Research Report

Automated Essay Scoring With E-rater® v.2.0

Yigal Attali

Jill Burstein

Research & Development

November 2005
RR-04-45

Automated Essay Scoring With E-rater$^{®}$ v.2.0

Yigal Attali and Jill Burstein

ETS, Princeton, NJ

November 2005

Automated Essay Scoring With E-rater® v.2.0

Yigal Attali and Jill Burstein

ETS, Princeton, NJ

Abstract

E-rater® has been used by ETS for automated essay scoring since 1999. This paper describes a new version of e-rater (v.2.0) that differs from the previous one (v.1.3) with regard to the feature set and model building approach. The paper describes the new version, compares the new and previous versions in terms of performance, and presents evidence on the validity and reliability of scores produced by the new version.

Key words: Automated essay scoring, e-rater, Criterion^{SM}

E-rater® has been used by ETS for automated essay scoring since February 1999. Burstein, Chodorow, and Leacock (2003) described the operational system put in use for scoring the Graduate Management Admission Test® Analytical Writing Assessment (GMAT® AWA) and for essays submitted to ETS's writing instruction application, Criterion℠ Online Essay Evaluation Service. Criterion is a Web-based service developed by ETS to evaluate a student's writing skill and provide instantaneous score reporting and diagnostic feedback. Criterion contains two complementary applications. The scoring application uses the e-rater engine. The second application, Critique, is composed of a suite of programs that evaluates and provides feedback for errors in grammar, usage, and mechanics; identifies the essay's discourse structure; and recognizes undesirable stylistic features. Additional details on these programs will be provided in the description of the new version of e-rater that follows.

The operational system of e-rater (v.1.3) was trained on a sample of essays written on the same topic that had been scored by human readers. It measured more than 50 features in all and then computed a stepwise linear regression to select those features that made a significant contribution to the prediction of essay scores. For each essay question or prompt, the result of training was a regression equation that could be applied to the features of a new essay written to the specific topic to produce a predicted value. This value was rounded to the nearest whole number to yield the score.

This paper describes a newer automated essay scoring system that will be referred to in this paper as e-rater version 2.0 (e-rater v.2.0). This new system differs from e-rater v.1.3 with regard to the feature set used in scoring, the model building approach, and the final score assignment algorithm. These differences result in an improved automated essay-scoring system.

The New Feature Set

The development of the new feature set used with e-rater v.2.0 was based on information extracted from e-rater v.1.3 and from the qualitative feedback of Criterion's writing analysis tools. In e-rater v.1.3, the feature set included approximately 50 features, and typically 8 to 12 features were selected and weighted for a specific model using stepwise linear regression. An analysis of the e-rater v.1.3 features revealed that some of them were implicitly measuring essay length (i.e., number of words in the essay) or had a nonmonotonic relationship with the human score. Features in e-rater v.2.0 were created by standardizing some features with regard to essay length, by altering the definition of others to take into account the nonmonotonic relationship with the human score, and also by creating new features. Below is a description, by category, of the 12 features included in the new feature set.

Errors in Grammar, Usage, Mechanics, and Style (Four Features)

Criterion produces feedback about a total of 33 errors in four categories: grammar, usage, mechanics, and comments about style. Most of these features are identified using natural language processing techniques (Burstein et al., 2003). Counts of the errors in the four categories form the basis for four features in e-rater v.2.0, referred to herein as *grammar*, *usage*, *mechanics*, and *style*. However, the features actually computed are the rates of errors in the four categories and are calculated for each category by counting the total number of errors in that category and dividing this by the total number of words in an essay.

Organization and Development (Two Features)

In addition to the various errors, the Criterion feedback application automatically identifies sentences in the essay that correspond to the following essay-discourse categories, using natural language processing: background, thesis, main ideas, supporting ideas, and conclusion (Burstein, Marcu, & Knight, 2003). Two features were derived from this feedback information.

An overall development score (referred to in what follows as *development*) is computed by summing up the counts of the thesis, main points, supporting ideas, and conclusion elements in the essay. An element is the longest consecutive number of sentences assigned to one discourse category. There are two exceptions in making these counts: Supporting ideas elements are counted only when they immediately follow a main point element, and main point elements are restricted to three different elements per essay. These restrictions follow the five-paragraph essay strategy for developing writers that was adopted in Criterion. According to this strategy, novice writers should typically include in their essay an introductory paragraph, a three-paragraph body in which each paragraph consists of a main point and supporting idea elements, and a concluding paragraph. Notice that the background element is not included in the five-paragraph strategy.

