Automated Essay Scoring: Writing Assessment and Instruction

Mark D. Shermis

The University of Akron

Jill Burstein

Derrick Higgins

Klaus Zechner

Educational Testing Service

Introduction

This chapter documents the advent and rise of automated essay scoring (AES) as a means of both assessment and instruction. The first section discusses what AES is, how it works, and who the major purveyors of the technology are. The second section describes outgrowths of the technology as it applies to on-going projects in measurement and education.

In 1973, the late Ellis Page and colleagues at the University of Connecticut programmed the first successful automated essay scoring engine, “Project Essay Grade (PEG)” (1973). The technology was foretold some six years earlier in a landmark *Phi Delta Kappan* article entitled, “The Imminence of Grading Essays by Computer” (Page, 1966). At the time the article was provocative and a bit outrageous, though in hindsight, it can only be deemed prophetic. As a former high school English teacher , Page was convinced that students would benefit greatly by having access to technology that would provide quick feedback on their writing. He also realized that the greatest hindrance to having secondary students write more was the requirement that, ultimately, a teacher had to review stacks of papers. While PEG produced impressive results, the technology of the time was too primitive to make it a practical application. Text had to be typed on IBM 80-column punched cards and read into a mainframe computer before it could be evaluated. As a consequence, the technology sat dormant until the early 1990s and was revitalized with the confluence of two technological developments: microcomputers and the Internet. Microcomputers permitted the generation of electronic text from a regular keyboard and the internet provided a universal platform to submit text for review (Shermis, Mzumara, Olson, & Harrington, 2001).

Automated essay scoring is a measurement technology in which computers evaluate written work (Shermis & Burstein, 2003). Most of the initial applications have been in English, but past work has been applied to Japanese (Kawate-Mierzejewska, 2003, March), Hebrew (Vantage Learning, 2001), and Bahasa Malay (Vantage Learning, 2002). Computers do not "understand" the written text being evaluated. So, for example, the computer would not "get" the following joke.

Q-When is a door not a door?

A-When it is ajar.

Unlike humans, a computer cannot interpret the play on words, and infer that the predicate in the answer (i.e., "ajar") is being cleverly used as a noun (i.e., "a jar").

What the computer does in an AES context is to analyze the written text into its observable components. Different AES systems evaluate different numbers of these components. Page and Peterson (1995) referred to these elements as "proxes" or *approximations* for underlying "trins" (i.e., *intrinsic* characteristics) of writing. It is the observable components that automated essay scoring engines, identify, computationally, and subsequently use to compute essay scores. AES statistical models are developed by weighting the various observable components as they relate to intrinsic characteristics of writing. For example, a model in the PEG system, might be formed by taking five intrinsic characteristics of writing (*content*, *creativity*, *style*, *mechanics*, and *organization*) and linking proxes. An example of a *simple prox* is *essay length*. The empirical evidence suggests that the longer an essay is, the more highly valued it is by a rater. This could be because the writer provides additional details which improves the essay's standing in the eyes of the grader. However, this relationship is not linear, but logarithmic. More specifically, it appears as if essay length is important up to a point, but beyond a certain theshold it carries little additional weight. Where that "point" is becomes a funtion of the average essay length. If most essays have 150 words and a candidate essay has 200 words, then the writer has probably already taken advantage of any contribution that essay length would influence in a rater's decision-making.

Again, in the PEG framework, an example of a *complex prox* might be a count of the number of times "because" is used in an essay. Though this counting method is admittedly superficial, it is an indicator of *sentence complexity*, and is also tied conceptually to the intrinsic characteristic of *style*. It is useful to note that this one prox is joined with a number of others to estimate the trin, and would not normally be used as a single indicator for the trait.

How are models formed? There are three general approaches, two of which involve the collection of empirical data, and one that is currently more of a theoretical option. In the first approach, two samples of essays are collected—one for model building and the other for model evaluation/confirmation--each of which has been rated by multiple raters. For the purpose of creating a model one might utilize the field distribution of ratings or select a fixed sample size (e.g., 300) with approximately an equal number of essays at each "cut" point. Using the later technique if there were six points on the rating rubric that raters used, the model would be most discriminating if there were 50 essays for each point of the scale (300/6 = 50). Because rating points at the extremes of the scale are more difficult to come by, one may need a pool of more than

500 essays from which to draw the data if using the field distribution of ratings. As alluded to previously, the prox (or prox cluster) variables are regressed against the essay ratings (or in the case where the model is not formulated in an *a priori* basis, to select the variables and optimize the weights). The validation sample (e.g., 200 cases) is used to evaluate the results from the first set of estimates. Most AES developers use multiple regression to create their models, but one developer uses multiple statistical techniques, and then selects the one that explains the most variance.

