45,179	451

Note.

1. As in Tables 7 and 12, the "Mean" column shows the mean response on a 5-point scale, where "Strongly Disagree" is 1, "Neutral" is 3, and "Strongly Agree" is 5.
2. *P*-values indicate the probability that the sample median is equal to 3 (neutral) based on a two-tailed test. Since all the *p*-values are highly significant (*p* < .001), this null hypothesis is rejected for all eight questions.

Students were also less critical or less aware of scoring flaws than teachers. In response to "The essay scores My Access gives are fair" students' mean response was 3.4 (Table 13), whereas teachers' mean response to a similar question, "My Access gives fair and accurate scores," was only 2.8.

In contrast to the motivating effect of holistic scores, domain (trait) scores seemed to confuse some students. Possible reasons include hard-to-understand distinctions among domains; inaccurate domain scoring; errors in the low-level feedback; and verbose, generic high-level feedback.

Amount of Revision

The amount of revision with MA varied greatly with time, among teachers, and between districts. The first year of large-scale AWE use in Farrington District showed no evidence of increased revising, as only 12% of essays had more than one draft. By the third year of the program that figure had risen to 53% (Table 3, page 15), suggesting that teachers in Farrington may have gradually learned to allow more time for writing assignments in order to exploit the program's potential for encouraging revision. In the third year at Sunrise District, 62% of essays in the district were revised at least once. Revision practices varied widely among teachers and students.

As mentioned, the number of revisions would have been higher if students had been given time and permission to revise all assignments. Many of the revisions were superficial, e.g., quick attempts to correct errors highlighted by the feedback on writing mechanics or word choice.

Interviews and surveys with teachers and students indicated increased amounts of revision. For example, 30 of 40 respondents on the teacher survey agreed or strongly agreed that students revised more when writing with MA (Table 12, page 25).

Use of Feedback

Students generally used the low-level feedback (on spelling, punctuation, grammar, and word choice) before they attempted to use the high-level feedback (on organization and development). However, terms such as "<adverb placement>" were frequently over students' heads, especially for ELL students. Two teachers said the amount of low-level feedback was so overwhelming that they had to mark just a few priorities for students to focus on.

Teachers often supplemented the MA revision tools with handouts. At least one required students to print their first draft, mark it with intended revisions, and show it to the teacher before entering the changes in MA. A number of teachers said that the high-level feedback in My Tutor was too verbose and generic. One wrote:

> I often find that students want to know their scores but are reluctant to go back and read the pages of suggestions from the computer. I have actually developed handouts that are checklists for revisions that force students to read the comments and suggestions that are available from My Access.

Students' Writing Development

As mentioned, AWE has been criticized for distorting the writing process. When we gave teachers the opportunity to criticize MA, answers tended to be roughly balanced between positive and negative. Overall, teachers did not feel that it "encourages mechanistic, formulaic writing" (Table 7, page 20) or that it "discourages thoughtful, reflective writing" (Table 7). Nor did they feel that wrote more mechanically and superficially with MA than with a word processor (Table 12, page 25).

Our classroom observations and teacher interviews confirmed the survey results indicating increasing student motivation to write and revise with MA. However, we observed some classrooms in which automated scoring appeared to shift the goal of writing from communication to improving the score, at least temporarily. This is a concern because the motivational literature is clear on the advantages of internal motivators (e.g., satisfaction from learning or from self-expression) over external motivators (e.g., grades or praise) (Ames, 1992; Ames & Archer, 1988; Elliott & Dweck, 1988; Lepper & Cordova, 1992; Lepper & Malone, 1987; Meece, Anderman, & Anderman, 2006; Pajares, 2001; Warschauer, 1996).

Discussion

We discuss the findings in five sections related to teachers, students, technologies, human-computer interactions, and recommendations.

Teachers: Easier Classroom Management

For the majority of teachers in this study, using AWE simplified classroom management for several reasons: students were more motivated to write and revise, they were more autonomous, and their writing portfolios were conveniently organized.

The potential for AWE to ease teachers' stress sometimes backfired when technical problems exceeded a teacher's troubleshooting skill, wasting precious class time. We heard of one large school district on the East coast that obviated such mishaps by sending a computer technician to the classroom whenever a class was scheduled to use AWE. This practice seems to us an excellent precaution, at least until the network is reliable and teachers develop confidence in their ability to deal with it.