In e-rater v.2.0, the development score is defined as the above sum minus 8. This development score may be interpreted as the difference or discrepancy between actual and optimal development. A score of –8 means that there are no required elements, whereas a score of 0 means that all required elements (thesis, conclusion, three main points, and corresponding supporting ideas) are present and there is no discrepancy between optimal and existing development.

The second feature derived from Criterion's organization and development module is the average length (in number of words) of the discourse elements in the essay (referred to as *AEL,* average discourse element length).

Lexical Complexity (Three Features)

Three features in e-rater v.2.0 are related specifically to word-based characteristics. The first is the ratio of number of word types (or forms) to tokens in an essay (referred to as *type/token*). For example, in "This essay is a long, long, long essay." there are five word types (*this*, *essay*, *is*, *a*, and *long*) and eight tokens (*this, essay, is, a, long, long, long,* and *essay*). So the type/token ratio is 5/8, or .625. The purpose of this feature is to count the number of unique words in the essay and standardize this count with the total number of words in the essay.

The second feature is a measure of vocabulary level (referred to as *vocabulary*). Each word in the essay is assigned a vocabulary level value based on Breland's standardized frequency index (Breland, Jones, & Jenkins, 1994), and the fifth lowest standardized frequency index value is used in e-rater v.2.0 to estimate the vocabulary level of the essay.

The third feature is the average word length in characters across all words in the essay (referred to as *AWL*, average word length).

Prompt-Specific Vocabulary Usage (Two Features)

Vocabulary usage features were used in e-rater v.1.3 to evaluate word usage in a particular essay in comparison to word usage in essays at the different score points. (The number of score points is usually 6.) To do this, content vector analysis (Salton, Wong, & Yang, 1975) was used. Content vector analysis is applied in the following manner in e-rater. First, each individual essay and a set of training essays from each score point are converted to vectors whose elements 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* is determined as follows:

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

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.2.0, 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, usually 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.

Essay Length (One Feature)

As the following analyses will show, essay length is the single most important objectively calculated variable in predicting human holistic scores. In e-rater v.2.0, it was decided to explicitly include essay length (measured in number of words) in the feature set, thus making it possible to control its importance in modeling writing ability, and at the same time making an effort to minimize the effect of essay length in the other features in the feature set.

E-rater v.2.0 Model Building and Scoring

In e-rater v.1.3, models were always prompt-specific. That is, models were built specifically for each topic, using data from essays written to each of the particular prompts and scored by human raters. This process requires significant data collection and human reader scoring—both time-consuming and costly efforts. In addition, e-rater v.1.3 models were based on a variable subset of from 8 to 12 predictive features that were selected by a stepwise linear regression from a larger set of approximately 50 features.

E-rater v.2.0 models are more uniform in nature across prompts and testing programs. Consequently, it is easier for any audience to understand and interpret these models. The most important aspect of the new system that contributes to this uniformity is the use of a reduced feature set that is fixed in nature. Because the number of features is small, and each one significantly contributes to the goal of predicting the human score, it is possible to use a multiple regression approach for modeling whereby the fixed feature set is present in all of the models. One advantage of this aspect of the new system is that since one knows what features will be in a model, it is possible to specify the weight of some or all of the features in advance, instead of using regression analysis to find optimal weights. It is important to be able to control feature weights when there are theoretical considerations related to various components of writing ability.

The following discussion outlines the way optimal and fixed weights are combined in e-rater v.2.0.

Combining Optimal and Fixed Weights in Multiple Regression

Below is the procedure for producing a regression equation that predicts the human score with n features of which the first k will have optimized weights and the last $n - k$ will have fixed predetermined weights.

1. Apply a suitable linear transformation to the features that have negative correlations with the human score in order to have only positive regression weights.

2. Standardize all features and the predicted human score.

3. Apply a linear multiple regression procedure to predict the standardized human score from the first k standardized features and obtain standardized weights for these features, labeled s_i ($1 \leq i \leq k$).

