Language Teaching Research 10,2 (2006); pp. 1–24

Automated writing evaluation: defining the classroom research agenda

Mark Warschauer *University of California* and
Paige Ware *Southern Methodist University*

With the advent of English as a global language, the ability to write well in English across diverse settings and for different audiences has become an imperative in second language education programmes throughout the world. Yet the teaching of second language writing is often hindered by the great amount of time and skill needed to evaluate repeated drafts of student writing. Online *Automated Writing Evaluation* programmes have been developed as a way to meet this challenge, and the scoring engines driving such programmes have been analysed in a considerable array of psychometric studies. However, relatively little research has been conducted on how AWE is used in the classroom and the results achieved with such use. In this article, we analyse recent developments in automated writing evaluation, explain the bases on which AWE systems operate, synthesize research with these systems, and propose a multifaceted process/product research programme on the instructional use of AWE. We explore this emerging area of inquiry by proposing a range of potential questions, methodologies and analytical tools that can define such a research agenda.

I Writing: the neglected R

The backdrop to the development of *Automated Writing Evaluation* (AWE; also referred to as *Automated Essay Scoring*[1]) is the changing role and importance of English language writing. Approaches to teaching English arise from the broader social, economic, cultural, and political contexts that shape the needs for using the language (Warschauer, 2000). Thus prior to the second world war, much foreign language instruction in English focused on reading, as this represented a principal use of foreign languages in many countries (Brown, 2001). In the postwar

Address for correspondence: Mark Warschauer, Dept. of Education and Dept. of Informatics, 2001 Berkeley Place, University of California, Irvine, CA 92697-5500, USA; e-mail: markw@uci.edu

 10.1191/1362168806lr190oa

period, with increased migration and cross-border travel, and the corresponding need for more oral fluency in foreign languages, audio-lingual and communicative approaches gained in popularity (Littlewood, 1981; Medgyes, 1986). Greater attention to second language writing began in the 1980s, and has continued to crescendo until today, due in large measure to the linguistic demands of a globalized informational society. With the Internet serving as a battering ram against economic and social boundaries, the educated urban population in much of the world now needs to write in English for vocational or professional purposes.

An increased emphasis on writing is also seen within the English-speaking countries, as educational leaders increasingly recognize the connections between writing and knowledge production in an information era, as well as the challenges of teaching writing to populations with large and growing numbers of immigrants (The National Commission on Writing, 2003). Reflecting this shift in emphasis, the US SAT exam for college admissions now requires a timed essay, as do state high school exit exams in California and elsewhere. With proficiency in English language writing being used as a gatekeeper for exit from high school and entry into institutions of higher learning, this places particular challenges in the way of English language learners and those who teach them.

Unfortunately, the need for expanded and improved instruction in English-language writing is seldom matched by the capacity of educational institutions to offer this instruction. Since effective writing instruction involves providing individual feedback on multiple drafts per student, it is extraordinarily time-consuming. A typical secondary teacher in most countries will have well over 100 students in their various courses. Numbers of students in university instruction vary widely, but large class sizes are the norm in many EFL contexts. In addition, the ability to provide fair and detailed feedback on writing demands a level of skill and training that is beyond the capacity of many instructors.

II Automated Writing Evaluation

AWE systems have been under development since the 1960s, when a national network of US universities, known as the College Board, supported the development of Project Essay Grade to help score thousands of high school student essays (Page, 2003). According to Page, who was

involved with the development, initial enthusiasm for the results was tainted by practical considerations; in the 1960s computer technology was simply not stable enough or accessible enough to expand into a larger scale. When microcomputers were introduced in the 1980s, interest in Project Essay Grade was once again renewed, and a second product was brought to the market: the Writer's Workbench (MacDonald *et al.*, 1982). The Writer's Workbench did not score essays; rather, it took steps towards providing feedback to writers on their writing quality. Although technology at that time operated on a narrow definition of quality, as it only allowed for flagging misused or misspelled words and for identifying long or short sentences, the Writer's Workbench pointed the field in an important direction: focus on feedback.

The current global emphasis on writing instruction, combined with technological developments of the last twenty years, has now allowed AWE software to achieve commercial viability. Three main companies – all based in the USA – are promoting AWE products (see Table 1), and a comparative examination of their programmes and scoring engines will provide an overview of the state of AWE today.[2] Two of the three companies, Vantage Learning and Educational Testing Services (ETS), have developed commercial products that make use of their scoring engines, while the third company leases their engine to commercial partners for inclusion in a variety of projects. Products from all of the companies are able to provide both numerical scores and some type of evaluative feedback. The scores are usually comparable to that provided

Table 1 Automated Writing Evaluation software

Company	Software Engine	Evaluation Mechanism	Commercial Product	Scoring	Feedback
Vantage Learning	Intellimetric	Artificial Intelligence	MY Access!	Holistic and component scoring	Limited individualized feedback
Educational Testing Service	E-Rater and Critique	Natural Language Processing	Criterion	Single holistic score	Wide range of individualized feedback
Pearson Knowledge Technologies	Intelligent Essay Assessor	Latent Semantic Analysis	Holt Online Essay Scoring (and others)	Holistic and component scoring	Limited individualized feedback

by human raters (see discussion later in this paper). The feedback varies in quality, but none of it is remotely similar to what a trained writing instructor can provide, and the ability of students to make use of the feedback is also questionable.