Collaborative student-teacher relationships have been a goal of educational reformers since John Dewey and Lev Vygotsky (Dewey, 1966; Vygotsky, 1986; Vygotsky, Cole, John-Steiner, & Scribner, 1978). We observed that teachers were able to temporarily relax their role as judges of student performance and play a more sympathetic, coaching role with students by offloading the preliminary evaluation of essays to the machine. The potential for AWE to facilitate this shift from an adversarial role of teacher-as-judge toward a supportive role of teacher-as-coach has been noted elsewhere (Myers, 2003), and has been disputed (Rothermel, 2007, p. 203). It runs counter to criticisms of automated scoring as non-social, dehumanizing, or even "Orwellian" (CCCC Executive Committee, 2004, 2006; Cheville, 2004; Huot, 1996; Ziegler, 2007).

We suspect that whether AWE humanizes or dehumanizes writing instruction depends much more on the teacher than the machine. If a teacher relates poorly to students, is preoccupied with technical concerns, or uses automated scoring to determine grades, using AWE is likely to dehumanize instruction. On the other hand, if a teacher uses the software to overcome students' reluctance to write and to help with low-level errors so that she can focus on high-level concerns like ideas and style, then it is likely to contribute to more human-oriented writing.

It could be argued that teachers were biased to rate the software highly in the hope that it would increase their chances of keeping it, because they were in the habit of defending familiar practices, or due to the social desirability bias of using the latest educational technology. We believe these potential biases were probably negligible because many teachers who endorsed the product overall did not hesitate to note its limitations, and the minority of teachers who criticized or avoided it evinced no fear in so doing. However, we cannot rule out the possibility of a Hawthorne effect for the teachers we interviewed and observed in class, especially in Sunrise district, which had a partnership agreement with the vendor to demonstrate MA use to potential customers.

Students: Mechanistic Metrics Motivate Message

We address students' writing in terms of three issues: *audience, scoring*, and *revision*.

Audience

The process-oriented composition literature generally supports writing for authentic audiences (Atwell, 1998; DeJoy, 1999; Faigley, 1986; Olson, 2003; Tobin, 2002). Even though automated scoring is the antithesis of an authentic audience, it provides a simulated audience that responds faster than most human readers. As with video games where a player can be injured or killed repeatedly without pain, machine scoring of essays was less threatening than human critics. Several teachers said that submitting early drafts to the machine boosted students' confidence in their writing, even when they realized the automated score did not reflect human-like judgment.

Is students' writing development distorted when they learn to write for a machine that relies on countable features of a text and ignores the non-countable traits that characterize good writing? As social informatics would predict, the answer appears to depend more on the teacher and the setting than the tool. If teachers present the software to help students simply to develop an essay to the point that it is worthy of human readers' attention, we see no reason to fear the writing process will be distorted.

Scoring

Many authors have noted the potential vulnerability of educational software to "gaming," or exploiting weaknesses in automated evaluation to score well without mastering the intended pedagogical content (Baker, 2005; Powers, Burstein, Chodorow, Fowles, & Kukich, 2001). We studied 20 essays in depth and found no evidence of gaming attempts. However, as mentioned, several teachers said that some students tried to spoof the scoring engine. We feel that when students experiment with spoofing, they are likely to exercise critical thinking skills and enhance their awareness of the differences between human and machine readers. Therefore we are not concerned about spoofing, provided that students do not use a "cheat sheet" and a spoofing-vigilant human reader scans the final draft of each essay.

A prominent member of the composition instruction community, Julie Cheville (2004) wrote, *"The standards of correctness that constitute the diagnostic domains of a program like Criterion are arbitrary"* (original italics). Many publications, most published since Cheville's statement, testify to great efforts by Criterion's developers to develop accurate domain scoring (Attali, 2007; Davey, 2009; Lee et al., 2008; Quinlan et al., 2009). Further research is needed on the accuracy of each AWE system's domain scoring. Our experience suggests that domain scoring in MA is less reliable than holistic scoring, a distinction that that teachers and students often miss, at least when they are new to the system.

The students in this study, like the much older students in the study by Scharber et al. (2008), felt strongly about AWE whether they liked it or not. The same was true for teachers and even administrators in this study.

Revision

The composition community generally concurs that revising leads to improved writing (Butler-Nalin, 1984; Fitzgerald, 1987; Flower & Hayes, 1986; Scardamalia & Bereiter, 1986; Sommers, 1980), although the evidence is stronger for high-school age and older writers than for younger ones (Fitzgerald, 1987). The revisions we saw support the prevailing view that “students do not voluntarily revise school-sponsored writing” (Emig, 1971, p. 93, cited in Hillocks, 1986, p. 41), and that younger students are especially disinclined to revise unless prompted to by teacher or peers (Scardamalia & Bereiter, 1986).

