J·T·L·A

The Journal of Technology, Learning, and Assessment

Volume 4, Number 3 · February 2006

Automated Essay Scoring With e-rater® V.2

Yigal Attali & Jill Burstein

www.jtla.org

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

J·T·L·A

Volume 4, Number 3

Automated Essay Scoring With e-rater® V.2

Yigal Attali & Jill Burstein

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

Copy Editor: Kevon R. Tucker-Seeley
Design: Thomas Hoffmann
Layout: Aimee Levy

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

Preferred citation:

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

Abstract:

E-rater® has been used by the Educational Testing Service for automated essay scoring since 1999. This paper describes a new version of e-rater (V.2) that is different from other automated essay scoring systems in several important respects. The main innovations of e-rater V.2 are a small, intuitive, and meaningful set of features used for scoring; a single scoring model and standards can be used across all prompts of an assessment; modeling procedures that are transparent and flexible, and can be based entirely on expert judgment. The paper describes this new system and presents evidence on the validity and reliability of its scores.

J·T·L·A

Automated Essay Scoring With e-rater® V.2

Yigal Attali & Jill Burstein
Educational Testing Service

Introduction

Constructed-response assessments have several advantages over traditional multiple-choice assessments (see, for example, Bennett & Ward, 1993), but the greatest obstacle for their adoption in large-scale assessment is the large cost and effort required for scoring. Developing systems that can automatically score constructed responses can help reduce these costs in a significant way and may also facilitate extended feedback for the students.

Automated scoring capabilities are especially important in the realm of essay writing. Essay tests are a classic example of a constructed-response task where students are given a particular topic (also called a prompt) to write about[1]. The essays are generally evaluated for their writing quality. This task is very popular both in classroom instruction and in standardized tests—recently the SAT® introduced a 25-minute essay-writing task to the test. However, evaluating student essays is also a difficult and time-consuming task.

Surprisingly for many, automated essay scoring (AES) has been a real and viable alternative and complement to human scoring for many years. As early as 1966, Page showed that an automated "rater" is indistinguishable from human raters (Page, 1966). In the 1990's more systems were developed; the most prominent systems are the Intelligent Essay Assessor (Landauer, Foltz, & Laham, 1998), Intellimetric (Elliot, 2001), a new version of the Project Essay Grade (PEG, Page, 1994), and e-rater (Burstein et al., 1998).

AES systems do not actually read and understand essays as humans do. Whereas human raters may directly evaluate various intrinsic variables of interest, such as diction, fluency, and grammar, in order to produce an essay score, AES systems use approximations or possible correlates of these intrinsic variables. Page and Petersen (1995) expressed this distinction with the terms *trin* (for intrinsic variables) and *prox* (for an approximation).

With all of the AES systems mentioned above, scoring models are developed by analyzing a set of typically a few hundred essays written on a specific prompt and pre-scored by as many human raters as possible. In this analysis, the most useful *proxes* for predicting the human scores, out of those that are available to the system, are identified. Finally, a statistical modeling procedure is used to combine these proxes and come up with a final machine-generated score of the essay.

Skepticism and criticisms have accompanied AES over the years, often related to the fact that the machine does not understand the written text. Page and Petersen (1995) list three such general objections to AES, the humanistic, defensive, and construct objections. The humanistic objection is that computer judgments should be rejected out of hand since they will never understand or appreciate an essay in the same way as a human. This objection might be the most difficult to reconcile and reflects the general debate about the merits and criteria for evaluation of artificial intelligence. In practice, improvements in the systems, empirical research, and better evaluations may contribute to increasing use of AES systems. In the meantime, for some implementations of high-stakes AES this objection (as well as others) is managed by including a human rating in the assessment of all essays. For example, e-rater is used as a second rater combined with a human rater in the essay writing section of the Graduate Management Admission Test (GMAT).

The defensive objection states that the computer may be successful only in an environment that includes only "good faith" essays. Playful or hostile students will be able to produce "bad faith" essays that the computer will not be able to detect. Only one published study addressed the defensive argument directly. In Powers, et al. (2001) students and teachers were asked to write "bad faith" essays deliberately in order to fool the e-rater system into assigning a higher (and lower) score than they deserved. Surely more such studies are needed in order to delineate more clearly the limitations of AES systems.

The last objection Page and Petersen (1995) list is the construct objection, arguing that the *proxes* measured by the computer are not what is really important in an essay. In this respect, an improved ability to provide specific diagnostic feedback on essays in addition to an overall score should serve as evidence for the decreasing gap between *proxes* and *trins*. It is important to distinguish, however, between canned feedback and specific suggestions for improvement of the essay. Although the former may be useful to students, the latter is preferred.

Partly in response to critiques of AES, there is a growing body of literature on the attempts to validate the meaning and uses of AES. Yang et al. (2002) classify these studies into three approaches. The first approach,

which is the most common in the literature, focuses on the relationship between automated and human scores of the same prompt. These studies compare the machine-human agreement to the human-human agreement and typically find that the agreements are very similar (Burstein et al., 1998; Elliot, 2001; Landauer et al., 2001). Some studies also attempt to estimate a true score, conceptualized as the expected score assigned by many human raters (Page, 1966).

We believe, however, that this type of validation is not sufficient for AES. In the case of AES, the significance of comparable single-essay agreement rates should be evaluated against the common finding that the simplest form of automated scoring which considers only essay length could yield agreement rates that are almost as good as human rates. Clearly, such a system is not valid. On the other hand, a "race" for improving single-essay agreement rates beyond human rates is not theoretically possible, because the relation between human ratings and any other measure is bound by the human inter-rater reliability.

Consequently, Yan et al.'s (2001) second approach to validation, examining the relationship between test scores and other measures of the same or similar construct, should be preferred over the first approach. Nonetheless, only two studies examined the relationship between essay scores and other writing measures. Elliot (2001) compared Intellimetric scores with multiple-choice writing test scores and teacher judgments of student writing skills and found that the automated scores correlated about as well with these external measures as human essay scores. Powers et al. (2002) also examined the relation between non-test writing indicators and between human and automated (e-rater) scores. They found somewhat weaker correlations for automated scores than for human scores.

Surprisingly, an "external" measure that was not studied in previous research on AES is the relationship of essay scores with essay scores from other prompts. Relationships across different prompts make more sense from an assessment perspective (where different forms use different prompts), they allow the estimation of a more general type of reliability (alternate-form reliability), and can serve as the basis for an estimation of the shared variance and true-score correlation between human and machine scores.