MY Access!™, developed by Vantage Learning, is a Web-based programme that has been used in public schools across the United States for several years. MY Access! uses an artificial intelligence scoring engine called Intellimetric™ to analyse some 300 semantic, syntactic and discourse features of an essay and compare them to features of sample essays that were previously scored by humans (Elliot, 2003). The Intellimetric engine does not claim to independently evaluate either the content or the organization of an essay, but rather to determine its similarity along a number of dimensions to the sample essays. Drawing on a set of about 300 human-scored essays on the same prompt, MY Access! provides a holistic score on a 1–6 or 1–4 scale, as well as individual scores on the same scale for focus and meaning, organization, content and development, language use and style, and mechanics and convention (though experience with the programme suggests that the individual scores are largely dependent on the overall score, with little variation among them). Generic feedback is then provided based on grade level, genre, and score; a fourth grade student who gets a score of 2 for organization on a persuasive essay will receive the same generic feedback as anyone else of the same score and level (e.g. pay attention to introduction and conclusion). A separate tool called *My Editor* provides advice on spelling, grammar and word usage, similar to that provided by Microsoft Word's spelling and grammar check, though in greater detail.

Criterion℠, developed by Educational Testing Services (ETS), is another well-known Web-based automated evaluation programme (see overview in Burstein *et al.*, 2004). Criterion's *e-rater*® software engine was developed in the mid-1990s with the primary function of scoring online essays and has been used since 1999 to score the essay portion of the Graduate Management Admissions Test (Burstein, 2003; Kukich, 2000). In such high-stakes exams, e-rater is used as one of two raters (the other a human rater), and any discrepancies in scoring are automatically sent to a human rater to resolve the score. For contexts with lower stakes, such as Web-based systems whose primary function is to provide students with opportunities to practise their writing, only a single reading by an automated system is required. The e-rater engine compares

essays along a number of dimensions to essays that had previously been scored by humans (Attali and Burstein, 2004; Burstein *et al.*, 1998; Shermis *et al.*, 2005). Specifically, e-rater analyses the rate of errors in grammar, usage, mechanics and style; the number of required discourse elements (such as thesis statement, main idea, or supporting idea); the lexical complexity (defined by type-token ratio, i.e. the number of unique words divided by the number of total words); the relationship of vocabulary used to that used in top-scoring essays on the same prompt; and the essay length (Attali and Burstein, 2004; Chodorow and Burstein, 2004). The Natural Language Processing (NLP) techniques in e-rater and a companion tool called *Critique* allow Criterion to make a broad range of linguistic diagnoses, and thus provide individualized feedback not only on spelling, grammar and word usage, but also on style, organization and development. Criterion can thus be used to provide feedback (albeit without numerical scores) on any writing, even in response to prompts that have not previously been normed by human scores; this is in contrast to MY Access!, which is designed to be used only with previously normed prompts.

A third main scoring engine, the *Intelligent Essay Assessor*™ (IEA), was originally developed at the University of Colorado and has recently been purchased by Pearson Education. IEA is not incorporated into a single flagship commercial product, such as MY Access (incorporating Intellimetric) or Criterion (incorporating e-rater). Rather, IEA is leased to commercial partners and included in a variety of products, such as the Holt Online Essay Assessor. (A somewhat different product incorporating IEA is Summary Street, which evaluates the summaries that students write and submit online by comparing them semantically to the original texts; see Kintsch *et al.*, 2002; Steinhardt, 2001.) IEA makes use of a distinct procedure known as *latent semantic analysis*, in which the semantic meaning of a given piece of writing is compared to a broader corpus of textual information (Landauer *et al.*, 2000). Using such analysis, IEA trains the computer on an informational database of textbook material, sample essays, or other sources rich in semantic variation surrounding a particular essay. The system is trained to focus specifically on evaluating conceptual content, and less attention is paid to mechanical correctness, stylistic control, or organizational structure (Landauer *et al.*, 2003). In comparison with Intellimetric and e-rater, IEA requires fewer human-scored essays in its training set because

scoring is accomplished based on semantic analysis rather than statistical comparison to a large set of previously scored work (Landauer *et al.*, 2003). The system maps student work against the informational database and identifies a wide range of acceptable semantic possibilities in the student responses. After the student work is compared against this body of data, the degree of relatedness can be converted into either a holistic score or into comments (Foltz *et al.*, 2000).

The main AWE programmes, such as MY Access! and Criterion, contain a wide range of other resources and features beyond automated scoring and feedback, including model essays, scoring rubrics, graphic organizers, dictionaries and thesauri. The programmes also generate a variety of progress reports for students and teachers, indicating individual and group scores in different areas and needs for further instructional attention. Since all three of the main AWE software engines are of similar quality in reliability of scoring (see further discussion below), it is often these added resources and features–plus the availability of previously normed prompts suitable for a particular population–that help make a specific product appealing.

None of these products were designed originally for ESL or EFL students, but all are being developed and marketed with those populations in mind. MY Access!, for example, allows the instructor to vary the language or level of the feedback provided so that English language learners can receive feedback in Spanish, Chinese, or a simplified English. And Criterion has a number of prompts designed to prepare students for the Test of English as a Foreign Language (TOEFL). The provision of such multilingual feedback systems might facilitate students' ability to interpret the feedback, but the question still remains open for empirical investigation about how effective such AWE software is in its ability to provide accurate, useful information for beginning-level writers from second language backgrounds. Translations, while useful, do not necessarily address the underlying issue of how best to support the development of ESL or EFL student writing with quality, fine-grained feedback.