What is the pedagogical value, if any, of domain scores and feedback that are significantly better than random chance, but far from perfect? They appear to serve a pointing function, which Vygotsky (1978) noted as one of the earliest and most fundamental uses of language. Spelling and grammar checkers in word processors are error prone in spite of decades of development (Galletta, Durcikova, Everard, & Jones, 2005; Kaplan et al., 1998, p. 3; Vernon, 2000), yet they remain enormously popular, arguably because they reduce the cognitive load in identifying constructions *likely* to be in error. The same argument explains why some teachers highlighted just a few of the many feedback suggestions for students to work on. It also implies that focused feedback is more valuable than generic feedback.

Learners need to be aware that their achievement falls short of their goal before they will take corrective action to close the gap (Black & Wiliam, 1998). Self-assessment skills are notoriously weak among learners in general, especially among novices (Dunning, Heath, & Suls, 2004). It is arguable that any reasonably reliable indicator of less-than-desired achievement is likely to motivate learners to review and revise their work. Peer feedback among novice writers has been found to be helpful as early as third grade (Daiute & Dalton, 1993). The effects also depend on the learner’s expertise and whether the feedback is accompanied by grades (Lipnevich & Smith, 2008). Further research is needed to determine the conditions under which error-prone feedback is helpful, harmless, or harmful. We suspect that erroneous feedback is more frequently miseducative when it is presented as authoritative, and when no human expert is available to override dubious scores and feedback.

Technologies: The Limitations of Automated Scoring

Although many of the technical details of current holistic and analytic scoring engines are proprietary and thus difficult to scrutinize, it is reasonable to assume for at least four reasons that current technologies are approaching their practical limits in validity and reliability. First, in order to accurately assess the objective, denotative features of an essay, including content and organization, the software would have to convert English (or other natural language) to a formal symbolic representation that could be manipulated computationally. That task, called “computational semantics” (roughly equivalent to “natural language understanding,” or NLU) has been a major challenge for computational linguists. The successful extant NLU systems are limited to focused questions within very narrow knowledge domains, such as a specialized branch of medicine or engineering, or narrow tasks such as placing an airline reservation. For NLU to accurately "understand" open-ended essays on the broad topics used in AWE programs, it would require a vast contextual web of situational knowledge that exists only in science fiction, not functioning software systems.

Ellis Page attacked this Gordian knot by going around it (Kaplan et al., 1998, p. 2). A spokesperson for ETS confirmed that *Criterion* and *e-rater* do not build a structured representation of meaning because current NLU techniques are not sufficiently developed. However, they make ample use of other statistical techniques of natural language processing, including discourse analysis (personal communication, D. Williamson, Oct., 2009). A spokesperson for Vantage learning said MA uses NLU techniques, but cannot "understand" the meaning of a text as well as a human reader (personal communication, B. Maguire, Oct., 2009).

Second, natural language understanding technologies primarily or exclusively address denotative meaning, and are unlikely to ever approach a skilled reader's sensitivity to connotative or subjective features of natural language, such as rhythm, voice, style, aesthetics, and author's intent.

Third, since the reliability of automated scoring is measured in terms of human scoring, the reliability of human scoring sets an upper limit on that of automated scoring (Quinlan et al., 2009). This limitation applies to both holistic and analytic scoring, but is more severe with analytic scoring because human inter-rater reliability for it is lower (Lee et al., 2008; personal communication, B. Maguire, Oct., 2009).

Fourth, student essays are frequently "noisy" (containing many errors in punctuation, grammar, spelling, and diction). Although NLU techniques have been developed to minimize the ambiguities introduced by such disfluencies (Bos, 2005; Rosé, Roque, & Bhembe, 2002), we are skeptical of the ability of such techniques to extract authors' intended meaning from the noise.

Although we have discussed the accuracy of automated scoring and feedback at length, we propose that accuracy is secondary to utility. Certainly accuracy (validity plus reliability, roughly speaking) contributes to utility, but so do local social and organizational factors. For example, it appears that My Access was received far more positively in the current study than in an earlier study in Taiwan (Chen & Cheng, 2008) at least partly due to better training of and support for teachers.