4. Find the standardized weights for the last $n - k$ features. Initially, the fixed weights are expressed as percentages of the sum of standardized weights for all features, and labeled p_i ($k+1 \leq i \leq n$). For example, if there are two predetermined weights in a set of 12 features (the 11th and 12th features are predetermined) and p_{11} and p_{12} are .1 and .2,

respectively, then s_{11}, the standardized weight for the eleventh feature (which is still unknown), should be equal to 10% of the sum of s_1 - s_{12}, s_{12} should be equal to 20% of s_1 - s_{12}, and the sum of s_1 - s_{10} should account for the remaining 70% of the standardized weights. This means that the standardized weights for the last *n* - *k* features should be found by applying the following formula to the last *n* - *k* features:

$$s_i = \frac{p_i \sum_{j=1}^{k} s_j}{1 - \sum_{j=k+1}^{n} p_j} \qquad k+1 \le i \le n$$

5. To find the unstandardized weights for all features (labeled w_1 - w_n), multiply s_i by the ratio of the standard deviation for the human score to the standard deviation for the feature.

6. Compute an interim predicted score as the sum of the product of feature values and weights w_1 - w_n.

7. Regress the interim predicted score to the human score and obtain an intercept, *a*, and a weight, *b*. The intercept will be used as the final intercept.

8. Determine the final unstandardized weights by multiplying *a* by w_i ($1 \le i \le n$).

The above procedure can also be applied when all weights are predetermined—for example, when content experts set the relative importance of all features judgmentally based on theory and experience. The combined use of a small feature set and more uniform modeling procedures may have a beneficial effect on the interpretability, reliability, and validity of automated scores.

Generic Model Building

Conventionally, e-rater has used a prompt-specific modeling approach in which a new e-rater model is built for each topic. To build prompt-specific models, however, requires a sample of scores from human-rated essays for each topic. In e-rater v.1.3, a sample of at least 500 scores from essays rated by human readers was required in a predetermined score distribution. At least 265 essays were used for model building, and the remaining set was used for cross-validating the model. In addition, such prompt-specific models obviously differ for different prompts from the same program, even though these prompts are scored by the same rubrics and scoring standards by human raters. In prompt-specific modeling, the relative weights of features vary between prompts and even the sign of weights might be reversed for some prompts. This surely cannot contribute to the interpretability of automated essay scores in the writing experts' community.

However, with e-rater's new small and fixed feature set, such extensive prompt-specific modeling may not be necessary or even useful. A single regression model that is used with all prompts from one program may perform, statistically, as well as models that are optimized for each prompt separately, especially when the models' performance is evaluated in a cross-validation sample.

The primary reason that such generic models might work as well as prompt-specific models is that most aspects of writing ability measured by e-rater v.2.0 are topic-independent. For example, if eight discourse units in a GMAT essay are interpreted as evidence of good writing ability, then this interpretation should not vary across different GMAT prompts. The same is true with rates of grammar, usage, mechanics, and style errors: The interpretation of 0%, 1%, or 5% error rates as evidence of writing quality should stay the same across different GMAT prompts. It also is important to note that the rubrics for human scoring of essays are themselves generic in that the same rules apply to all prompts within a program.

Consistent with this, we have found that it is possible to build generic models based on the feature set in v.2.0 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.

It is important to note that a generic regression model does not mean that different prompts have the same difficulty level, since only the *interpretation* of levels of performance for individual features are the same across prompts, not the levels of examinee performance themselves. For example, it may be true that a specific topic is harder for students to write about, and this will be reflected in the development scores of these students. E-rater generic modeling can also incorporate prompt-specific vocabulary usage information through the two vocabulary-based features (*max. cos.* and *cos. w/6*). Again, it is important to distinguish between the *computation* of these two features, which must be based on a prompt-specific training sample, and the *interpretation* of the values of these features, which can be based on generic regression weights.

It also is possible to build generic e-rater models that do not contain the two prompt-specific vocabulary usage features and can thus be applied to new prompts with no training at all. This can be done by setting the weights for these two features to zero, thus excluding these features in model building. Recall that in e-rater v.2.0 it is possible to set the weights of the features instead of estimating them in the regression analysis. Setting some feature weights to zero is analogous to discarding these features from the feature set.

Score Assignment

The last step in assigning an e-rater score is the rounding of the continuous regression model score to the six scoring guide categories. In e-rater v.1.3, the cutoff values were simply the half points between whole values. For example, an essay receiving an e-rater score in the range of 3.50 to 4.49 would be assigned a final score of 4. However, this method of rounding may not be optimal with respect to the goal of maximizing machine-human agreement. In the framework of signal detection theory (Swets, 1996), the problem of determining the optimal cutoff between two adjacent human scores i and j can be modeled by the two distributions of (unrounded) e-rater scores (one for the lower i score, which can be labeled *noise*, one for the higher j score, which can be labeled *signal*) that partially overlap with respect to these scores.