Finally, as noted earlier, each of the three main companies developing and distributing AWE products is based in the USA. This is not surprising, given the large ESL market in the USA, as well as the country's technological edge in areas related to digital media and artificial intelligence. However, due to the huge global need for EFL writing instruction,

it is likely that these products will soon diffuse elsewhere, either through the international reach of companies like ETS, or through competitive products being developed elsewhere.

III Research on AWE

Research to date on AWE has largely been funded by the companies that have produced commercial AWE products and carried out by the staff of these companies. This is not unexpected as early studies of software use are often commissioned by firms as part of the research–development cycle and to support launching and refinement of a product. Once products gain a hold in the marketplace and thus begin to have a broader educational impact, the broader scholarly community starts to pay attention and get engaged in research. We expect that this transition–from principally vendor-sponsored research to independent research–will occur with AWE as it has occurred with other commercial applications, and, indeed, this article is intended to help stimulate such independent research. In the meantime, readers should bear in mind that most of the studied described below have been carried out with industry funding or participation. In addition, though these studies have often been presented at conferences, only rarely have they been submitted to the more rigorous peer review of academic journal publication. For all these reasons, research conducted to date should be considered with a highly critical eye.

IV Psychometric research on AWE

Most industry-sponsored research has focused on psychometric issues, and in particular on the validity of AWE programmes as determined by their correlations with scores given by human evaluators. The main way to validate AWE scores is by comparing the correlations between computer-generated and human rater scores to the correlations attained from two human raters. Based on this measure, most AWE programmes fare well (see summaries in Cohen *et al.*, 2003; Keith, 2003), with correlations with a single human judge usually ranging in the 0.80 to 0.85 range, approximately the same range as correlations between two human judges (see Table 2). This means that a computer-generated score will either agree with or come within a point of a human-rated score more

Table 2 Correlations between computer and human reader scores

Engine	Essay type	Sample size	Human–human correlation	Human–computer correlation
Intellimetric	K-12 Norm-Referenced Test (NRT)	102	0.84	0.82
Intelligent Essay Assessor	Graduate Management Admissions Test (GMAT)	1363	0.86–0.87	0.86
E-Rater	GMAT (13 prompts)	500–1000 per prompt	0.82–0.89	0.79–87

Source: Adapted from Cohen, et al., 2003

than 95% of the time, about the same rate of agreement as that between two human judges (Chodorow and Burstein, 2004; Elliot and Mikulas, 2004). These validity studies have for the most part taken place on large-scale standardized tests. The human–computer interrater reliability is expected to be lower in classroom contexts, where the actual content of student writing is likely to be more important than for the standardized tests (see discussion in Keith, 2003).

Researchers affiliated with ETS have also investigated whether computer-generated scores correlate with a broad range of non-test measures, such as Grade Point Average in writing courses, self-comparison in writing ability with peers, reported success with various kinds of writing, professors' grades of writing samples, and reported accomplishments in writing (Powers *et al.*, 2002b). They found that e-rater scores tended to correlate with these non-test criteria less strongly than did the scores assigned by human readers, but that the rank order of correlations with the non-test indicators was similar for e-rater and human-based scoring. This study thus supported the notion that the further one moves from standardized tests, the more comparative strengths there are to human readers–but that, nevertheless, computerized scores retain a good deal of validity.

ETS staff have also published a number of studies investigating the reliability of e-rater or Criterion in evaluating particular aspects of writing, such as coherence (Higgins, 1988), discourse structure (Burstein *et al.*, 2003), vocabulary usage (Burstein and Wolska, 2003; Leacock

and Chodorow, 2001), content (Shermis *et al.*, 2005) and grammar (Chodorow and Leacock, 2000). These technical studies are of special value to researchers in the area of natural language processing, computational linguistics and AWE software development, and are also useful for helping educational researchers understand better the evaluatory mechanisms involved. For example, the Shermis *et al.* (2005) study found that content plays a relatively minor role in the overall score that Criterion assigns to essays, ranging from a contribution of 1% of the variance in scores of persuasive and expository essays to about 6% of the variance in scores of descriptive essays.

Another important psychometric issue is whether AWE software can be tricked. One study has shown that expert writers can fool AWE software programmes and get relatively high scores on polished nonsensical essays (Powers *et al.*, 2002a). However, as Shermis and Burstein (2003) point out, while a bad essay can get a good score, it takes a good writer to produce the bad essay to get the good score. They convincingly argue that it is almost impossible for a bad writer to produce the bad essay that gets the good score.

Finally, Cohen *et al.* (2003) examine the applicability of AWE software toward evaluated essays written in languages other than English. They found that the Intellimetric software engine was successful in scoring essays written in Hebrew, though not to the same degree of validity as in English (as measured by correlations of human–computer versus human–human scoring). The authors also point out that software based on latent semantic analysis, such as the Intelligent Essay Theory, should be quite promising for other languages, since the underlying procedures of such analysis are language blind, though they can be affected by structural features, such as the morphological complexity of a language. They conclude that 'the path to a working system of AES [automated essay scoring] in languages other than English seems quite arduous, but given the head-start in English and the amazing tools of modern NLP, it seems a realistic and achievable goal' (2003: 26). Researchers in Japan have taken up this challenge and developed the Automated Japanese Essay Scoring System (jess), which grades essays based on their rhetoric, organization and content (Ishioka and Kameda, 2004). Initial studies have shown that jess delivers scores roughly similar to those of e-rater when grading essays that have been translated from English to Japanese.