In the second approach, the evaluation of content may be accomplished through the specification of vocabulary (i.e., the evaluation of the other aspects of writing is performed as described above). Latent Semantic Analysis and its variants are employed by some developers to provide estimates as to how close the vocabulary in the candidate answer is to a targeted vocabularly set (Landauer, Foltz, & Laham, 1998).

The third approach is to develop models that are based on a "gold standard" formulated by experts. To date this mechanism for developing models is more theoretical than applied. If normative models can be created for the relevant dimensions of writing (i.e., age and writing genre), then other variables could be tailored on an *a priori* basis to generate a statistical blueprint for the ideal response. The blueprint may or may not be aligned with human ratings. For example, the guidelines for a few high-stakes writing programs explicity direct raters to ignore expressions of non-standard English. However, raters find this a difficult challenge, even when exposed to comprehensive training programs. When presented with an expression of non-standard English, a typical rater will inevitably undervalue the essay even though an answer may be functionally equivalent to a response given in standard English. AES would have the capacity to overcome this human limitation *if* the relevant affected variables associated with non-standard English can be isolated and adjusted.

AES Programs

Presently there are three major developers of automated essay scoring. The Educational Testing Service (ETS) has *e-rater*® which is a component of *Criterion*^{SM}, a comprehensive electronic portfolio administration system (http://www.ets.org/criterion). E-rater is also used as a scoring application for high- and low-stakes assessments, as well as a number of test practice applications. Vantage Learning has developed *Intellimetric*™ which is also part of an electronic portfolio administration system called *MyAccess!*^{TM} (http://www.vantagelearning.com). Finally, Pearson Knowledge Technologies supports the *Intelligent Essay Assessor*™ which is used by a variety of proprietary electronic portfolio systems (http://www.pearsonkt.com). All of the products have the capacity to receive text via a web page and return feedback to both a student user and comprehensive data base that may be accessed by teachers. In the paragraphs below, a short description is given that illustrate the kinds of factors/dimensions/variables used in building AES scoring models. References are provided for a more comprehensive descripiton of the process.

The construction of *e-rater v. 2.0*[1] models is given in detail in Attali and Burstein (2006). It is composed of up to 12 features used by e-rater v.2.0 to score essays. These 12 features are associated with six areas of analysis: errors in grammar, usage, and mechanics (Leacock & Chodorow, 2003); style (Burstein, 2003); identification of

[1] E-rater v.2.0 is the infrastructure for all subsequent version of e-rater. Currently, e-rater v.9.1 has been released. Some of the mathematical transformations implemented in the earlier e-rater v.2.0 have been modified in more current versions.

organizational segments, such as thesis statement (Burstein, Chodorow, & Leacock, 2003); and vocabulary content (Attali & Burstein, 2006).

Eleven of the individual features reflect essential characteristics in essay writing and are aligned with human scoring criteria. These features include related to: (1) proportion of errors in grammar, (2) proportion of word usage errors, (3) proportion of mechanical errors, (4) proportion of style comments, (5) number of required discourse elements, (6) average length of discourse elements, (7) score assigned to essays with similar vocabulary, (8) similarity of vocabulary to essays with score "6", (9) number word types divided by number of word tokens, (10) log frequency of least common words, (11) average length of words, and (12) total number of words (Attali & Burstein, 2006).[2] E-rater uses a sample of human-scored essay data for model building purposes. E-rater identifies features and feature weights are assigned using a multiple regression procedure. E-rater models can be built at the topic level, in which case a model is built for a specific essay prompt. However, more often, e-rater models are built at the grade-level. So, for instance, a model is built for 6th grade writers in *Criterion*. Writers can respond to topics selected by the teacher from the set of Criterion prompts, or the teacher can assign his or her own topic, and the 6th-grade model will be used to score these teacher-topic responses.