Human Trust in Pseudo-Human Software

Teachers and students we observed frequently appeared to place unwarranted trust in AWE technology in at least two ways. First, a halo effect seemed to apply, as users extended their impressions of the accuracy of holistic scoring into domain scoring, in spite of prominent clues that domain scoring was less accurate. Second, when users witnessed a modest simulation of human-like ability, they assumed the software judged essays in a manner similar to humans. (For an explanation of how AES can use countable features of a text to score essays as accurately as humans using very different criteria, see Shermis, Garvan, & Diao, 2008.) Students' trust in the software was socially reinforced by teachers and the MA user interface, which did little to suggest anything other than totally reliable scoring and feedback. Teachers' trust in the software was socially reinforced by administrative support for AWE, professional training, and peer support, especially in Sunrise District, where senior teachers (the Site Technology Coordinators and the writing project director) encouraged MA use. In contrast, users were less inclined to place unwarranted trust in computer hardware and networks, which made no semblance of emulating human judgment, and whose failures, when they occurred, were more obvious and disruptive.

Reeves and Nass (2003) explain unwarranted trust as one of many ways in which people treat computers and other media as social actors. Unwarranted trust is particularly likely when the users are inexperienced (Shapiro & McDonald, 1992, p. 556) or the media purportedly has specialized expertise (Fogg & Tseng, 1999; Shapiro & McDonald, 1992):

> It is as if the "ghost in the machine" has conferred special authoritative status on such sources, perhaps because it is presumed that any information appearing in broadcast or computer-mediated form has already been through appropriate gatekeepers who have verified its accuracy, relevance, validity, and appropriateness. Therefore, its very availability in mediated form is taken as prima facie evidence of its authenticity. (Burgoon et al., 2000, p. 556)

Recommendations

Our study suggests that AWE software can be useful in the classroom if it is deployed in a way that does not distract students from the communicative purpose of writing. Based on our experience, we recommend that teachers take several steps mentioned above: reviewing the final draft of essays, assigning grades, circulating around the classroom to answer questions while students are writing, balancing AWE and non-AWE writing, and having students address other audiences besides the teacher and machine. Teachers may also wish to limit the number of allowed revisions, encourage higher-level revision, schedule sufficient time for multiple drafts, and turn off domain scoring.

Conclusion

Many attempted implementations of educational technology have failed due to teacher resistance rather than deficiencies in the technologies per se (Cuban, 2003, 2005). The success of the AWE projects in Farrington and Sunrise were the result of many local factors that are not easy to replicate, including a mature AWE technology, strong administrative support, excellent professional training, teachers ready to experiment with technology, and in Sunrise, strong peer support among teachers.

Automated writing evaluation has been a topic of hot debate, with some advocates considering it a magic bullet for dramatically improving writing and others viewing it as a threat to the very fabric of education. Our study found little basis for either of these extreme perspectives. We found that mindful use of AWE can help motivate students to write and revise, increase writing practice, and allow teachers to focus on higher-level concerns instead of writing mechanics. However, those benefits require sensible teachers who integrate AWE into a broader writing program emphasizing authentic communication, and who can help students recognize and compensate for the limitations of software that appears more intelligent at first than on deeper inspection. Thus, like many educational technologies, it is unlikely to improve ineffective teaching, but it can help good teachers be more effective.

References

Ames, C. (1992). Classrooms: Goals, Structures, and Student Motivation. *Journal of Educational Psychology, 84*(3), 261–271.

Ames, C., & Archer, J. (1988). Achievement Goals in the Classroom: Students' Learning Strategies and Motivation Processes. *Journal of Educational Psychology, 80*(3), 260–267.

Attali, Y. (2007). *Construct Validity of e-rater® in Scoring TOEFL® Essays* (ETS RR-07-21). Princeton, NJ: ETS.http://www.ets.org/Media/Research/pdf/RR-07-21.pdf.

Attali, Y., & Burstein, J. (2004, June 13 to 18, 2004). *Automated Essay Scoring With E-rater V.2.0*. Paper presented at the Conference of the International Association for Educational Assessment (IAEA), Philadelphia, PA.

Attali, Y., & Burstein, J. (2006). Automated Essay Scoring With e-rater® V.2. *Journal of Technology, Learning, and Assessment, 4*(3).

Atwell, N. (1998). *In the Middle: New Understanding About Writing, Reading, and Learning* (2nd ed.). Portsmouth, NH: Boynton/Cook.

Baker, W.D. (2005). "Layers and Layers" of Teaching Writers' Workshop: A Response to Katie Wood Ray's The Writing Workshop [Electronic Version]. *Pedagogy, 5*, 348-352. Retrieved 9/10/05 from http://muse.jhu.edu/journals/pedagogy/v005/5.2baker.html.