V A classroom research agenda

Most of the research to date has been funded or carried out by developers of AWE software so as to demonstrate that it 'works' either in the sense that its scores correlate with those of human raters or that particular features or aspects function as desired. As writing instructors are well aware, however, other useful forms of feedback such as peer assessment and dialogue do not fall neatly into the categories used by psychometric research. AWE software provides a specific form of feedback that may be useful for particular, but not all, pedagogical purposes. None the less, the type of research currently conducted is a prerequisite for classroom use of the software, as the use of automatically generated scores would be highly problematic if they were not at least minimally valid. However, while this psychometric research is thus a necessary condition for adopting AWE software in the classroom, it is far from sufficient for making maximal use of the software, or indeed, even for understanding if it 'works' in the broader sense of contributing to positive outcomes for student learning.

In the remainder of this paper, we focus on development of a classroom research agenda that can help evaluate and guide the use of AWE software for the teaching of writing.[3] In doing so, we analyse the few studies that have yet been done on instructional use of AWE, and then propose a range of research questions, approaches and methodologies for carrying this agenda forward. Drawing on a categorization developed by Long (1984), we divide the research possibilities into three areas: product, process and process/product. In this case, *product* refers to educational outcome (i.e. what results from using the software), *process* refers to learning and teaching process (i.e. how the software is used), and *process/product* refers to the interaction between use and outcome. Since none of the prior research has been conducted in exclusively second language contexts, we report on the general research, making note of the presence of second language learners in cases it was specifically referred to in prior studies.

VI Product research: does AWE improve writing?

The first question of most educational administrators – i.e. those responsible for deciding whether to purchase AWE systems and implement

them in their schools or districts – will be 'Does the software improve student writing?' A few preliminary studies have been conducted in this area. Elliot and Mikulas (2004) of Vantage Learning report on four outcome-based studies. In the first study, 25 ninth-grade students piloted MY Access! during a six-week summer session, writing an average of four essays with 2–5 revisions per essay. The students' overall writing scores improved from 2.00 on a four-point scale in the beginning of the programme to 2.84 at the end of the programme. The school district then extended used of MY Access! to all ninth and tenth grade students for an academic year. All five high schools in the district improved their Academic Performance Index (API) scores over the year, usually to a larger extent than scores had improved in previous years. However, since there were no control groups in either the pilot or follow-up study, and since the API measures a broad range of academic skills, including those in other subjects such as mathematics, it is difficult to assess the particular extent to which MY Access! use contributed to score improvement.

In the second study, 496 high school students took classes using MY Access! over an academic year and were compared to 306 whose classes did not use MY Access!. Division between the two groups was described as 'voluntary' (2004: 23), though the authors do not specify whether it was the teachers who volunteered to use the software or the students who volunteered to take courses from those teachers. Some 65% of the 812 students in total were English language learners, but no information is provided as to how many were in each group. At the end of the academic year, all 812 students took the California High School Exit Examination, with 81% of the MY Access! group passing as compared to 46% of the comparison group. However, with no information available as to which teachers or students volunteered for each group or of the baseline level of the students using MY Access! compared to those not using it, it is again difficult to assess what contributed to these differential results.

A third study compared 100 fifth-grade students using MY Access! and 95 counterparts not using MY Access!, again on a by-class voluntary distribution. There were actually three groups in all, as the MY Access! group was divided into those using the software 4–5 times a week and those using it 2–3 times a week. After five months of English language arts instruction with or without the software, the students took a district writing test, with 86% of the medium-use (2–3 times a week)

students passing, 67% of the high-use (4–5 times a week) students passing, and 43% of the none-use students passing. Again, the lack of random assignment or baseline scores makes the results difficult to interpret, though it is interesting to note that greater use of the software was not associated with higher test scores.

In the last of the studies reported on by Elliot and Mikulas, 709 eleventh grade students used MY Access! from September 2003 to March 2004 and then took a state school assessment that 91% of the students passed. The authors compared that passage rate to the year of 2000, in which only 76% of the school's ninth graders passed the exam. The authors offer no explanation why a comparison is made with students three years earlier, or from a different grade level, and the results reported thus raise more questions than answers.

In summary, though Elliot and Mikulas claim that 'students who use MY Access! tend to show greater improvement in writing and language arts skills that those students who do not use MY Access!' (2004: 30), their research is not methodologically sound enough to justify that claim. This raises the question of how to design a more rigorous study to measure outcomes from the use of AWE. Outcome-based research with the use of technology is traditionally very difficult to accomplish, since the introduction of new technology often changes the broader ecology of the classroom, making comparisons to a control group difficult (see discussion in Warschauer, 1998), and the outcomes of technology can be so broad that they are difficult to assess. However, in this case, what is being introduced is not computers *per se*, but a specific use of computers, and the desired outcome is the specific improvement of writing test scores, which is relatively easy to measure. Nevertheless, there are still substantial problems in carrying out true experiments in school settings, due in part to the difficulty in carrying out random assignment.

One methodologically rigorous study that did employ random assignment was carried out by Shermis *et al.* (2004). In this study, 1072 urban high school students were randomly assigned to a treatment group, which wrote on up to seven Criterion writing prompts (n = 537), or a control group, which participated in the same classes but did not use the Criterion software and instead completed alternate writing assignments during the class periods (n = 535). Following the treatment, students from both the treatment and control groups took a state writing

assessment. No significant differences were noted between the two groups. This outcome is not that surprising, due to the within-class random assignment design. Simply put, any classroom innovation, and especially those using technology, will likely have its best effect if it is fully integrated into instruction (see, for example, Warschauer, 1996). A teacher will only likely fully explain and support the use of technology with her students if all of them are using it; if, instead, only some are using it, those students will likely not be receiving the full range of explanation and support to use the software well. And indeed, in this study, teacher compliance was low, and as a result, most of the students failed to even complete all of the writing prompts, as discussed in the process/product section below.