A comprehensive specification of the Intellimetric model is given in Elliot (2003). The model selects from 500 component features (i.e., proxes) (and clusters the selected elements into at least five consolidates sets. These sets include content, word variety, grammar, text complexity, and sentence variety. Other dimensions of writing may be used, but these five are common to *Intellimetric* models. *Intellimetric* uses word nets

[2] This feature is not used in the current version of e-rater.

based on Latent Semantic Dimension which is similar in nature to LSA (i.e., it determines how close the candidate response is, in terms of content, to a modeled set of vocabulary). Word variety refers to word complexity or word uniqueness. The grammar composite that evaluates things like subject-verb agreement, and text complexity is similar in nature to ascertaining the reading level of the text. The information gleaned is then used by a series of independent mathematical judges, or mathematical models, to "predict" the expert human scores and then optimized to produce a final predicted score. Typically, the judge that explains the largest proportion of rater variance will be employed in model development.

The technical details of the *Intelligent Essay Assessor* are highlighted in Landauer, Laham, & Foltz (2003). IEA is modeled using a two-pronged approach. The content of the essay is assessed by using a combination of external databases and LSA. For example, if the writing prompt had to do with the differentiation among Freud's concepts of superego, ego, and id, the reference database might include the electronic version of an introductory text in psychology. From that database, LSA would determine the likelihood of encountering certain words (e.g.., the term "conscience" as a synonym for "superego") given the constellation of vocabulary in the reference text. A candidate essay with more relevant vocabulary will be awarded a higher score. In setting up the models, *IEA* incorporates a validation procedure to check that LSA scores are aligned with those that might be given by human raters.

In contrast to *e-rater* and *Intellimetric*, the non-content features (e.g., mechanics, style, organization) of IEA are not fixed, but rather are constructed as a function of the domains assessed in the rating rubric. The weights for prox variables associated with these domains are predicted based on human ratings, and then are combined with the score calculated for content.

Reliability and Validity

Because AES models often formed by using more than two raters, studies that have evaluated inter-rater agreement have usually showed that the agreement coefficients between the computer and human raters is at least as high or higher than among human raters themselves (Elliot, 2003; Landauer et al., 2003; Page & Petersen, 1995). All AES engines have obtained exact agreements with humans as high as the mid-80’s and adjacent agreements in the mid-high 90’s--slightly higher than the agreement coefficients for trained human raters. Several validity studies have suggested that AES engines tap the same construct as that being evaluated by human raters. Page, Keith, & LaVoie (1995) examined the construct validity of AES, Keith (2003) summarized several discriminant and true score validity studies of the technology, and Attali & Burstein (2006) demonstrated the relationship between AES and instructional activities associated with writing.

AES is not without its detractors. Ericcson & Haswell (2006) performed a comprehensive critique of the technology from the perspective of those who teach post-secondary writing. Objections to the technology ranged from a concern about the ethics of using computers rather than humans to teach writing to the lack of synchronicity between how human graders approach the rating task and the process by which AES evaluates a writing sample to failed implementations of AES in university placement testing programs. And clearly there are certain types of stylized text writing that AES may never be able to evaluate. Nevertheless, AES is now used as a scoring process for high-stakes tests (e.g., GMAT) and is provided as a common instructional intervention for writing.

AES was a technology trigger that has spawned several related, and new innovative education technologies. In the next section we provide descriptions of emerging technology that, based on AES, has migrated to other measurement domains.

Transformations into New Applications

The success of AES and short-answer scoring (Leacock & Chodorow, 2003) has set the stage for a number of new capabilities developed for text-based analysis for enhanced feedback related to technical and organizational writing quality to help both native and non-native English speakers, and applications that incorporate text-analysis capabilities to provide reading comprehension support for English language learners (ELLs). Until now, the majority of AES and related capabilities have focused on text. However, speech-based capabilities are also making their way into commercial applications. In light of this, the second half of this section is focused a discussion of a speech-based, instructional capability currently used for scoring the speech of ELLs.

Criterion^{SM} Online Essay Evaluation Service

As mentioned above, automated essay scoring engines are typically combined with electronic portfolio systems to provide a full-spectrum set of services for those involved with writing instruction and assessment. In this section, a description of the *Criterion* online essay evaluation service is provided. The application is designed to help teachers in K-12 classrooms, and in community college, and university classrooms who typically have a large number of writing assignments to grade. This limits the number of writing assignments that teachers can offer to students. In an effort to offer additional writing practice to students, researchers have sought to develop applications not only for automated essay scoring, but that also offer more descriptive essay feedback similar to teacher feedback of student writing: indications of grammar, usage, and mechanics errors, stylistic, and organization and development issues. Pioneering work in automated feedback of this kind was initiated in the 1980's with the Writer's Workbench (MacDonald, Frase, S., & Keenan, 1982), and continues in applications, including the *Criterion*^{SM} *online essay evaluation service* (Burstein, Chodorow, & Leacock, 2004), the AOL$^{®}$ Writing Wizard, and Vantage Learning MY Access!$^{®}$.