This kind of validation also changes the usual perspective on human and machine scores. In the first validation approach, the human scores are seen as the ultimate criterion against which the machine scores are evaluated. However, Bennett and Bejar (1998) urge against the reliance on human ratings as the criterion for judging the success of automated scoring. As they point out, human raters are highly fallible, and this is

especially true for human essay scoring. Using human ratings to evaluate and refine automated scoring could produce a suboptimal result.

This brings us to the third validation approach of Yan et al. (2001). In addition to validation efforts based on demonstrating statistical relationship between scores, the importance of understanding the scoring processes that AES uses should be stressed. However, these kinds of studies are not common. As an example, consider the question of the relative importance of different dimensions of writing, as measured by the machine, to the automated essay scores. This is a fundamental question for establishing the meaning of automated scores. However, few answers to it exist in the literature. Burstein et al. (1998) report on the most commonly used features in scoring models, but apart from four features, the features actually used varied greatly across scoring models. Landauer et al. (2001) suggest that the most important component in scoring models is content.

The lack of studies of this kind may be attributed to the data-driven approach of AES. Both the identification of *proxes* (or features) and their aggregation into scores rely on statistical methods whose aim is to best predict a particular set of human scores. Consequently, both what is measured and how it is measured may change frequently in different contexts and for different prompts. This approach makes it more difficult to discuss the meaningfulness of scores and scoring procedures.

There is, however, a larger potential for AES. The same elastic quality of AES that produces relatively unclear scores can be used to control and clear up the scoring process and its products. This paper describes a new approach in AES as it is applied in e-rater V.2. This new system differs from the previous version of e-rater and from other systems in several important ways that contribute to its validity. The feature set used for scoring is small and the features are intimately related to meaningful dimensions of writing. Consequently, the same features are used for different scoring models. In addition, the procedures for combining the features into an essay score are simple and can be based on expert judgment. Finally, scoring procedures can be successfully applied on data from several essay prompts of the same assessment. This means that a single scoring model is developed for a writing assessment, consistent with the human rubric that is usually the same for all assessment prompts in the same mode of writing. In e-rater V.2 the whole notion of training and data-driven modeling is considerably weakened.

These characteristics of the new system strengthen the standardization and communicability of scores, contribute to their validity, and may contribute to greater acceptability of AES. As Bennett and Bejar (1998) note, automated scoring makes it possible to control what Embretson (1983) calls the *construct representation* (the meaning of scores based on internal

evidence) and *nomothetic span* (the meaning of scores based on relationships with external variables). With e-rater V.2, it is possible to control the construction of scores both based on the meaning of the different dimensions of scoring and on external evidence about the performance of the different dimensions and e-rater as a whole.

The paper will start with a description of the features in the new system and its scoring procedures. Then performance results for the new system are presented from a variety of assessment programs. In addition to reporting the usual agreement statistics between human and machine scores, this paper adds analyses based on alternate-form results. This allows a better comparison of human and machine reliabilities and makes it possible to estimate the human-machine true-score correlation. Results show that e-rater scores are significantly more reliable than human scores and that the true-score correlation between human and e-rater scores is close to perfect. The paper concludes with an introduction to two new developments that are made possible with e-rater V.2.

The Feature Set

AES was always based on a large number of features that were not individually described or linked to intuitive dimensions of writing quality. Intellimetric (Elliot, 2001) is based on hundreds of undisclosed features. The Intelligent Essay Assessor (Landauer, Laham, and Foltz, 2003) is based on a statistical technique for summarizing the relations between words in documents, so in a sense it uses every word that appears in the essay as a mini-feature. The first version of e-rater (Burstein et al., 1998) used more than 60 features in the scoring process. PEG (Page, 1994) also uses dozens of mostly undisclosed features. One of the most important characteristics of e-rater V.2 is that it uses a small set of meaningful and intuitive features. This distinguishing quality of e-rater allows further enhancements that together contribute to a more valid system.

The feature set used with e-rater V.2 include measures of grammar, usage, mechanics, style, organization, development, lexical complexity, and prompt-specific vocabulary usage. This feature set is based in part on the NLP foundation that provides the instructional feedback to students who are writing essays in *Criterion*^{SM}, ETS's writing instruction application. Therefore, a short description of *Criterion* and its feedback systems will be given before the detailed description of the feature set.

Criterion is a web-based service that evaluates a student's writing skill and provides instantaneous score reporting and diagnostic feedback. The e-rater engine provides score reporting. The diagnostic feedback is based on a suite of programs (writing analysis tools) that identify the essay's

discourse structure, recognize undesirable stylistic features, and evaluate and provide feedback on errors in grammar, usage, and mechanics.

The writing analysis tools identify five main types of grammar, usage, and mechanics errors – agreement errors, verb formation errors, wrong word use, missing punctuation, and typographical errors. The approach to detecting violations of general English grammar is corpus based and statistical, and can be explained as follows. The system is trained on a large corpus of edited text, from which it extracts and counts sequences of adjacent word and part-of-speech pairs called *bigrams*. The system then searches student essays for bigrams that occur much less often than would be expected based on the corpus frequencies (Chodorow & Leacock, 2000).

The writing analysis tools also highlight aspects of style that the writer may wish to revise, such as the use of passive sentences, as well as very long or very short sentences within the essay. Another feature of undesirable style that the system detects is the presence of overly repetitious words, a property of the essay that might affect its rating of overall quality (Burstein & Wolska, 2003).

Finally, the writing analysis tools provide feedback about discourse elements present or absent in the essay (Burstein, Marcu, and Knight, 2003). The discourse analysis approach is based on a linear representation of the text. It assumes the essay can be segmented into sequences of discourse elements, which include introductory material (to provide the context or set the stage), a thesis statement (to state the writer's position in relation to the prompt), main ideas (to assert the author's main message), supporting ideas (to provide evidence and support the claims in the main ideas, thesis, or conclusion), and a conclusion (to summarize the essay's entire argument). In order to identify the various discourse elements, the system was trained on a large corpus of human annotated essays (Burstein, Marcu, and Knight, 2003). Figure 1 (next page) presents an example of an annotated essay.