What other designs are then possible? Random assignment of AWE software use to half the teachers in a school or district could provide a methodological solution, but would probably be very unappealing to both administrators and teachers. More realistic would be a regression-discontinuity design (Trochim, 2002). In this design, all the students who score below a certain level on writing or an English language arts test could be assigned to classes that were using AWE software, with students above the cut-off continuing traditional writing instruction. Regression analysis is then used to see if there is a discontinuity in score gains between the two groups at the cut-off point (i.e. do students just below the cut-off point assigned to the AWE classes indicate gains significantly greater than do students just above the cut-off point assigned to traditional classes?).

In most cases, though, administrators will prefer to implement the innovation with all the students in a school or district. In that case, a quasi-experimental design can be used, in which the test score gains in the school or district using AWE are compared to a similar group of students (e.g. all those in a comparable neighbouring school or district) over a given year, or to students in the same location in a prior year. One possible complicating factor in this methodology is the broader ecological effect of use of AWE software. For example, many districts purchase large amounts of new computers in order to implement use of AWE, thus reducing the student–computer ratio in the district. In such a case, care would need to be made to assess whether any possible gains in writing scores came from use of the AWE software or other uses of the new computers.

Though scores on a standardized writing test are likely to be the most important outcome measure, such scores are not unproblematic for evaluation of AWE software's impact. One of the main purposes of AWE software is to teach students to revise their work, but standardized writing tests are almost always based on submission of single drafts of timed essays. Correlation of gains in AWE scores with other writing outcome measures would thus be valuable. These might include scores on essays or portfolios that have gone through multiple revisions, or on non-test score data such as that described in Powers *et al.* (2002b), discussed above.

Finally, if test score data is rising, it is valuable to know what aspects of writing are improving, and how writing is changing overall. Standardized writing assessments are known to have a large washback effect on writing instruction (Hillocks, 2002). One of the main concerns voiced about AWE software is that it will further encourage students to 'write to the test' by focusing on those features that are most easily detected by the software (see, for example, discussion in Cheville, 2004). Though, as pointed out earlier, bad writers are unlikely to have the skills to trick AWE software, students of all levels may nevertheless consciously or unconsciously adjust their writing to attempt to meet the assessment criteria of the software. And teachers may feel pressured to lead them in this direction in the attempt to raise test scores (see statement opposing electronic scoring by the Conference on College Composition and Communication, 2004). This could result in writing that is correctly structured and mechanically accurate, but lacking creativity, originality, a clear purpose, or a sense of audience. Comparative examination of final writing samples of students who have and have not used AWE software, not only for the numerical scores achieved, but also for the quality and nature of the writing according to diverse measures, can allow a more nuanced understanding of what students learn about writing through the use of automated scoring programmes.

VII Process-research: how is AWE software used?

One of the major problems with any product-based educational research is that it leaves the educational process involved as a black box. It is often not very helpful to know that an innovation brought, or did not bring, results, if you have little information about how or even if the

innovation was implemented. Recently, for example, the Los Angeles Unified School District conducted a major evaluation of a $50 million implementation of a software programme for reading instruction. The failure of the programme to bring positive results was largely tied to the effect that it was not implemented by teachers to the extent expected (Slayton and Llosa, 2002).

Fortunately, AWE software itself yields a great deal of data indicating how the software is used. For example, MY Access! reports provide the number of essays and revisions submitted in a given time period; average student scores per essay both at the overall level and in five individual areas; the number of grammatical, mechanical and stylistic errors of specific types committed in the essays (or, for example, in original drafts compared to final revised versions); and the performance in each of these areas by student group (whether divided by class, school, grade level, economic status, ethnicity, participation in special programmes, or English language level). AWE programmes also save archives of all drafts and final versions of papers that have been submitted, thus providing another source of data for additional computer-based or human analysis.

A study by Attali (2004) gives a detailed account of how Criterion was used by sixth to twelfth graders throughout the USA during the 2002–2003 school year, and in particular what kinds of changes occurred through revision of essays. The dataset included 33,171 student essay submissions to Criterion of 50 or more words. A strong majority of the student essays (71%) had been submitted only one time without revision, indicating that the programme is not being used in practice in ways consistent with how AWE software is often touted (i.e. as a motivator and guide for revision). A total of 13% were submitted twice (an original and one revision), 8% were submitted three times, and the remaining 8% were submitted four or more times.

Comparing the first and last submissions for essays that had been revised, Attali found statistically significant improvement in the total holistic score (from 3.7 to 4.2 on a six-point scale) as well as in computerized evaluations of development, grammar, usage, mechanics and style. Scores in each of these areas rose consistently over the first five revisions. Students were able to significantly lower the rate of most of the 30 specific error types that were identified by the system and reduced their error rates by about one quarter. The greatest numbers of errors

corrected were in the areas of spelling and capitalization. When measured by effect size (defined as the percentage of errors in a particular category that were corrected), the greatest improvements were in garbled sentences, fused words, capitalization, confused words and wrong form of words. As for discourse elements, students in their revised essays significantly increased the rate of occurrence of background and conclusion elements, main points and supporting ideas.