The optimal cutoff criterion depends on two factors: the base rates (i.e., prior probabilities) of signal and noise and the relative costs of errors (or benefits of successes). If the base rates of signal and noise are equal and the relative costs of errors are equal, too, then the optimal criterion lies at the balance point of the e-rater score where the two distributions intersect. However, the less frequent the signal becomes (relative to the noise), the higher the criterion must become, because there are a relatively large number of false alarms (deciding that the stimulus is a signal when it is really noise) from the huge distribution of noise. In most writing assessments, the base rates of midscores are higher than either high or low scores.

The second important factor to decide on an optimal cutoff point concerns the relative costs of false alarms and misses and the benefits of hits and rejections. If the costs of false alarms (relative to misses) and the benefits of rejections (relative to hits) are large, then again the criterion must become higher. With most writing assessments, the correct assignment of high scores is considered more important than correct assignment of low scores. If $P(x \mid s)$ and $P(x \mid n)$

are denoted as the height of the signal and noise distributions at a particular point *x*, *P*(*s*) and *P*(*n*) as the base rates of signal and noise, and *v*(·) as the value (positive for benefits and negative for costs) of each decision outcome, the optimal cutoff *c* is found from the equation:

$$\frac{P(c \mid s)}{P(c \mid n)} = \frac{P(n)}{P(s)} \times \frac{v(rej.) + v(f.a.)}{v(hit) + v(miss)}$$

In e-rater v.2.0, this formula for finding optimal cutoffs is applied by assuming a normal distribution of (unrounded) e-rater scores for each human score level and by estimating the mean and standard deviation of the distribution from the field distributions of the training samples. The benefits of correct identification of the highest scores are sometimes increased relative to the benefits of middle scores. This may raise the rate of correct assignments of high scores when the base rate of these scores is very low. The resulting cutoffs between two adjacent scores *i* and *j* are usually found around the midpoint between the average (unrounded) e-rater score for essays that are scored by humans as *i* and between the average (unrounded) e-rater score for essays that are scored by humans as *j*.

Analyses of the Performance of E-rater v.2.0

The analyses that will be presented in this paper are based on essays from various user programs. We used 6th- through 12th-grade user data for Criterion and human-scored essay data for GMAT issue essays (where students are asked to present their opinion on an issue), GMAT argument essays (where students are asked to analyze the reasoning in an argument and find its logical flaws), and TOEFL® (Test of English as Foreign Language™) essays.. The 6th- through 12th-grade essays were extracted from the Criterion database and scored by trained human readers (two to three readers per essay; third readers were used to resolve score discrepancies of 2 or more points) according to grade-specific rubrics. The GMAT and TOEFL data also included similar scoring by humans. All human scoring rubrics are on a 6-point scale from 1 to 6.

Table 1 presents descriptive statistics of these essays. The average human score (AHS) was computed by averaging the first two human scores that were available for each of the essays. Overall, 64 different prompts and almost 18,000 essays were analyzed.

Table 1

Descriptive Statistics on Essays and Average Human Score (AHS)

Program	Number of prompts	Mean number of essays per prompt	Mean AHS	STD AHS
Criterion 6th grade	5	203	3.01	1.16
Criterion 7th grade	4	212	3.21	1.20
Criterion 8th grade	5	218	3.50	1.29
Criterion 9th grade	4	203	3.65	1.24
Criterion 10th grade	7	217	3.39	1.23
Criterion 11th grade	6	212	3.90	1.08
Criterion 12th grade	5	203	3.61	1.22
GMAT argument	7	493	3.54	1.18
GMAT issue	9	490	3.56	1.17
TOEFL	12	197	3.60	1.17
Overall	64	278	3.53	1.20

The mean AHS across all prompts for most programs is around 3.5, except for somewhat lower mean scores for 6th and 7th grade and higher mean scores for 11th grade. The standard deviations (STD AHS) are also quite similar between programs.

Table 2 presents average correlations (across prompts in a program) of each feature for each of the 10 programs analyzed. Correlations proved to be very similar across programs. One exception may be the apparent trend in correlations for the maximum cosine value (10th feature in Table 2), with lower correlations for the lower grades.