The Criterion^{SM} *online essay evaluation service* combines e-rater automated essay scoring, advisories indicating anomalies, such as off-topicness (Higgins, Burstein, & Attali, 2006), and descriptive feedback. The descriptive feedback is comprised of a suite of programs that evaluate and, subsequently, flag essays for errors in grammar, usage, and mechanics; identify an essay's discourse structure; and, recognize undesirable stylistic features. As the population of English language learners (ELL) grows, researchers are working on enhancements to the grammatical error detection component to accommodate the kinds of mistakes more common in the ELL population. This kind of feedback includes determiner and preposition errors (Han, Chodorow, & Leacock, 2004), and collocation errors (e.g., "*strong computer*" instead of "*powerful computer*") (Futagi, Deane, Chodorow, & Tetreault (2008); Pantel & Lin, 2000; Shei & Pain, 2000)..

Criterion offers a pre-writing (planning) utility. This emphasis on planning was a logical outgrowth of the process-writing approach that *Criterion* embodies. Both earlier literature (Elbow, 1973) and more recent research have suggested that making plans can help students produce better quality writing, just as revising drafts can (Chai, 2004;

Goldstein & Carr, 1996) In light of this research, it is advisable to incorporate formal planning activities into writing instruction applications. The ability to collect student planning data through formal planning applications provides a new and authentic data source that can be used in writing research. Other computer-based instructional systems also offer a planning tool, including Inspiration Software®, Inc., which offers elaborate graphic organizers for writing and research projects, while online writing-instruction applications such as the AOL® Writing Wizard, CompassLearning® Odyssey Writer, and Vantage Learning MY Access!® provide on-screen planning tools to aid in the process of composition.

Generally speaking, researchers continue to develop capabilities for online writing instruction programs that are aligned with different populations of students and their respective needs with regard to their becoming more proficient writers.

Text Adaptor: Technology to Support English Language Learners

Authentic texts for the classroom that are grade-level appropriate and accessible to English language learners are often unavailable, especially in middle school and high school. As a result, the time-consuming practice of manual text adaptation has become a required task for both ESL and content-area teachers. Research suggests that certain kinds of text modifications, specifically vocabulary expansion and elaboration (i.e., providing synonyms and native language cognates) can facilitate students' comprehension of content in a text (Perez, 1981)(Carlo et al., 2004; Fitzgerald, 1995; Hancin-Bhatt & Nagy, 1994; Ihnot, 1997; James & Klein, 1994; Jimenez, Garcia, & Pearson, 1996; Nagy, Garcia, Durgunoglu, & Hancin-Bhatt, 1993; Oh, 2001; Yano, Long, & Ross, 1994).

Text Adaptor, a web-based tool, was designed as an authoring tool to support K-12 teachers in the text adaptation practice. While we continue to develop the tool, it currently incorporates several natural language processing (NLP) capabilities to support automated generation suggested text modifications for classroom texts. Tool suggestions are similar to the kinds of adaptations that teachers might implement for English language learners in their content-area classes.

Text Adaptor allows users to import a text or web page, and subsequently, to generate the following types of adaptations of the imported text: English and Spanish text summaries, vocabulary support, including synonym (Lin, 1998), antonym[3], and Spanish/English cognate identification. *Text Adaptor* also identifies complex phrasal and sentence structures, and academic vocabulary, fixed phrases (for example, phrasal verbs and collocations), and cultural references. Teachers can then modify the text accordingly, given the learning needs of the ELL students in their classrooms. NLP capabilities used to generate these adaptations include, automatic summarization (Marcu, 2000), machine translation[4], and synonym and antonym identification. The adaptation capabilities include strategies used by teachers to manually create adaptations, such as summaries and varied vocabulary support, as well as translating a text into another language. Teachers can use *Text Adaptor* to author any kind of classroom text, including reading texts, activities, and assessments.