Attali's study makes no differentiation between students by grade level, school, or language background. However, the software easily allows these types of analysis to be carried out. AWE programmes can yield a wealth of valuable data, for example, of the types of revisions that second language learners make in their writing compared to those of native speakers, or even simply the number of revisions carried out. Similar comparisons could be made between beginning and advanced second language learners, or learners from different native language backgrounds.

There is of course also much additional information that could be gained from direct observation. We have begun observations of K-12 classrooms using AWE software in five schools in California (see an early report on this research by Grimes, 2005). Our initial indications demonstrate that teachers use AWE software in diverse ways. For example, while some teachers encouraged students to post multiple revisions, others used AWE software for timed essay writing to simulate high-stakes testing situations. Most of the teachers integrated the assessment component of AWE, but they none the less provided hand-written feedback on student essays in addition to the electronic feedback. These uses reflect their underlying approach to writing as well as the pressures and constraints that they feel in their school and district.

Overall, though, Attali's finding that relatively few revisions are carried out has been confirmed by our own observations. This reflects the general classroom environment in the era of standards high-stakes testing, with teachers feeling pressured to cover as much material as possible. Little time is thus left for multiple revisions of a single essay, especially when the standardized tests that teachers are preparing their students for all involve tightly timed essays, thus obviating the need for much revision. In addition, many of the students that we have observed are not capable of understanding anything but the most simple of computer-generated feedback. When revisions do take place, students

commonly limit themselves to correcting as many as possible of the indicated spelling errors (though even this is difficult for limited-English students) and they ignore the narrative advice about improving development or organization. Our work in this area is just beginning. Much more systematic observation of AWE software being used in diverse social and linguistic contexts will yield more detailed accounts about how students make use of AWE. It can also yield information about instructional strategies, such as how teachers integrate AWE into the curriculum, how this differs by social context and student population, and what appear to be best practices with AWE.

Classroom observations can be supplemented by other forms of qualitative data collection to more fully explore the teaching and learning process with AWE. Think-aloud protocols (see discussion in Smagorinsky, 1994), in which students record their thoughts as they work with AWE programmes can shed additional light on the writing and revision process. Video documentation of student work, either from a video camera or from software that records movies of screen shots (e.g., Snapz Pro X, Ambrosia Software, 2005), can provide further details of individual student behaviour.

Surveys can document attitudes towards AWE among a large numbers of students or teachers (see Warschauer, 1996, for an example of a survey-based study of technology and writing) and can readily be conducted online since all users of AWE by definition have regular school computer and Internet access. Interviews with teachers can illuminate the pedagogical decisions they make with AWE software and can help explain, for example, why little revision takes place in many classrooms using the software.

Finally, longitudinal case studies of individual teachers or students (cf. Lam, 2000), or ethnographic studies of districts or schools or classrooms (cf. Warschauer, 1999), can provide a broader holistic view of what is lost and gained with AWE use over time. What types of students become most excited, or most turned off, by computer-generated feedback? How does the students' sense of identity develop as writers in these environments? Is the use of AWE software consistent with the development of reflective writers with a sense of audience and pupose? Does AWE relieve teachers of drudgery to allow more time for creativity, or does it serve to depersonalize the classroom environment? How do decisions to implement AWE intersect with broader goals of

teaching, learning and assessment? Qualitative attention to these issues, supplemented by the quantitative data that the AWE software itself yields, will give a much fuller view of the role of AWE in first and second language writing instruction.

VIII Process/product research

The most desirable form of language learning research involves a product–process approach (Long, 1984). In this way, it is possible to get inside the black box of classroom activity to analyse what the teaching and learning processes were, and then to tie these processes to specific learning outcomes.

The third of four studies reported in the Elliot and Mikulas (2004) paper discussed above make a small step in this direction by reporting test outcomes in relationship to how often the software was used per week. A more detailed example of product–process research on this topic is seen in the study by Shermis *et al.* (2004), also mentioned earlier. The study reports, for example, that only 112 of the 537 treatment students completed all seven prompts, and that reasons for this low level of completion included teacher non-compliance, student absences, problems in scheduling the laboratories, and changes in the high school level. (A learning curve developed by the researchers predicted that if students had written approximately five more writing assignments each, differences in performance would have been significant.) The study also attempted to analyse correlations between scores on the individual prompts, as well as between individual prompt scores and score on the final writing test, and also examine the types of error corrections made. Though the particular results found in the study are problematic for understanding the full potential of AWE (due to non-compliance issues arising in part from the problems of a mixed in-class design discussed earlier), the study does point to ways that the statistical data yielded by AWE software can be used in process/product research.

There are numerous questions that could be addressed from a process/product perspective. Making use of AWE software reports, test score data, and demographic data, researchers could address whether the total number of revisions that students carry out correlate with higher final scores on a separate outcome test, whether the types of revisions made correlate with higher final scores, or whether the relationship

between revisions and scores on a final outcome measure test is mediated by students' language background, grade level, beginning writing level, or socioeconomic status. Adding observation, surveys, and/or interviews, researchers could also assess whether particular approaches to integrating AWE software in instruction by teachers tend to correlate with greater improvements in writing test scores, or whether different learning styles and strategies by individual students in their use of AWE correlate with writing score improvement.

IX Conclusion

Douglas Reeves, in his investigation of highly successful schools with traditionally at-risk students, claims that key factors enabling such success include a razor-like focus on measuring educational achievement, assessments that are frequent and external to the classroom teacher, multiple opportunities for student success, and a strong focus on writing development as a lever for overall academic gains (Reeves, 2002). The use of AWE software appears to be custom-made to meet all these criteria. It is thus no surprise that a number of educational technology enthusiasts are looking to AWE as a silver bullet for language and literacy development.