Figure 1: A Student Essay With Annotated Discourse Elements

<Introductory Material>"You can't always do what you want to do!," my mother said. She scolded me for doing what I thought was best for me. It is very difficult to do something that I do not want to do.**</Introductory Material> <Thesis>**But now that I am mature enough to take responsibility for my actions, I understand that many times in our lives we have to do what we should do. However, making important decisions, like determining your goal for the future, should be something that you want to do and enjoy doing.**</Thesis>**

<Introductory Material>I've seen many successful people who are doctors, artists, teachers, designers, etc.**</Introductory Material> <Main Point>**In my opinion they were considered successful people because they were able to find what they enjoy doing and worked hard for it.**</Main Point> <Irrelevant>**It is easy to determine that he/she is successful, not because it's what others think, but because he/she have succeed in what he/she wanted to do.**</Irrelevant>**

<Introductory Material>In Korea, where I grew up, many parents seem to push their children into being doctors, lawyers, engineer etc.**</Introductory Material> <Main Point>**Parents believe that their kids should become what they believe is right for them, but most kids have their own choice and often doesn't choose the same career as their parent's.**</Main Point> <Support>**I've seen a doctor who wasn't happy at all with her job because she thought that becoming doctor is what she should do. That person later had to switch her job to what she really wanted to do since she was a little girl, which was teaching.**</Support>**

<Conclusion>Parents might know what's best for their own children in daily base, but deciding a long term goal for them should be one's own decision of what he/she likes to do and want to do **</Conclusion>**

In addition to the information extracted from the writing analysis tools, e-rater V.2 features are also based on measures of lexical complexity and of prompt-specific vocabulary usage. Great care has been taken to calculate measures that are relatively independent of essay length and that are each related to human holistic evaluations of essays. Below is a description, by category, of the features included in the new feature set.

Grammar, Usage, Mechanics, and Style Measures (4 features)

Feedback and comments about grammar, usage, mechanics, and style are outputted from *Criterion*. Counts of the errors in the four categories form the basis for four features in e-rater V.2. Since raw counts of errors are highly related to essay length, the rates of errors are then calculated by dividing the counts in each category by the total number of words in the essay.

The distribution of these rates is highly skewed (with few essays that have high rates of errors). In particular, because there are essays that do not have any errors in a certain category, simple statistical transformations of the rates will not yield less skewed distributions that are desirable from a measurement perspective. To solve this problem, we add "1" to the total error count in each category (so that all counts are greater then 0) before the counts are divided by essay length. In addition, a log transformation is then applied to the resulting modified rates. Note that the modified rates for essays that originally did not have any errors will be different depending on the length of the essay. A short essay will have a higher modified rate than the modified rate for a long essay (when both initially had no errors). The distribution of log-modified rates is approximately normal. These four measures are referred to, henceforth, as *grammar, usage, mechanics,* and *style*.

Organization and Development (2 features)

There are many possible ways to use the discourse elements identified by the writing analysis tools, depending upon the type of prompt and the discourse strategy that is sought by the teacher or assessment. Prompts in standardized tests and in classroom instruction often elicit persuasive or informative essays. Both genres usually follow a discourse strategy that requires at least a thesis statement, several main and supporting ideas, and a conclusion.

The overall organization score (referred to in what follows as *organization*) was designed for these genres of writing. It assumes a writing strategy that includes an introductory paragraph, at least a three-paragraph body with each paragraph in the body consisting of a pair of main point and supporting idea elements, and a concluding paragraph. The organization score measures the difference between this minimum five-paragraph essay and the actual discourse elements found in the essay. Missing elements could include supporting ideas for up to the three expected main points or a missing introduction, conclusion, or main point. On the other hand, identification of main points beyond the minimum three would not contribute to the score. This score is only one possible use of the identified discourse elements, but was adopted for this study.

The second feature derived from *Criterion*'s organization and development module measures the amount of development in the discourse elements of the essay and is based on their average length (referred to as *development*).

Lexical Complexity (2 features)

Two features in e-rater V.2 are related specifically to word-based characteristics. The first is a measure of vocabulary level (referred to as *vocabulary*) based on Breland, Jones, and Jenkins' (1994) Standardized Frequency Index across the words of the essay. The second feature is based on the average word length in characters across the words in the essay (referred to as *word length*).

Prompt-Specific Vocabulary Usage (2 features)

E-rater evaluates the lexical content of an essay by comparing the words it contains to the words found in a sample of essays from each score category (usually six categories). It is expected that good essays will resemble each other in their word choice, as will poor essays. To do this, content vector analysis (Salton, Wong, & Yang, 1975) is used, where the vocabulary of each score category is converted to a vector whose elements are based on the frequency of each word in the sample of essays.

Content vector analysis is applied in the following manner in e-rater: first, each essay, in addition to a set of training essays from each score point, is converted to vectors. These vectors consist of elements that are weights for each word in the individual essay or in the set of training essays for each score point (some function words are removed prior to vector construction.). For each of the score categories, the weight for word *i* in score category *s*:

$$W_{is} = (F_{is} / MaxF_s) * \log(N / N_i)$$

Where F_{is} is the frequency of word *i* in score category *s*, $MaxF_s$ is the maximum frequency of any word at score point *s*, N is the total number of essays in the training set, and N_i is the total number of essays having word *i* in all score points in the training set.

For an individual essay, the weight for word *i* in the essay is:

$$W_i = (F_i / MaxF) * \log(N / N_i)$$

Where F_i is the frequency of word *i* in the essay and MaxF is the maximum frequency of any word in the essay.

Finally, for each essay, six cosine correlations are computed between the vector of word weights for that essay and the word weight vectors for

each score point. These six cosine values indicate the degree of similarity between the words used in an essay and the words used in essays from each score point.

In e-rater V.2, two content analysis features are computed from these six cosine correlations. The first is the *score point value* (1-6) for which the maximum cosine correlation over the six score point correlations was obtained (referred to as *max. cos.*). This feature indicates the score point level to which the essay text is most similar with regard to vocabulary usage. The second is the *cosine correlation value* between the essay vocabulary and the sample essays at the highest score point, which in many cases was 6 (referred to as *cos. w/6*). This feature indicates how similar the essay vocabulary is to the vocabulary of the best essays. Together these two features provide a measure of the level of prompt-specific vocabulary used in the essay.

Additional Information

In addition to essay scoring that is based on the features described above, e-rater also includes systems that are designed to identify anomalous and bad-faith essays. Such essays are flagged and not scored by e-rater. These systems are not discussed in this paper.