As part of this research, a 2008 pilot study was conducted in two online teacher professional development (TPD) for ELL teachers in the United States: one at a large, private university on the west coast, and another at a large, private university on the east

[3] WordNet® lexical database.
[4] Language Weaver's English-to-Spanish machine translation system: http://www.languageweaver.com.

coast (Burstein, 2009; Shore, Burstein, and Sabatini, 2009). A central purpose of the pilot was to gauge if *Text Adaptor* increased teachers' linguistic awareness, resulting in improved text adaptations for ELLs. A pre-posttest design was implemented with approximately 70 teachers enrolled in the TPD courses. The pilot activities were integrated into the respective TPD courses. All participants completed online background surveys about their educational and teaching experiences, and post-surveys that asked the control group about their experience adapting texts in the pilot study, and asked the treatment group about their experiences using *Text Adaptor*. All participants completed manual (pre-) adaptations to gauge baseline adaptation knowledge and ability. Both completed a mid- and post-adaptation activity. *Control* teachers completed these activities manually, while *treatment* teachers were trained to use *Text Adaptor* and completed their adaptations using the tool. An important outcome indicated that *all* teachers who participated in this pilot gained knowledge about linguistic features as a result of the TPD training. Teachers were better able to articulate how to modify content to make it accessible to all students. In comparative analyses of the pre- and post-adaptations, we found that teachers who used *Text Adaptor* modified features in texts that were closely associated with *best practices* in modifying materials for non-native English speakers, and modified the language and content of the text more comprehensively than teachers who did not have access to *Text Adaptor*. Positive outcomes suggesting teachers' increased knowledge and linguistic awareness around text adaptations when they used the tool has inspired additional research toward developing additional tool features to support authoring of content-area texts for English language learners.

Automated scoring of spoken responses

Automated scoring of speech follows a paradigm similar to that of automated essay scoring. First, language related features are extracted, and in a second step a scoring model is used to compute a score based on a combination of these features. The main difference between text and speech is that the word identities are unknown in speech, and additional programming is needed for a speech recognizer to generate word hypotheses from the digitized candidate’s speech response to an item prompt. (Another difference from an assessment perspective is that speech testing is generally done for non-native speakers.)

Ordinate, a subsidiary of Harcourt, has been developing language tests since the 90’s where basic language abilities such as reading or repeating are tested (Bernstein, 1999). This is another way of avoiding the high error rate in open-ended speech recognition for spontaneous speech. They showed correlations around 0.80 between their tests and other widely used language tests such as ETS®’s TOEFL® (Bernstein, DeJong, Pisoni, & Townshend, 2000).

Cucchiarini et al. (Cucchiarini, Strik, & Boves, 1997a, 1997b) developed a speech recognition based automatic pronunciation scoring system for Dutch by using features such as log likelihood Hidden Markov Model scores, various duration scores, and information on pauses, word stress, syllable structure, and intonation. They also found good agreement (correlations above 0.70) between machine scores and human ratings of pronunciation.

Stanford Research Institute (SRI) International, similarly, has been developing an automatic pronunciation scoring system, EduSpeak™, which measures phone accuracy, speech rate, and duration distributions for non-native speakers who read English texts (Franco et al., 2000). Unlike in Ordinate's test, the texts being read need not be known to the system for prior training.

At Educational Testing Service (ETS®), research in automated speech scoring has been conducted since 2002. In 2006 a first speech scoring system, *SpeechRater*^{SM}, was successfully deployed to score the speaking section of *TOEFL® Practice Online* (TPO). This is an environment that helps students prepare for the Test of English as a Foreign Language (TOEFL®). Unlike for the aforementioned predecessors, the goal for developing *SpeechRater* is to provide scoring for assessments that cover a wide range of speaking proficiency (i.e., not only pronunciation) and to elicit spontaneous and natural speech from the test candidates as opposed to mere reading or repetition (Xi, Higgins, Zechner, & Williamson,2008; Xi, Zechner, & Bejar, 2006; Zechner & Bejar, 2006; Zechner, Higgins, Xi & Williamson (in press).

The tasks scored by *SpeechRater* are modeled on those used in the Speaking section of the TOEFL® iBT (internet-based test) These tasks ask the examinee to provide information or opinions on familiar topics based on their personal experience or background knowledge, as well as to respond to read or audio stimuli related to campus life and academic situations, such as lectures. The speaking time per item is about a minute. They are scored on a scale of 1-4, with a score of zero assigned to responses which do not address the task.