At the same time, a hundred-year history of introducing educational technology in the classroom – from film to radio to television to the language laboratory – should teach us that there is no such thing as a silver bullet (for a historical overview, see Cuban, 1986). Every technology brings with it both good and bad. And technologies do not exert an 'impact' but instead help reshape the broader ecology of the classroom and school (Warschauer, 1998). High-stakes testing regimes themselves are reshaping teaching and learning, and software that provides automated writing assessment may amplify those tendencies, drawing students away from composition that is purposeful, audience focused, and expressive toward that which is formulaic and stilted. As Hamp-Lyons (in press) explains:

> The contexts and needs of classrooms and teachers are not the same as those of large scale testing. The large scale needs to discriminate, to separate, to categorize and label. It seeks the general, the common, the group identifier, the scaleable, the replicable, the predictable, the consistent, the characteristic. The teacher, the classroom, seeks the special, the individual, the changing, the changeable, the surprising, the subtle, the textured, the unique. Neither is *better* but they *are* different.
>
> (italics in orginial)

Our own prior work has focused on examining the new literacies involved in online writing and multimedia development for real audiences. We expect that many educators carrying out research on these 'electronic literacies' (Warschauer, 1999: 1) will be wary of automated writing assessment, which at face value seems to negate much of what we celebrate as liberating in new media use. But at the same time we are cognizant of the high stakes for both ESL and EFL students in attaining English language writing proficiency, and we are painfully aware of the steep odds that learners face in reaching this goal. We thus think that tools and approaches that promise to promote academic literacy by providing frequent, individualized feedback on student writing ought to be considered, but not without keeping a critical eye turned onto carefully designed research.

In summary, we believe that both of the above-described potentials – technology that empowers by providing instant evaluation and feedback, and technology that dehumanizes by eliminating the human element – exist in automated writing evaluation software. Where on the continuum between these two outcomes AWE might fall depends on how such software is used in specific teaching and learning contexts, a matter subject to empirical investigation. Well-designed product-, process- and product/process studies of AWE use in the classroom can help educators make informed decisions about utilization of such programmes, and can also signal developers on ways that AWE software should be improved. May the research continue.

Acknowledgements

We thank Jill Burstein, Doug Grimes, and Mark Shermis for their thoughtful comments on an earlier draft of this paper.

Notes

[1] The term automated Essay Scoring is used more broadly and reflects the historical development of this software, which originally was designed to score essays. The programs used today handle not only essays but also other types of writing, and provide not only numerical scores but also other forms of feedback. We thus prefer the term Automated Writing Evaluation.

[2] Other programs may be in existence or development, but we focus on those that are currently most widely commercially available.

[3] By *classroom research*, we mean any research done on the classroom use of AWE, whether or not that research incorporates direct classroom observation.

X References

Attali, Y. 2004: Exploring the feedback and revision features of Criterion. Paper presented at the National Council on Measurement in Education conference, April 2004, San Diego, CA.

Attali, Y. and **Burstein, J.C.** 2004: Automated essay scoring with E-rater 2.0. Paper presented at the Conference of the International Association for Educational Assessment, June 2004, Philadelphia, PA.

Brown, D. 2001: *Teaching by principles: an interactive approach to language pedagogy*. Addison Wesley Longman.

Burstein, J., Kukich, K., Wolff, S., Lu, C., Chodorow, M., Braden-Harder, L. and **Harris, L.** 1998: Automated scoring using a hybrid feature identification technique. Paper presented at the Annual Meeting of the Association of Computational Linguistics, August 1998, Montreal.

Burstein, J.C. 2003: The e-rater scoring engine: automated essay scoring with natural language processing. In Shermis, M.D. and Burstein, J.C. editors, *Automated essay scoring: a cross-disciplinary perspective.* Lawrence Erlbaum, 113–21.

Burstein, J.C., Chodorow, M. and **Leachock, C.** 2004: Automated essay evaluation: the Criterion online writing service. *AI Magazine* 25(3): 27–36.

Burstein, J.C., Marcu, D. and **Knight, D.** 2003: Finding the WRITE stuff: automatic identification of discourse structure in student essays. *IEEE Intelligent Systems* 18(1): 32–39.

Burstein, J.C. and **Wolska, M.** 2003: *Toward evaluation of writing style: finding repetitive word use in student essays.* Paper presented at the Proceedings of the 11th Conference of the European Chapter of the Association for Computational Linguistics, April 2003, Budapest.

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

Chodorow, M. and **Burstein, J.C.** 2004: *Beyond essay length: evaluating e-rater's performance on TOEFL essays* (Research Report No. 73). Educational Testing Service.

Chodorow, M. and **Leacock, C.** 2000: *An unsupervised method for detecting grammatical errors.* Paper presented at the Conference of the North American Chapter of the Association for Computational Linguistics, April–May 2000, Seattle.

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

Conference on College Composition and Communication 2004: *CCCC position statement on teaching, learning and assessing writing in digital environments*. Retrieved 6 January 2006 from <http://www.ncte.org/groups/cccc/positions/115775.htm>

Cuban, L. 1986: *Teachers and machines: the classroom use of technology since 1920*. Teachers College Press.

Elliot, S. 2003: IntelliMetric: from here to validity. In Burstein, J.C., editor, *Automated essay scoring: a cross-disciplinary perspective.* Lawrence Erlbaum, 71–86.

Elliot, S. and **Mikulas, C.** 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, April 2004, San Diego, CA.