E-rater Model Building and Scoring

Scoring in e-rater V.2 is a straightforward process. E-rater scores are calculated as a weighted average of the standardized feature values, followed by applying a linear transformation to achieve a desired scale. A scoring model thus requires the identification of the necessary elements for this scoring process. There are three such elements: identifying the standardized feature weights (or relative weights; in e-rater they are commonly expressed as percentages of total standardized weight), identifying the means and standard deviations to be used in standardizing each feature values, and identifying appropriate scaling parameters. The following sections present the approach of e-rater V.2 for identification of these elements that contribute to user control over the modeling process and to standardization of scores.

Control and Judgment in Modeling

Typically, AES is based entirely on automated statistical methods to produce essay scores that resemble as much as possible human scores. Intellimetric (Elliot, 2001) is using a blend of several expert systems based on different statistical foundations that use hundreds of features. The Intelligent Essay Assessor (Landauer, Laham,

and Foltz, 2003) is based on Latent Semantic Analysis, a statistical technique for summarizing the relations between words in documents. The first version of e-rater (Burstein et al., 1998) used a stepwise regression technique to select the best features that are most predictive for a given set of data. PEG (Page, 1994) is also based on regression analysis.

The advantage of these automated statistical approaches to scoring is that they are designed to find optimal solutions with respect to some measure of agreement between human and machine scores. There are, however, several disadvantages to such automated approaches to scoring. One problem is that such statistical methods will produce different solutions in different applications, both in terms of the kind of information included in the solution and in the way it is used. A writing feature might be an important determinant of the score in one solution and absent from another. Moreover, features might contribute positively to the score in one solution and negatively in another. These possibilities are common in practice because AES systems are based on a large number of features that are relatively highly correlated among themselves and have relatively low correlations with the criterion (the human scores). These variations between solutions are a real threat to the validity of AES.

Another disadvantage of statistically-based scoring models is that such models may be difficult to describe and explain to users of the system. Difficulty in communicating the inner structure of the scoring model is a threat to the face validity of AES.

Finally, the use of statistical optimization techniques in an automated way might produce other undesirable statistical effects. For example, many statistical prediction techniques including regression produce scores that have less variability than the scores they are supposed to predict (in this case the human scores). This effect may be unacceptable for an assessment that considers using AES.

The above disadvantages of statistical modeling illustrate the importance of having judgmental control over modeling for AES. In e-rater V.2 an effort is made to allow such control over all elements of modeling. This is made possible by having a small and meaningful feature set that is accessible to prospective users and by using a simple statistical method for combining these features (weighted average of standardized feature scores).

Judgmental control is enhanced further by making it possible to determine relative weights judgmentally, either by content experts or by setting weights based on other similar assessments (assessments that have similar prompts, rubrics, or are designed for students with similar writing abilities). This allows control over the importance of the different dimensions of e-rater scoring when theoretical or other considerations are

present. Although weights can be determined from a multiple regression analysis, we repeatedly find (and will show below) that "non-optimal" weights are not less efficient than the optimal weights found through statistical analysis. Note that it is possible to combine statistical and judgment-based weights and that it is possible to control weights partially by setting limits on statistical weights. This is particularly useful in order to avoid opposite-sign weights (e.g., a negative weight when a positive weight is expected) or a very large relative weight that might also lower the face validity of scores.

Another aspect of modeling that is controlled in e-rater is scaling parameters. Since typically AES is expected to emulate human scoring standards, an appropriate scaling of machine scores should result in the same mean and standard-deviation as those of the human scores. In e-rater this is done by simply setting the mean and standard deviation of e-rater scores to be the same as the mean and standard deviation of a single human rater on the training set. Although this solution seems obvious it is in fact different than what will be achieved through the use of regression analyses, because regression analyses produce scores that are less variable than the predicted score. It is also important to note that scaling machine scores to have a smaller variation (as with the use of regression analysis) will generally result in *higher* agreement results. However, we feel that when machine scores are to be used as a second rater, in addition to a human rater, it is important for machine scores to have the same variation as the human rater.

Although the typical application of AES is for emulating human scoring standards there may be other applications that call for other scaling approaches. For example, when e-rater is used as the only rater it might be more appropriate to scale the scores to some arbitrary predefined parameters (e.g., to have the mean score of an equating group to be set to the middle of the score range).

Finally, e-rater achieves more control over modeling through enhanced standardization that will be discussed in the next section.

Standardization of Modeling

AES models have always been prompt-specific (Burstein, 2003; Elliot, 2003; Landauer, Laham, and Foltz, 2003). That is, models are built specifically for each topic, using data from essays written to each of the particular topics and scored by human raters. This process requires significant data collection and human reader scoring—both time-consuming and costly efforts. However, it also weakens the validity of AES because it results in different models with different scoring standards for each prompt that belongs to the assessment or program. There is generally only one set

of human scoring rubrics per assessment, reflecting the desire to have a single scoring standard for different prompts.

A significant advance of e-rater V.2 is the recognition that scoring should and could be based on program-level models. A program is defined here as the collection of all prompts that are supposed to be interchangeable in the assessment and scored using the same rubric and standards. The primary reason that program-level models might work as well as prompt-specific models is that the aspects of writing performance measured by e-rater V.2 are topic-*independent*. For example, if a certain organization score in a particular prompt is interpreted as evidence of good writing ability, then this interpretation should not vary across other prompts of the same program. The same is true with the other e-rater scores. Consistent with this, we have found that it is possible to build program-level (or generic) models without a significant decrease in performance. In other words, idiosyncratic characteristics of individual prompts are not large enough to make prompt-specific modeling perform better than generic modeling.

Analyses of the Performance of e-rater V.2

In this section we will present an evaluation of the performance of e-rater V.2 on a large and varied dataset of essays. This section will start with descriptive statistics of the human scores of the essays; then descriptive statistics and evidence on the relation between individual features and human scores will be presented. Next, the performance of e-rater scoring will be compared to human scoring in terms of inter-rater agreement; and finally a subset of the data will be used to evaluate and compare the alternate-form reliability of e-rater and human scoring and to estimate the true-score correlation between e-rater and human scoring.

The analyses that will be presented in this paper are based on essays from various user programs. We analyze sixth through twelfth grade essays submitted to *Criterion* by students, GMAT essays written in response to issue and argument prompts, and TOEFL® (Test of English as Foreign Language) human-scored essay data. All essays were scored by two trained human readers according to grade-specific or program rubrics. All human scoring rubrics are on a 6-point scale from 1 to 6.