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

Ben-Simon, A., & Bennett, R.E. (2007). Toward More Substantively Meaningful Automated Essay Scoring. *Journal of Technology, Learning, and Assessment, 6*(1), 47. http://www.pedagogy.ir/images/pdf/meaningful-scoring.pdf.

Black, P., & Wiliam, D. (1998). Assessment and classroom learning. *Assessment in Education: Principles, Policy & Practice, 5*(1), 7–68.

Bos, J. (2005). *Towards Wide-Coverage Semantic Interpretation*. Paper presented at the Proceedings of Sixth International Workshop on Computational Semantics IWCS-6, Tilburg, The Netherlands.

Breland, H.M. (1996). Computer-Assisted Writing Assessment: The Politics of Science versus the Humanities. In E. M. White, W. D. Lutz & S. Kamusikiri (Eds.), *Assessment of Writing: Policies, Politics, and Practices* (pp. 249–256). New York: Modern Language Association of America.

Burgoon, J.K., Bonito, J.A., Bengtsson, B., Cederberg, C., Lundeberg, M., & Allspach, L. (2000). Interactivity in human-computer interaction: a study of credibility, understanding, and influence. *Computers in Human Behavior, 16*(6), 553-574.http://www.sciencedirect.com/science/article/B6VDC-41NK8J3-6/2/06e0de01fe8d6b372351733c32b24876.

Burstein, J., Chodorow, M., & Leacock, C. (2004). Automated essay evaluation: the Criterion Online writing service. *AI Magazine, 35*(3), 27–36.

Burstein, J., & Marcu, D. (2003). Automated Evaluation of Discourse Structure in Student Essays In M. D. Shermis & J. Burstein (Eds.), *Automated Essay Scoring: A Cross-Disciplinary Perspective*. Mahwah, New Jersey: Lawerence Erlbaum, Associates.

Butler-Nalin, K. (1984). Revising Patterns in Students' Writing. In A.N. Applebee (Ed.), *Contexts for Learning to Write* (pp. 121–133): Ablex Publishing Co.

CCCC Executive Committee. (2004). CCCC Position Statement on Teaching, Learning, and Assessing Writing in Digital Environments. *NCTE (National Council of Teachers of English) Position Papers*.

CCCC Executive Committee. (2006). Writing Assessment: A Position Statement. *NCTE*.

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*(3), 94–112.

Cheville, J. (2004). Automated Scoring Technologies and the Rising Influence of Error. *English Journal, 93*(4), 47–52.

Chung, G., & Baker, E. (2003). Issues in the reliability and validity of automated scoring of constructed responses. In M. D. Shermis & J. Burstein (Eds.), *Automated essay grading: A cross-disciplinary approach Mahwah, NJ: Erlbaum* (pp. 23–40). Mahwah, NJ: Erlbaum.

Cohen, Y., Ben-Simon, A., & Hovav, M. (2003, October, 2003). *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. http://www.aqa.org.uk/support/iaea/papers/ben-cohen-hovav.pdf.

Cuban, L. (2003). *Oversold and Underused: Computers in the Classroom*. Cambridge, Mass: Harvard University Press.

Cuban, L. (2005). *Teachers and Machines: The Classroom Use of Technology Since 1920*: Teachers College Press.

Daiute, C., & Dalton, B. (1993). Collaboration between children learning to write: Can novices be masters? *Cognition and Instruction, 10*, 281–333.

Davey, T. (2009). *Principles for Building and Evaluating e-rater Models*. Paper presented at the Annual Meeting of the National Council on Measurement in Education.

DeJoy, N.C. (1999). I was a Process-Model Baby. In T. Kent (Ed.), *Post-Process Composition* (pp. 163–178): Southern Illinois University Press.

Dewey, J. (1966). *Democracy And Education: An Introduction To The Philosophy Of Education*. New York: The Free Press.

Dunning, D., Heath, C., & Suls, J.M. (2004). Flawed Self-Assessment: Implications for Health, Education, and theWorkplace. *Psychological Science in the Public Interest, 5*(3), 69–106.

Elbow, P. (1981). *Writing with Power*. Oxford: Oxford University Press.

Elliot, S.M. (2003). IntelliMetric: From Here to Validity. In M. D. Shermis & J. Burstein (Eds.), *Automatic Essay Scoring: A Cross-Disciplinary Approach* (pp. 71–86): Lawrence Erlbaum Associates.