Foltz, P., Gilliam, S. and **Kendall, S.** 2000: Supporting content-based feedback in on-line writing evaluation with LSA. *Interactive Learning Environment* 8(2): 111–29.

Grimes, D. 2005: Assessing automated assessment: essay evaluation software in the classroom. Proceedings of the Computers and Writing Conference, Stanford, CA.

Hamp-Lyons, L. (in press): The impact of testing practices on teaching: ideologies and alternatives. In Cummins, J. and Davison, C., editors, *The international handbook of English language teaching*, Volume 1. Springer.

Higgins, J. 1988: *Language, learners and computers*. Longman.

Hillocks, G., Jr. 2002: *The testing trap: how state writing assessments control learning*. Teachers College Press.

Ishioka, T. and **Kameda, M.** 2004: *Automated Japanese essay scoring system: jess.* Proceedings of the 15th International Workshop on Database and Expert Systems Applications, Zaragosa, Spain, 4–8.

Keith, T.Z. 2003: Validity and automated essay scoring systems. In Shermis, M.D. and Burstein, J.C., editors, *Automated essay scoring: a cross-disciplinary perspective.* Lawrence Erlbaum, 147–67.

Kintsch, E., Steinhart, D., Stahl, G., LSA Research Group, Matthews, C. and **Lamb, R.** 2002: Developing summarization skills through the use of LSA-based feedback. *Interactive Learning Environment* 8(2): 87–109.

Kukich, K. 2000: Beyond automated essay scoring. *IEEE Intelligent Systems* 15(5): 22–27.

Lam, W.S.E. 2000: Second language literacy and the design of the self: a case study of a teenager writing on the Internet. *TESOL Quarterly* 34: 457–82.

Landauer, T., Laham, D. and **Foltz**, **P.** 2000: The Intelligent Essay Assessor. *IEEE Intelligent Systems* 15(5): 27–31.

Landauer, T., Laham, D. and **Foltz, P.** 2003: Automated scoring and annotation of essays with the Intelligent Essay Assessor. In Shermis, M.D. and Burstein, J.C., editors, *Automated essay scoring: a cross-disciplinary perspective*. Lawrence Erlbaum Associates, 87–112.

Leacock, C. and **Chodorow, M.** 2001: *Automatic assessment of vocabulary usage without negative evidence* (Research Report No. 67). Educational Testing Service.

Littlewood, W. 1981: *Communicative language teaching*. Cambridge University Press.

Long, M.H. 1984: Process and product in ESL programme evaluation. *TESOL Quarterly* 18(3): 409–25.

MacDonald, N.H., Frase, L.T., Gingrich, P.S. and **Keenen, S.A.** 1982: Writer's workbench: computer aid for text analysis. *IEE Transactions on Communications* 30(1): 105–10.

Medgyes, P. 1986: Queries from a communicative teacher. *ELT Journal* 40(2): 107–12.

Page, E.B. 2003: Project essay grade: PEG. In Shermis, M.D. and Burstein, J.C., editors, *Automated essay scoring: a cross-disciplinary perspective.* Lawrence Erlbaum, 43–54.

Powers, D.E., Burstein, J.C., Chodorow, M., Fowles, M.E. and **Kukich, K.** 2002a: Stumping e-rater: challenging the validity of automated essay scoring. *Computers in Human Behavior* 18: 103–34.

Powers, D.E., Burstein, J.C., Chodorow, M.S., Fowles, M.E. and **Kukich, K.** 2002b: Comparing the validity of automated and human scoring of essays. *Journal of Educational Computing Research* 26(4): 407–25.

Reeves, D. 2002: *Accountability in action*. Advanced Learning Press.

Shermis, M.D. and **Burstein, J.C.** 2003: Introduction. In Shermis, M.D. and Burstein, J.C., editors, *Automated essay scoring: a cross-disciplinary perspective*. Lawrence Erlbaum.

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

Shermis, M.D., Shneyderman, A. and **Attali, Y.** 2005: How important is content in the ratings of essay assessments? Paper presented at the annual meeting of the National Council of Measurement in Education, March 2005, Montreal.

Slayton, J. and **Llosa, L.** 2002: *Evaluation of the Waterford Early Reading Programme 2001–2002: Implementation and student achievement* (Planning, Assessment and Research Division Publication No. 144). Los Angeles Unified School District Programme Evaluation and Research Branch.

Smagorinsky, P. 1994: *Speaking about writing: reflections on research methodology*. Sage.

Steinhardt, D.J. 2001: Summary Street: An intelligent tutoring system for improving student writing through the use of Latent Semantic Analysis. PhD thesis, University of Colorado, Boulder.

The National Commission on Writing 2003: *The neglected "R": the need for a writing revolution*. College Entrance Examination Board.

Trochim, W.M.K. 2002: *The regression-discontinuity design*. Retrieved 6 January 2006 from <http://www.socialresearchmethods.net/kb/quasird.htm>

Warschauer, M. 1996: Motivational aspects of using computers for writing and communication. In Warschauer, M., editor, *Telecollaboration in foreign language learning: proceedings of the Hawai'i symposium.* University of Hawai'i, Second Language Teaching and Curriculum Center, 29–46.

—— 1998: Researching technology in TESOL: determinist, instrumental, and critical approaches. *TESOL Quarterly* 32(4): 757–61.

—— 1999: *Electronic literacies: language, culture, and power in online education*. Lawrence Erlbaum Associates.

—— 2000: The changing global economy and the future of English teaching. *TESOL Quarterly* 34(3): 511–35.