Descriptive Statistics

Table 1 presents descriptive statistics of these essays. Data is presented for the first of two human scores available for each essay (H1). Overall there were 64 different prompts and more than 25,000 essays in the data.

Table 1: Descriptive Statistics on Essays and Single Human Score (H1)

Program	Prompts	Mean # of essays per prompt	Mean H1	STD H1
Criterion 6th Grade	5	203	2.93	1.24
Criterion 7th Grade	4	212	3.22	1.29
Criterion 8th Grade	5	218	3.58	1.41
Criterion 9th Grade	4	203	3.70	1.35
Criterion 10th Grade	7	217	3.39	1.32
Criterion 11th Grade	6	212	3.93	1.17
Criterion 12th Grade	5	203	3.66	1.30
GMAT argument	7	758	3.57	1.37
GMAT issue	9	754	3.58	1.34
TOEFL	12	500	4.05	1.09
Overall	**64**	**401**	**3.67**	**1.31**

The mean H1 score across all prompts for most programs is around 3.5, except for somewhat lower mean scores for sixth and seventh grade and higher mean scores for eleventh grade and TOEFL essays. The standard deviations (STD H1) are also quite similar across programs except for TOEFL with lower standard deviation.

Table 2 presents average correlations (across prompts in a program) of each feature for each of the 10 programs analyzed. Correlations proved to be quite similar across programs. Relatively larger differences in correlations can be observed for the mechanics, development, word length, and maximum cosine features.

Table 2: Average Correlations (Across All Prompts in a Program) of Feature Values With H1

Feature	6th	7th	8th	9th	10th	11th	12th	GMAT argument	GMAT issue	TOEFL
Grammar	.58	.59	.64	.56	.64	.61	.60	.58	.60	.56
Usage	.59	.64	.64	.64	.67	.67	.70	.66	.67	.64
Mechanics	.36	.45	.47	.21	.39	.30	.38	.31	.38	.38
Style	.41	.52	.54	.48	.49	.54	.52	.42	.44	.50
Organization	.49	.62	.57	.60	.61	.54	.63	.51	.54	.46
Development	.19	.29	.39	.27	.20	.37	.31	.15	.22	.29
Vocabulary	.49	.59	.59	.57	.56	.62	.64	.56	.58	.58
Word Length	.18	.08	.33	.11	.29	.25	.35	.21	.15	.14
Max. Cos.	.27	.24	.38	.33	.38	.38	.40	.48	.46	.40
Cos. w/6	.48	.41	.46	.42	.38	.44	.41	.66	.66	.56

Table 3 presents the mean, standard deviation, and skewness of scores for each feature. Most of the features have relatively small skewness values.

Table 3: Descriptive Statistics for Feature Distributions

Feature	Mean	STD	Skewness
Grammar	5.16	0.69	-0.85
Usage	5.33	0.60	-1.13
Mechanics	3.86	0.92	0.23
Style	0.91	0.08	-1.58
Organization	1.72	0.50	-1.27
Development	3.76	0.41	0.36
Vocabulary	56.33	5.77	-0.26
Word Length	4.53	0.43	-0.14
Max. Cos.	4.06	1.33	-0.01
Cos. w/6	0.19	0.06	0.27

Modeling Results

Model-building in this section was based on the first human score (H1) and all results are based on a comparison with the second human score (H2). Several types of e-rater models were built to demonstrate the advantages of generic models and non-optimal feature weights. In addition to prompt-specific models (PS), results are shown for generic program-level models based on all 10 features (G10), generic program-level models based on 8 features (G8) without the two prompt-specific vocabulary usage features (max. cos. and cos. w/6), and generic program-level models based on all 10 features but with a single fixed set of relative weights (G10F).

To give a sense of the relative importance of the different features in the regression models, Table 4 presents the relative weights of the features for the G10 models. In general the organization and development features show the highest weights. The last column shows the coefficient of variation of the weights across programs as a measure of the stability of weights. This coefficient is calculated as the ratio of the standard deviation to the average and is expressed in percentages. The table shows relatively small fluctuations in weights across programs, with most coefficients in the range of 20% to 40% except for 66% for style and 96% for cos. w/6. The average weights across all programs in Table 4 (second to last column) were taken as the fixed set of weights for models G10F.

Table 4: Relative Feature Weights (Expressed as Percent of Total Weights) From Program-Level Regression for Prediction of H1

Feature	6th	7th	8th	9th	10th	11th	12th	GMAT argument	GMAT issue	TOEFL	Average	Coef. Var.
Grammar	.12	.02	.09	.04	.11	.07	.05	.09	.09	.09	.08	39
Usage	.16	.09	.06	.08	.10	.08	.08	.11	.07	.10	.09	28
Mechanics	.08	.09	.09	.04	.08	.04	.05	.02	.04	.07	.06	39
Style	.04	.06	.11	.04	.03	.06	.02	.01	.01	.05	.04	66
Organization	.16	.33	.24	.31	.32	.29	.37	.23	.27	.22	.27	22
Development	.09	.20	.19	.17	.19	.22	.26	.14	.17	.18	.18	24
Vocabulary	.04	.09	.08	.05	.07	.06	.06	.07	.08	.10	.07	25
Word Length	.06	.04	.08	.10	.05	.06	.08	.05	.06	.06	.06	26
Max. Cos.	.12	.09	.02	.08	.05	.11	.03	.12	.08	.05	.08	46
Cos. w/6	.14	.00	.05	.09	.01	.00	.00	.16	.13	.09	.07	96

One of the interesting aspects of the results is the relatively minor role of "content" or prompt-specific vocabulary in the scoring models. On average, the two prompt-specific vocabulary features accounted for 15% of the total weights. Even for the GMAT argument prompts, which could

be expected to have a strong content influence[2], the non-content features accounted for more than 70% of the total weights. This suggests that for many types of prompts scoring of structure that is well measured leaves little residual impact to the specific words used in the essay.

Tables 5 and 6 present Kappa and exact agreement results for all types of models. The first column presents Kappa or exact agreement values between H1 and H2 scores. Subsequent columns present each model's result (Kappa or exact agreement value) between the e-rater and H2 scores. The two tables show a similar pattern, whereby the different e-rater scores agree at least as much with H2 as H1 does (except for the GMAT argument program) and the generic models show the same performance as the prompt-specific models. Even the restrictive generic models, the fully generic G8 and the fixed weights G10F, show similar performance.