Elliot, S.M., & Mikulas, C. (2004, April 12-16, 2004). *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.

Elliott, E.S., & Dweck, C.S. (1988). Goals: An Approach to Motivation and Achievement. *Journal of Personality and Social Psychology, 54*(1), 5–12.

Ericsson, P.F., & Haswell, R. (Eds.). (2006). *Machine Scoring of Human Essays: Truth and Consequences*: Utah State University Press.

Faigley, L. (1986). Competing Theories of Process: A Critique and a Proposal. *College English, 48*(6), 527–542.

Fitzgerald, J. (1987). Research on Revision in Writing. *Review of Educational Research, 57*(4), 481–506.

Flower, L., & Hayes, J.R. (1986). Detection, Diagnosis, and the Strategies of Revision. *College Composition and Communication, 37*(1), 16–55.

Fogg, B.J., & Tseng, H. (1999, May 15-20, 1999). *The Elements of Computer Credibility*. Paper presented at the CHI 99.

Foltz, P.W., Laham, D., & Landauer, T.K. (1999). Automated Essay Scoring: Applications to Educational Technology [Electronic Version]. *Interactive Multimedia Electronic Journal of Computer-Enhanced Learning*.

Galletta, D.F., Durcikova, A., Everard, A., & Jones, B.M. (2005). Does spell-checking software need a warning label? *Commun. ACM, 48*(7), 82–86.

Grimes, D. (2008). *Middle School Use of Automated Writing Evaluation*. Unpublished Ph.D Dissertation, University of California, Irvine, Irvine, CA. http://douglasgrimes.com/windocs/Grimes–Middle%20School%20Use%20of%20AWE–Final%20Dissertation%20.doc.

Grimes, D., & Warschauer, M. (2006, April 7–11, 2006). *Automated Essay Scoring in the Classroom*. Paper presented at the American Educational Research Association (AERA) Annual Conference, San Francisco, CA.

Grimes, D., & Warschauer, M. (2008a). Learning with Laptops: A Multi-Method Case Study. *Journal of Educational Computing Research (JERIC), 38*(3).

Grimes, D., & Warschauer, M. (2008b). *Middle School Use of Automated Writing Evaluation. Paper presented at the Annual Convention of the American Education Research Association*. from http://douglasgrimes.com/windocs/Grimes+Warschauer–AERA%202008–Middle%20School%20Use%20of%20AWE.doc.

Hillocks, G.J. (1986). Research on Written Composition. Urbana, Illinois: 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.

Kaplan, R.M., Wolff, S., Burstein, J.C., Lu, C., Rock, D., & Kaplan, B. (1998). *Scoring Essays Automatically Using Surface Features*. Princeton, NJ: Graduate Record Examinations Board, Educational Testing Service. http://www.ets.org/Media/Research/pdf/RR-98-39-Kaplan.pdf.

Keith, T.Z. (2003). Validity of Automated Essay Scoring. In M. D. Shermis & J. Burstein (Eds.), *Automated Essay Scoring: A Cross-Disciplinary Perspective* (pp. 147–167).

Kelly, P.A. (2001). *Automated scoring of essays: Evaluating score validity*. Unpublished manuscript.

Kling, R. (1999). What is Social Informatics and Why Does it Matter? *D-Lib Magazine, 5*(1).

Kling, R. (2000). Learning About Information Technologies and Social Change: The Contribution of Social Informatics. *The Information Society, 16*, 217–232.

Landauer, T.K., Laham, D., & Foltz, P.W. (2003). Automated scoring and annotation of essays with the Intelligent Essay Assessor. In M. D. Shermis & J. Burstein (Eds.), *Automated Essay Sscoring: A Cross-Disciplinary Perspective* (pp. 87–112). Mahwah, NJ: Lawrence Erlbaum Associates.

Lee, Y., Gentile, C., & Kantor, R. (2008). *Analytic Scoring of TOEFL® CBT Essays: Scores From Humans and E-rater®* (TOEFL Research Rep. No. RR-81, ETS RR-08-01). Princeton, NJ: ETS.

Lepper, M.R., & Cordova, D.I. (1992). A Desire to Be Taught: Instructional Consequences of Intrinsic Motivation. *Motivation and Emotion, 16*(3), 187–208.

Lepper, M.R., & Malone, T.V. (1987). Intrinsic Motivation and Instructional Effectiveness in Computer-Based Education. In R.E. Snow & M. J. Farr (Eds.), Aptitude, Learning, and Instruction, vol. 3: *Cognative and Affective Process Analyses*. Hillsdale, NJ: Lawrence Erlbaum, Inc.