The design of *SpeechRater* is similar to that of e-rater, which underscores both the influence which the history of work in essay scoring has had on the development of speech scoring systems, and the fundamental similarities in the two domains. Both systems proceed by first extracting a vector of features to represent a response, and then using a machine learning system to predict the appropriate score based on those features.

In fact, there is a preliminary step in the case of *SpeechRater*: the response is first processed by a speech recognizer, the output of which provides a more pliable basis for the construction of scoring features than the raw speech stream. This speech recognizer is adapted to the speech of non-native English speakers from a wide variety of first-language backgrounds, but still manages a word accuracy rate of only around 50%. While this means that, on average, every other word is recognized incorrectly, this is the best that can currently be achieved by state-of-the art wide-coverage speech recognition systems on data from non-native speakers with multiple language backgrounds and proficiency levels, under variable recording conditions. (Since TPO is web-based, examinees may record their responses in their own homes, using their own microphones.)

This level of recognizer performance means that the features extracted by *SpeechRater* must not be highly dependent on recognition accuracy. By the same token, this means that *SpeechRater* cannot presently be expected to account for the full range of elements mentioned in the TOEFL rubric®. The rubric specifies three dimensions of attributes which contribute to the score of a response:

- Delivery (low-level technical features of speech production, such as pronunciation and fluency),
- Language Use (formal cues of linguistic competence, such as grammar and diction), and
- Topic Development (higher-level semantic, pragmatic, and organizational aspects of the response).

Delivery features can most easily be extracted from the state of the speech recognition system, while language use is more difficult to address, given the constraints of recognition accuracy. Of course, the appropriate development of the topic is even more challenging to assess without an accurate transcript of the response.

Most of the features actually used in *SpeechRater* address the delivery aspect of the TOEFL® speaking construct in one way or another. A subset of the features do, however, relate to the Language Use dimension of the construct as well. The statistical model *SpeechRater* uses to predict the score on the basis of these features is a multiple linear regression, although promising experiments have been performed using decision trees as well.

Currently, *SpeechRater* is in operational use to score the *TOEFL Practice Online* speaking section only. It provides the examinee with a predicted score for the entire section, comprised of six speaking items, and with a range within which their score is expected to fall with a certain probability. In the future, the application's use may be expanded to other testing programs, and work is being conducted to expand the construct coverage of the model, to bring it into closer alignment with the scoring of the operational TOEFL®.

Conclusion

As amazing as the invention of television was in the late 1940's, it was clear that the kinescopic pictures and mono-channel sound that reflected the technology of the times was an inadequate substitute for re-creating the "real thing". However, over time improvements were made to the broadcasting enterprise—tape for the kinescope, color, multi-channel sound, high definition clarity. In addition, new uses were made for television beyond entertainment (e.g., instruction, security). Do these developments make the experience any more authentic? Maybe. How often have you heard someone say, "I'd rather watch it on TV."?

In a similar vein, automated essay scoring might still be characterized as an emerging technology. Though it has been demonstrated to replicate human judgements in the grading of essays, over time it will be enhanced to do so with even more proficiency and accuracy. Moreover, it has branched out to perform other roles (instruction) and is now used as a conceptual platform for other applications (language proficiency ratings). Finally it has engendered a discussion about what constitutes good writing and how is it best achieved.

References

Ajay, H. B., Tillett, P. I., & Page, E. B. (1973). *Analysis of essays by computer (AEC-II)* (No. 8-0102). Washington, DC: U.S. Department of Health, Education, and Welfare, Office of Education, National Center for Educational Research and Development.

Attali, Y., & Burstein, J. (2006). Automated Essay Scoring With e-rater V.2. *Journal of Technology, Learning, and Assessment, 4*(3), Available from http://www.jtla.org.

Bernstein, J. (1999). *PhonePass Testing: Structure and Construct*. Menlo Park, CA: Ordinate Corporation.

Bernstein, J., DeJong, J., Pisoni, D., & Townshend, B. (2000). *Two experiments in automatic scoring of spoken language proficiency*. Paper presented at the InSTIL-2000, Dundee, Scotland.

Burstein, J. (2003). The E-rater scoring engine: Automated essay scoring with natural language processing. In M. D. Shermis & J. Burstein (Eds.), *Automated essay scoring: A cross-disciplinary perspective* (pp. 113-122). Mahwah, NJ: Lawrence Erlbaum Associates, Inc.