Table 5: Human Kappas (H1/H2) and H2/e-rater Kappas

		H2/e-rater			
Program	**H1/H2**	**PS**	**G10**	**G8**	**G10F**
Criterion 6th Grade	.27	.31	.30	.30	.33
Criterion 7th Grade	.38	.42	.41	.41	.42
Criterion 8th Grade	.38	.35	.36	.36	.36
Criterion 9th Grade	.33	.36	.37	.32	.36
Criterion 10th Grade	.35	.41	.38	.38	.35
Criterion 11th Grade	.34	.42	.44	.42	.43
Criterion 12th Grade	.39	.43	.43	.43	.41
GMAT argument	.37	.32	.32	.31	.32
GMAT issue	.38	.38	.38	.37	.38
TOEFL	.44	.44	.44	.43	.42

Table 6: Human Exact Agreement (H1/H2) and H2/e-rater Exact Agreements

		H2/e-rater			
Program	H1/H2	PS	G10	G8	G10F
Criterion 6th Grade	.43	.46	.46	.46	.48
Criterion 7th Grade	.52	.54	.54	.54	.54
Criterion 8th Grade	.50	.49	.49	.49	.49
Criterion 9th Grade	.47	.49	.51	.46	.50
Criterion 10th Grade	.49	.53	.51	.52	.49
Criterion 11th Grade	.50	.56	.58	.56	.58
Criterion 12th Grade	.52	.55	.55	.56	.53
GMAT argument	.49	.46	.46	.45	.46
GMAT issue	.50	.51	.51	.50	.51
TOEFL	.59	.59	.59	.58	.58

Alternate-Form Results

Evaluations of AES systems are usually based on single-essay scores. In these evaluations, the relation between two human rater scores and between a human and an automated score are usually compared. Although this comparison seems natural, it is also problematic in several ways.

In one sense this comparison is intended to show the validity of the machine scores by comparing them to their gold standard: the scores they were intended to imitate. However, at least in e-rater V.2, the dependency of machine scores on human scores is very limited since the set of writing features (and their relative importance) is not dependent on human holistic scores. E-rater scores can be computed and interpreted without the human scores.

In another sense the human-machine relation is intended to evaluate the reliability of machine scores, similar to the way the human-human relation is interpreted as reliability evidence for human scoring. However, this interpretation is problematic too. Reliability is defined as the consistency of scores across administrations, but both the human-human and the machine-human relations are based on a single administration of only one essay. Furthermore, in this kind of analysis the machine-human relation would never be stronger than the human-human relation, even if the machine reliability was perfect. This is because the relation between the scores of two human raters on essays written in response to one particular prompt is an assessment of the reliability of human scoring for this prompt, or in other words, of the rater agreement reliability. Any other measure or

scoring method for these prompt essays could not have a stronger relation with a human score than this rater reliability. Finally, this analysis takes into account only one kind of inconsistency between human scores: inter-rater inconsistencies within one essay. It does not take into account inter-task inconsistencies. The machine scores, on the other hand, have perfect inter-rater reliability. All this suggests that it might be better to evaluate automated scores on the basis of multiple essay scores.

The data for this analysis comes from the *Criterion* 6th to 12th grade essays that were analyzed in the previous section. The different prompts in each grade-level were designed to be parallel and exchangeable, and thus they could be viewed as alternate forms. The essays were chosen from the *Criterion* database to include as many multiple essays per student as possible. Consequently it was possible to identify in the set of 7,575 essays almost 2,000 students who submitted two different essays. These essays (almost 4,000 in total, two per student) were used to estimate the alternate-form reliability of human and e-rater scores. These analyses were based on the sub-optimal fixed weights model G10F.

Table 7 presents the alternate-form reliabilities of the e-rater scores, single human scores (H1 and H2), and for the average human score (AHS), for each grade and overall. The table shows that the e-rater score has higher reliabilities than the single human rater does in six out of seven grades. Further, the e-rater score also has equivalent overall reliabilities to the average of two human raters' scores (.59 vs. .58 for the AHS).

Table 7: Alternate-form Reliabilities of Human and e-rater Scores

Grade	N	G10F	H1	H2	AHS
Criterion 6th Grade	285	.68	.48	.63	.65
Criterion 7th Grade	232	.61	.52	.53	.59
Criterion 8th Grade	338	.58	.49	.55	.58
Criterion 9th Grade	280	.41	.45	.27	.41
Criterion 10th Grade	352	.57	.52	.52	.57
Criterion 11th Grade	280	.50	.33	.41	.44
Criterion 12th Grade	226	.74	.63	.70	.74
Overall	1993	.59	.50	.53	.58

The estimation of human and machine reliabilities and the availability of human-machine correlations across different essays make it possible to evaluate human and machine scoring as two methods in the context of a multi-method analysis. Table 8 presents a typical multi-method correlation table. The two correlations below the main diagonal are equal

to the average of the correlations between the first e-rater score and second human score (either single or average of two), and between the second e-rater score and first human score. (Both pairs of correlations were almost identical.) The correlations above the diagonal are the corrected correlations for unreliability of the scores. These correlations were almost identical for single and average of two human scores. The reliabilities of the scores are presented on the diagonal.

Table 8: Multi-method Correlations Across Different Essays

Score	e-rater	Single human rater	AHS
E-rater	.59	.97[3]	.97[3]
Single human rater	.53[2]	.51[1]	—
AHS	.56[2]	—	.58

Note: Diagonal values are alternate-form reliabilities: correlation between two essays.
[1]Average of H1 and H2 reliabilities.
[2]Average of correlations between e-rater on one essay and human scores on another.
[3]Correlations corrected for unreliability of scores: raw correlation divided by square-root of the product of reliabilities.

The main finding presented in Table 8 is the high corrected-correlation (or true-score correlation) between human- and machine-scores—.97. This high correlation is evidence that e-rater scores, as an alternative method for measuring writing ability, are measuring a very similar construct as the human scoring method of essay writing. These findings can be compared to the relationship between essay writing tests and multiple-choice tests of writing (direct and indirect measures of writing). Breland and Gaynor (1979) studied the relationship between the Test of Standard Written English (TSWE), a multiple-choice test, and performance on writing tasks, on three different occasions. 234 students completed all tasks and the estimate obtained for the true-score correlation between the direct and indirect measures of writing was .90. This study concluded that the two methods of assessment of writing skills tend to measure the same skills.