Lipnevich, A., & Smith, J. (2008). *Response to Assessment Feedback: The Effect of Grades, Praise, and Source of Information* (ETS RR-08-30). Princeton, NJ: ETS.

Meece, J. L., Anderman, E.M., & Anderman, L.H. (2006). Classroom Goal Structure, Student Motivation, and Academic Achievement. *Annual Revue of Psychology, 57*, 487–503.

Myers, M. (2003). Foreward. In M.D. Shermis & J. Burstein (Eds.), *Automated Essay Scoring: A Cross-Disciplinary Perspective*. Mahwah, New Jersey: Lawerence Erlbaum Associates, Publisher.

National Commission on Writing. (2003). *The Neglected "R": The Need for a Writing Revolution*. New York, NY: College Entrance Examination Board.

Olson, C.B. (2003). *The Reading/Writing Connection: Strategies for Teaching and Learning in the Secondary Classroom*: Pearson Education, Inc.

Page, E. (1967). *Statistical and linguistic strategies in the computer grading of essays*. Paper presented at the 1967 International Conference On Computational Linguistics.

Page, E. (1994). Computer Grading of Student Prose Using Modern Concepts and Software. *Journal of Experimental Education, 62*(2), 127–142.

Pajares, F. a. (2001). Toward a Positive Psychology of Academic Motivation. *The Journal of Educational Research, 95*(1), 27–35.

Phillips, S.M. (2007). *Automated Essay Scoring: A Literature Review*: Society for the Advancement of Excellence in Education.

Powers, D.E., Burstein, J.C., Chodorow, M., Fowles, M.E., & Kukich, K. (2001). *Stumping E-Rater: Challenging the Validity of Automated Essay Scoring*, (No. GRE Board Professional Report No. 98-08bP, ETS Research Report 01-03).

Quinlan, T., Higgins, D., & Wolff, S. (2009). *Evaluating the Construct Coverage of the E-rater Scoring Engine* (ETS RR-09-01). Princeton, NJ: ETS.

Riedel, E., Dexter, S.L., Scharber, C., & Doering, A. (2005, April 11–15, 2005). *Experimental Evidence on the Effectiveness of Automated Essay Scoring in Teacher Education Cases*. Paper presented at the 86th Annual Meeting of the American Educational Research Association, Montreal, Canada.

Rosé, C.P., Roque, A., & Bhembe, D. (2002). *An Efficient Incremental Architecture for Robust Interpretation*. Paper presented at the Human Language Technology, San Diego, CA.

Rothermel, B.A. (2007). Machine Scoring of Human Essays. In P.F. Ericsson & R. Haswell (Eds.), *Machine Scoring of Student Essays: Truth and Consequences* (pp. 199–210). Logan, Utah: University of Utah Press.

Scardamalia, M., & Bereiter, C. (1986). Research on written composition. In C. Wittrock (Ed.), *Handbook of research on teaching* (3rd ed.) (pp. 778–803). New York, NY: Macmillan.

Scharber, C., Dexter, S., & Riedel, E. (2008). Students' Experiences with an Automated Essay Scorer. *Journal of Technology, Learning, and Assessment,* 7(1), 4–44.

Schroeder, J., Grohe, B., & Pogue, R. (2008). The Impact of Criterion Writing Evaluation Technology on Criminal Justice Student Writing Skills. *Journal of Criminal Justice Education 19*(8), 432–445.

Shapiro, M.A., & McDonald, D.G. (1992). I'm Not a Real Doctor, but I Play One in Virtual Reality: Implications of Virtual Reality for Judgments about Reality. *Journal of Communication, 92–114*, 92–114.

Shermis, M.D., & Burstein, J. (Eds.). (2003). *Automated Essay Scoring: A Cross-Disciplinary Perspective*. Mahwah, New Jersey: Lawerence Erlbaum Associates, Publisher.

Shermis, M.D., Burstein, J.C., & Bliss, L. (2004). *The Impact of Automated Essay Scoring on High Stakes Writing Assessments*. Paper presented at the Annual Meetings of the National Council on Measurement in Education, San Diego, CA.

Shermis, M.D., Garvan, C.W., & Diao, Y. (2008). *The Impact of Automated Essay Scoring on Writing Outcomes*. Paper presented at the Paper presented at the Annual Meetings of the National Council on Measurement in Education.

Sommers, N. (1980). Revision Strategies of Student Writers and Experienced Adult Writers. *College Composition and Communication, 31*, 378–388.