Burstein, J., Chodorow, M., & Leacock, C. (2003). *Criterion: Online essay evaluation: An application for automated evaluation of test-taker essays.* Paper presented at the Fifteenth Annual Conference on Innovative Applications of Artificial Intelligence, Acapulco, Mexico.

Burstein, J., Chodorow, M., & Leacock, C. (2004). Automated essay evaluation: The Criterion Online writing service. *AI Magazine, 25*(3), 27-36.

Burstein, J. (2009). Opportunities for natural language processing in education. In A. Gebulkh (Ed.), *Springer lecture notes in computer science* (Vol. 5449, pp. 6-27). Springer: New York, NY.

Carlo, M. S., August, D., McLaughlin, B., Snow, C. E., Dressler, C., Lippman, D., et al. (2004). Closing the gap: Addressing the vocabulary needs of English language learners in bilingual and mainstream classrooms. *Reading Research Quarterly, 39*(2), 188-215.

Chai, C. O. L. (2004). *Development and validation of procedures to assess writing plans.* Unpublished dissertation, University of British Columbia, Vancouver.

Cucchiarini, C., Strik, H., & Boves, L. (1997a). *Automatic evaluation of Dutch pronunciation by using speech recognition technology*. Paper presented at the IEEE Automatic Speech Recognition and Understanding Workshop, Santa Barbara, CA.

Cucchiarini, C., Strik, H., & Boves, L. (1997b, September). *Using speech recognition technology to assess foreign speakers' pronunciation of Dutch.* Paper presented at the Third International Symposium on the Acquisition of Second Language Speech: NEW SOUNDS 97, Klagenfurt, Austria.

Futagi, Y., Deane, P., Chodorow, M., and Tetreault, J (2008). A computational approach to detecting collocation errors in the writing of non-native speakers of English. *Computer Assisted Language Learning, 21*, 353-367.

Elbow, P. (1973). *Writing without teachers*. New York, NY: Oxford University Press.

Elliot, S. (2003). Intellimetric: From here to validity. In M. D. Shermis & J. Burstein (Eds.), *Automated essay scoring: A cross-disciplinary perspective* (pp. 71-86). Mahwah, NJ: Lawrence Erlbaum Associates, Inc.

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

Fitzgerald, J. (1995). English-as-a-second-language learners' cognitive reading processes: A review of research in the United States. *Review of Educational Research*, *65*(2), 145-190.

Franco, H., Abrash, V., Precoda, K., Bratt, H., Rao, R., & Butzberger, J. (2000). *The SRI EduSpeak system: Recognition and pronunciation scoring for language learning.* Paper presented at the InSTiLL-2000, Dundee, Scotland.

Goldstein, A. A., & Carr, P. G. (1996). *Can students benefit from process writing?* Washington, D.C.: National Center for Educational Statistics (ERIC Document Reproduction Service Number ED 395 320).

Han, N.-R., Chodorow, M., & Leacock, C. (2004). *Detecting errors in English article usage with a maximum entropy classifier trained on a large, diverse corpus.* Paper presented at the 4th International Conference on Language Resources and Evaluation, Lisbon, Portugal.

Hancin-Bhatt, B., & Nagy, W. E. (1994). Lexical transfer and second language morphological development. *Applied Psycholinguistics, 15*(3), 289-310.

Higgins, D., Burstein, J., & Attali, Y. (2006). Identifying off-topic student essays without topic-specific training data. *Natural Language Engineering, 12*(2), 145-159.

Ihnot, C. (1997). *Read naturally*. St. Paul, MN: Read Naturally.

James, C., & Klein, K. (1994). Foreign language learners’ spelling and proof reading strategies. *Papers and Studies in Contrastive Linguistics*, *19*(31-46).

Jimenez, R. T., Garcia, G. E., & Pearson, D. P. (1996). The reading strategies of bilingual Latina/o who are successful English readers: Opportunities and obstacles. *Reading Research Quarterly, 31*(1), 90-112.

Kawate-Mierzejewska, M. (2003, March). *E-rater software.* Paper presented at the Japanese Association for Language Teaching, Tokyo, Japan.

Keith, T. Z. (2003). Validity and automated essay scoring systems. In M. D. Shermis & J. Burstein (Eds.), *Automated essay scoring: A cross-disciplinary perspective* (pp. 147-168). Mahwah, NJ: Lawrence Erlbaum Associates, Inc.