Table 9 shows the results from another interesting analysis that is made possible with the multiple-essay data, namely the reliability of individual features. The table presents the alternate-form reliability of each feature.

Table 9: Alternate-Form Reliabilities of Individual Features

Feature	Reliability
Grammar	.45
Usage	.45
Mechanics	.46
Style	.43
Organization	.48
Development	.36
Vocabulary	.44
Word Length	.47
Max. Cos.	.15
Cos. w/6	.09

The table shows that most of the features have reliabilities in the mid 40s. The only features that stand out are the prompt-specific vocabulary usage features with very low reliabilities.

Summary and Future Directions

e-rater V.2 is a new AES system with a small and meaningful feature set and a simple and intuitive way of combining features. These characteristics allow a greater degree of user judgmental control over the scoring process such as determination of the relative importance of the different writing dimensions measured by the system. It also allows greater standardization of scoring, specifically allowing a single scoring model to be developed for all prompts of a program or assessment. These aspects contribute to the validity of e-rater because they allow a greater understanding and control over the automated scores.

In this paper we have provided evidence for the validity of e-rater V.2 that cover all three of Yang's (2002) methods for system validation: single-essay agreement results with human scores, correlations between scores on different prompts, and descriptions of the scoring process and how it contributes to the validity of the system. The analysis results presented in the paper show that e-rater scores have significantly higher alternate-form reliability than human scores while measuring virtually the same

construct as the human scores. In the next sections, we mention two important ongoing efforts that should drive future AES development and introduce two specific enhancements that are made possible by e-rater's characteristics.

Improvements in Features

The e-rater measures described here do not obviously cover all the important aspects of writing quality and do not perfectly measure the dimensions that it does cover. Improving the features used for scoring and devising new features that tap more qualities of writing performance is an important ongoing process. Features that can provide useful instructional feedback to students should be preferred because they allow a more valid basis for AES.

Detection of Anomalous Essays

Detection of anomalous and bad-faith essays is important for improving the perception of AES in the public. More systematic efforts are needed to characterize the types of anomalies that can be created by students. One of the most obvious ways in which students can prepare for an essay-writing test (human- or machine-scored) is by memorizing an essay on a different topic than the (still unknown) test topic. AES systems must be able to detect such off-topic essays. Higgins, Burstein, and Attali (2006) studied different kinds of off-topic essays and their detection.

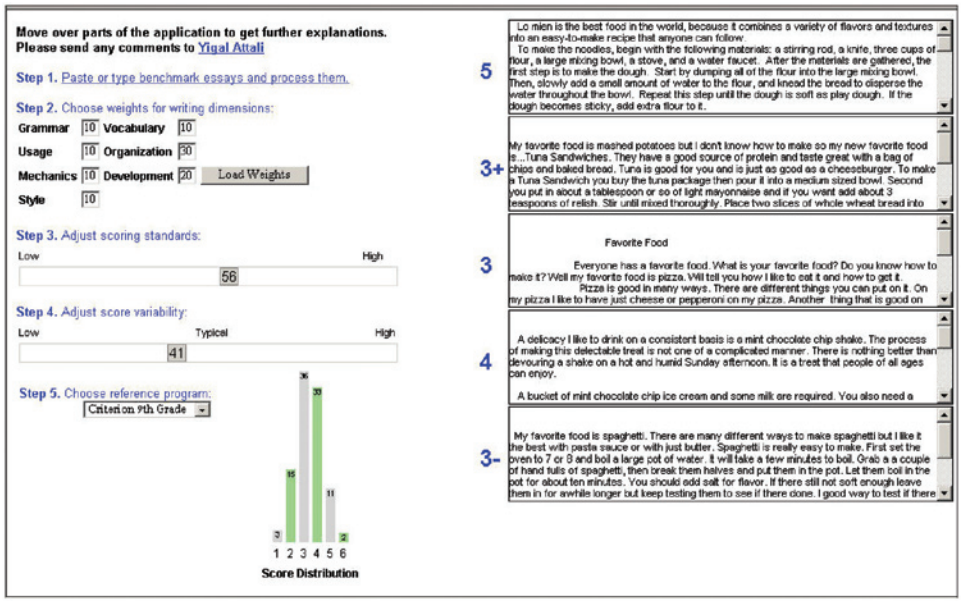
AES On-The-Fly

The score modeling principles of e-rater V.2 can be applied in a radical way to enable AES that is completely independent of statistical analysis of human-scored essays (Attali, 2006). To understand how this can be done one only needs to review the three elements of modeling. We have already suggested that relative weights can be determined judgmentally. The distributions of feature values are commonly estimated with new essay data (specific either to the prompt, or in program-level models of e-rater V.2 from other prompts). The idea is to use previously collected data from a diverse set of programs to construct distributions of features that can be used with any scoring standards. Finally, a small set of benchmark essays supplied by a user are used to identify appropriate scaling parameters, or in other words to set the appropriate scoring standards.

This approach can be characterized as adjusting an anchor model instead of the traditional way of developing models from scratch every time a scoring model is required. Figure 2 shows a screen-capture from a web-application that applies the on-the-fly modeling approach. After

loading a few benchmark essays (step 1), the user determines relative weights to each of the dimensions measured by e-rater (step 2). Then the scoring standards (step 3) and score variability (the difference in scores between essays with different qualities, step 4) are selected. Finally, the user can even select a reference program (*Criterion*'s 9th grade is shown) to immediately see the effect of the changing standards on the entire distribution of scores for this program. GRE® test developers successfully used this application to develop a scoring model for the "Present Your Perspective on an Issue" task based on five benchmark essays only (Attali, 2006).

Figure 2: On-The-Fly Modeling Application



Objective Writing Scales

An exciting possibility that a standards-based AES system like e-rater V.2 enables is the development of an objective writing scale that is independent of specific human rubrics and ratings (Attali, 2005a). Meaningful features that are used in consistent and systematic ways allow us to describe the writing performance of groups and individuals within these groups on such a single scale. This scale would be based on the feature distributions for these groups.

Collecting representative information on writing performance in different grades would enable us, for example, to acquire a better understanding of the development of writing performance along the school years. It might also enable us to provide scores on this developmental scale

("your performance is at the 8th grade level") instead of the standard 1–6 scale. It might also show differential growth paths for different dimensions of writing.