Spandel, V. (2008). *Creating Writers Through 6-Trait Writing Assessment and Instruction* (5th Edition): Allyn & Bacon.

Strauss, A.L., & Corbin, J.M. (1998). *Basics of Qualitative Research: Techniques and Procedures for Developing Grounded Theory*. Thousand Oaks, CA: SAGE Publications.

Suhr, K.A., Hernandez, D., Grimes, D., & Warschauer, M. (2010). Laptops and Fourth Grade Literacy: Assisting the Jump over the "Fourth Grade Slump". *Journal of Technology, Learning, and Assessment, 9*(5).

Tobin, L. (2002). Introduction: How the Writing Process Movement Was Born—and Other Conversion Narratives. In L. Tobin & T. Newkirk (Eds.), *Taking Stock: The Writing Process Movement in the '90's* (pp. 1–14). Portsmouth, New Hampshire: Boynton/Cook Publishers.

Vantage Learning, I. (2006). Vantage Learning's Award-Winning, Classroom-Proven Instructional Writing Program Now Available in Home Edition. Retrieved July 2, 2009, from http://www.vantagelearning.com/corporate/news/press_releases/20061203-1.html.

Vantage Learning, I. (2007). MY Access!® Efficacy Report. Newtown, PA: Vantage Learning, Inc.

Vernon, A. (2000). Computerized grammar checkers 2000: capabilities, limitations, and pedagogical possibilities. *Computers and Composition, 17*(3), 329-349. http://www.sciencedirect.com/science/article/B6W49-41Y87JP-6/2/f190de4406fd060ee927b588484dae03.

Vygotsky, L.S. (1986). *Thought and Language*. Cambridge, MA: Harvard University Press.

Vygotsky, L.S., Cole, M., John-Steiner, V., & Scribner, S. (1978). *Mind in Society: Development of Higher Psychological Processes*. Cambridge, MA: Harvard University Press.

Wang, J., & Brown, M.S. (2007). Automated Essay Scoring Versus Human Scoring: A Comparative Study. *Journal of Technology, Learning, and Assessment, 6*(1), 46.

Warschauer, M. (1996). *Motivational aspects of using computers for writing and communication*. Paper presented at the Telecollaboration in foreign language learning: Proceedings of the Hawai'i symposium Honolulu, HA, USA.

Warschauer, M. (2006a). *Laptops and Literacy*. Paper presented at the Annual Meeting of the American Educational Research Association, San Francisco, California.

Warschauer, M. (2006b). *Laptops and Literacy: Learning in the Wireless Classroom*. New York, NY: Teachers' College Press.

Warschauer, M., & Grimes, D. (2008). Automated Writing Assessment in the Classroom. *Pedagogies, 3*(1).

Williamson, D.M. (2009). *A Framework for Implementing Automated Scoring*. Paper presented at the Annual Meeting of the American Educational Research Association and the National Council on Measurement in Education, San Diego, CA.

Yin, R.K. (1989). *Case Study Research: Design and Methods* (2nd ed.). Newbury Park, London, New Delhi: Sage Publications.

Ziegler, W.W. (2007). Computerized Writing Assessment: Community College Faculty Find Reasons to say "Not Yet". In P. F. Ericsson & R. Haswell (Eds.), *Machine Scoring of Human Essays: Truth and Consequences* (pp. 138–153). Logan, Utah: Utah State University Press.

Author Biographies

Douglas Grimes is a software consultant and an independent researcher in educational technologies. He holds a B.A. in Economics from Yale, an M.S. in Management Information Systems from the University of Arizona, and a Ph.D in Information and Computer Sciences from the University of California, Irvine. He spent 15 years as a software developer and four years conducting educational research, focusing on one-to-one laptop programs and classroom use of automated writing evaluation software. His primary research interests are technologies for middle school literacy and informal science learning through video games. He can be contacted at doug@douglasgrimes.com.

Mark Warschauer is professor of education and informatics at the University of California, Irvine, and director of the Digital Learning Lab at the university. He is also director of UCI's Ph.D. in Education program, which includes a specialization in Language, Literacy, and Technology. His most recent book is Laptops and Literacy: Learning in the Wireless Classroom (Teachers College Press, 2006). He can be contacted via his website at http://www.gse.uci.edu/markw.

J·T·L·A

The Journal of Technology, Learning, and Assessment

www.jtla.org

Technology and Assessment Study Collaborative
Caroline A. & Peter S. Lynch School of Education, Boston College