Landauer, T. K., Foltz, P. W., & Laham, D. (1998). Introduction to latent semantic analysis. *Discourse Processes*, *25*, 259-284.

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 scoring: A cross-disciplinary perspective* (pp. 87-112). Mahwah, NJ: Lawrence Erlbaum Associates, Inc.

Leacock, C., & Chodorow, M. (2003). C-rater: Scoring of short-answer questions. *Computers and the Humanities, 37*(4), 389-405.

Lin, D. (1998). *Automatic retrieval and clustering of similar words.* Paper presented at the 35th Annual Meeting of the Association for Computational Linguistics, Montreal, Canada.

MacDonald, N. H., Frase, L. T., S., G. P., & Keenan, S. A. (1982). The Writer's Workbench: Computer aids for text analysis. *IEEE Transactions on Communications*, *30*(1), 105-110.

Marcu, D. (2000). *The theory and practice of discourse parsing and summarization*. Cambridge, MA: The MIT Press.

Nagy, W. E., Garcia, G. E., Durgunoglu, A. Y., & Hancin-Bhatt, B. (1993). Spanish-English bilingual students' use of cognates in English reading. *Journal of Reading Behavior*, *25*(3), 241-259.

Oh, S. Y. (2001). Two types of input modification and EFL reading comprehension: Simplification versus elaboration. *TESOL Quarterly*, *35*(1), 69-96.

Page, E. B. (1966). The imminence of grading essays by computer. *Phi Delta Kappan, 48*, 238-243.

Page, E. B., Keith, T., & Lavoie, M. J. (1995, August). *Construct validity in the computer grading of essays.* Paper presented at the annual meeting of the American Psychological Association, New York, NY.

Page, E. B., & Petersen, N. S. (1995). The computer moves into essay grading: Updating the ancient test. *Phi Delta Kappan, 76*(7), 561-565.

Pantel, P., & Lin, D. (2000). *Word-for-word glossing with contextually similar words.* Paper presented at the 1st Annual Meeting of the North American Chapter of Association for Computational Linguistics (NAACL2000), Seattle, WA.

Perez, E. (1981). Oral language competence improves reading skills of Mexican-American third graders. *Reading Teacher, 35*(1), 24-27.

Shei, C., & Pain, H. (2000). An ESL writer's collocational aid. *Computer Assisted Language Learning, 13*(2), 167-182.

Shermis, M. D., & Burstein, J. (2003). Introduction. In M. D. Shermis & J. Burstein (Eds.), *Automated essay scoring: A cross-disciplinary perspective* (pp. xiii-xvi). Mahwah, NJ: Lawrence Erlbaum Associates.

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

Shore, J., Burstein, J. and Sabatini, J. (2009, April). *Web-based technology that supports ELL reading instruction.* Paper presented at the Interactive Symposium for the Special Interest Group, Computers and Internet for Educational Applications, held in conjunction with the American Educational Research Association Annual Meeting, San Diego. Available at: http://ciaesig.ning.com/

Vantage Learning. (2001). *A Preliminary study of the efficacy of IntelliMetric™ for use in scoring Hebrew assessments*. Newtown, PA: Vantage Learning.

Vantage Learning. (2002). *A study of Intellimetric^{TM} for responses in scoring Bahasa Malay*. Newtown, PA: Vantage Learning.

Xi, X., Higgins, D., Zechner, K., & Williamson, D.M. (2008). Automated Scoring of Spontaneous Speech Using SpeechRater^{SM} v1.0. Educational Testing Service, Research Report RR-08-62, November.

Xi, X., Zechner, K., & Bejar, I. (2006, April). *Extracting meaningful speech features to support diagnostic feedback: An ECD approach to automated scoring.* Paper presented at the National Council on Measurement in Education, San Francisco, CA.

Yano, Y., Long, M., & Ross, S. (1994). The effects of simplified and elaborated texts on foreign language reading comprehension. *44*, 189-219.

Zechner, K., & Bejar, I. (2006). *Towards automatic scoring of non-native spontaneous speech.* Proceedings of the Human Language Technology Conference of the NAACL, New York, NY.

Zechner, K., Higgins, D., Xi, X. & Williamson, D.M. (in press). Automatic scoring of non-native spontaneous speech in tests of spoken English. Speech Communication.