Comparisons of different ethnic groups (and other background classifications) could reveal differences in the relative strength of the various writing dimensions. For example, Attali (2005b) found a significant difference in the patterns of writing performance of TOEFL examinees from Asia and from the rest of the world. Asian students show higher organization scores and lower grammar, usage, and mechanics scores, compared to other students. Consequently, decisions with regard to the relative weights of these dimensions will have an effect on the overall performance of different ethnic groups.

Endnotes

1 An example of a descriptive prompt could be "Imagine that you have a pen pal from another country. Write a descriptive essay explaining how your school looks and sounds, and how your school makes you feel." An example of a persuasive prompt could be "Some people think the school year should be lengthened at the expense of vacations. What is your opinion? Give specific reasons to support your opinion."

2 In this task the student is required to criticize a flawed argument by analyzing the reasoning and use of evidence in the argument.

References

Attali, Y. (2005a, November). *New possibilities in automated essay scoring.* Paper presented at the CASMA-ACT Invitational Conference: Current Challenges in Educational Testing, Iowa City, IA.

Attali, Y. (2005b). Region Effects in e-rater and human discrepancies of TOEFL Scores. Unpublished Manuscript.

Attali, Y. (2006, April). *On-the-fly automated essay scoring.* Paper presented at the Annual Meeting of the National Counsel on Measurement in Education, San Francisco, CA.

Bennett, R. E. & Bejar, I. I. (1998).Validity and automated scoring: It's not only the scoring. *Educational Measurement: Issues and Practice, 17*(4), 9–17.

Bennett, R. E. & Ward, W. C. (Eds.). (1993). *Construction versus choice in cognitive measurement: Issues in constructed response, performance testing, and portfolio assessment*. Hillsdale, NJ: Lawrence Erlbaum.

Breland, H. M., Jones, R. J., & Jenkins, L. (1994). *The College Board vocabulary study* (College Board Report No. 94–4; Educational Testing Service Research Report No. 94–26). New York: College Entrance Examination Board.

Breland, H. M. & Gaynor, J. L. (1979). A comparison of direct and indirect assessments of writing skill. *Journal of Educational Measurement, 16,* 119–128.

Burstein, J. & Wolska, M. (2003). Toward evaluation of writing style: Overly repetitious word use in student writing. In *Proceedings of the 10th Conference of the European Chapter of the Association for Computational Linguistics*. Budapest, Hungary.

Burstein, J., Marcu, D., & Knight, K. (2003). Finding the WRITE stuff: Automatic identification of discourse structure in student essays. *IEEE Intelligent Systems: Special Issue on Natural Language Processing, 18*(1): 32–39.

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. C., Kukich, K., Wolff, S., Lu, C., & Chodorow, M. (1998, April). *Computer analysis of essays.* Paper presented at the annual meeting of the National Council of Measurement in Education, San Diego, CA.

Chodorow, M. & Leacock, C. (2000). *An unsupervised method for detecting grammatical errors.* Paper presented at the 1st Annual Meeting of the North American Chapter of the Association for Computational Linguistics.

Elliot, S. M. (2001, April). *IntelliMetric: From here to validity.* Paper presented at the annual meeting of the American Educational Research Association, Seattle, WA.

Embretson, S. E. (1983). Construct validity: Construct representation versus nomothetic span. *Psychological Bulletin, 93,* 179–197.

Higgins, D., Burstein, J., & Attali, Y. (2006). Identifying Off-Topic Student Essays without Topic-Specific Training Data. In J. Burstein and C. Leacock (eds), *Special Issue of Natural Language Engineering on Educational Applications Using NLP,* to appear.

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. (2001, February). *The intelligent essay assessor: Putting knowledge to the test.* Paper presented at the Association of Test Publishers Computer-Based Testing: Emerging Technologies and Opportunities for Diverse Applications conference, Tucson, AZ.

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.

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

Page, E. B. (1994). Computer grading of student prose, using modern concepts and software. *Journal of Experimental Education, 62,* 127–142.

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

Powers, D. E., Burstein, J. C., Chodorow, M., Fowles, M. E., & Kukich, K. (2001). Stumping e-rater: Challenging the validity of automated essay scoring (GRE Board Professional Rep. No. 98–08bP, ETS Research Rep. No. 01–03). Princeton, NJ: Educational Testing Service.

Powers, D. E., Burstein, J. C., Chodorow, M., Fowles, M. E., & Kukich, K. (2002). Comparing the validity of automated and human scoring of essays. *Journal of Educational Computing Research, 26,* 407–425.

Salton, G., Wong, A., & Yang, C.S. (1975). A vector space model for automatic indexing. *Communications of the ACM, 18,* 613–620.

Yang, Y., Buckendahl, C. W., Juszkiewicz, P. J., & Bhola, D. S. (2002). A review of strategies for validating computer-automated scoring. *Applied Measurement in Education, 15,* 391–412.

Author Biographies

Yigal Attali is a Research Scientist at Educational Testing Service in Princeton, NJ. Yigal is interested in exploring cognitive aspects of assessment and in the implementation of technology in assessment. His current research includes development and evaluation of automated essay scoring (he is a primary inventor of e-rater® V.2) and assessment of feedback mechanisms in constructed response tasks. He received his B.A. in computer sciences and his Ph.D. in psychology from the Hebrew University of Jerusalem.

Jill Burstein is a principal development scientist in the Research & Development Division at Educational Testing Service in Princeton, NJ. Her area of specialization is computational linguistics. She led the team that first developed e-rater®, and she is a primary inventor of this automated essay scoring system. Her other inventions include several capabilities in *Criterion*^{SM}, including an essay-based discourse analysis system, a style feedback capability, and a sentence fragment finder. Currently, her research involves the development of English language learning capabilities that incorporate machine translation technology. Her research interests are natural language processing (automated essay scoring and evaluation, discourse analysis, text classification, information extraction, and machine translation); English language learning, and the teaching of writing. Previously, she was a consultant at AT&T Bell Laboratories in the Linguistics Research Department, where she assisted with natural language research in the areas of computational linguistics, speech synthesis, and speech recognition. She received her B.A. in Linguistics and Spanish from New York University, and her M.A. and Ph.D. in linguistics from the City University of New York, Graduate Center.

J·T·L·A

The Journal of Technology, Learning, and Assessment

www.jtla.org

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