Table 3 presents the average feature values for each AHS from 1.0 to 6.0. The average feature values of scores are presented relative to the average feature score for an AHS of 1.0. This was done to provide a common range of scores for comparison. The two last columns also present the original mean and standard deviation of scores for each feature. Except for one case (usage scores between AHS of 1.0 and 1.5), the average scores are monotonically decreasing as AHS is increasing.

Table 2

Average Correlations (Across All Prompts in a Program) of Feature Values With AHS

								GMAT	GMAT	
Feature	6th	7th	8th	9th	10th	11th	12th	argument	issue	TOEFL
Grammar	-0.22	-0.16	-0.13	-0.15	-0.23	-0.18	-0.21	-0.28	-0.28	-0.38
Usage	-0.11	-0.16	-0.16	-0.19	-0.24	-0.23	-0.26	-0.15	-0.12	-0.14
Mechanics	-0.38	-0.34	-0.35	-0.22	-0.39	-0.28	-0.41	-0.37	-0.40	-0.46
Style	-0.49	-0.56	-0.58	-0.52	-0.51	-0.57	-0.54	-0.40	-0.44	-0.54
Development	0.58	0.67	0.64	0.65	0.65	0.60	0.67	0.51	0.56	0.59
AEL	0.12	0.18	0.25	0.17	0.09	0.25	0.19	0.08	0.12	0.14
Type/token	-0.37	-0.49	-0.43	-0.49	-0.45	-0.42	-0.47	-0.44	-0.34	-0.28
Vocabulary	-0.49	-0.51	-0.50	-0.58	-0.44	-0.44	-0.49	-0.36	-0.48	-0.42
AWL	0.24	0.10	0.37	0.08	0.31	0.28	0.38	0.19	0.14	0.23
Max. cos.	0.15	0.07	0.15	0.09	0.20	0.24	0.40	0.42	0.32	0.41
Cos. w/6	0.43	0.40	0.37	0.32	0.32	0.32	0.43	0.31	0.34	0.58
Essay length	0.74	0.82	0.79	0.81	0.79	0.74	0.83	0.71	0.79	0.82

Note. AEL = average discourse element length; AWL = average word length.

To give a sense of the relative importance of the different features in the regression models, Table 4 presents the standardized weights of the first 11 features when a regression analysis for prediction of AHS was performed for each program separately. The analysis does not include the essay length feature because in all subsequent modeling the weight of this feature was controlled. The standardized weights presented in the table were scaled so that the total of the weights would sum to 1 (that is, each column sum is 1 or 100%). The table shows similar weights across programs with no significant trends (monotonically increasing or decreasing weights) from low to higher grades. The more important features in these models are the development score, followed by average element length, style, average word length, mechanics, and vocabulary.

Table 3

Average Feature Values (Relative to the Average of AHS of 1.0) per AHS and Overall Mean and Standard Deviation

	AHS												
	1.0	1.5	2.0	2.5	3.0	3.5	4.0	4.5	5.0	5.5	6.0	Mean	STD
Grammar	1.00	.50	.39	.35	.29	.24	.21	.18	.16	.14	.12	0.0048	0.0121
Usage	1.00	1.06	.92	.88	.74	.63	.51	.46	.40	.34	.30	0.0028	0.0054
Mechanics	1.00	.64	.46	.39	.32	.27	.22	.20	.17	.16	.14	0.0274	0.0381
Style	1.00	.86	.70	.56	.49	.41	.34	.28	.22	.			
Development	1.00	.87	.76	.62	.53	.41	.33	.24	.21	.17	.15	-3.0	2.4
AEL[a]	1.00	.75	.68	.67	.68	.68	.68	.65	.61	.56	.53	47.3	26.3
Type/token	1.00	.90	.86	.83	.81	.79	.78	.76	.75	.75	.75	0.6322	0.1038
Vocabulary	1.00	.91	.88	.84	.83	.80	.78	.75	.73	.71	.68	41.3	7.5
AWL [a]	1.00	1.00	.99	.98	.97	.96	.95	.94	.94	.94	.92	4.5	0.5
Max. cos. [a]	1.00	1.00	.95	.91	.88	.83	.82	.78	.75	.73	.71	4.1	1.2
Cos. w/6 [a]	1.00	.76	.65	.60	.54	.52	.49	.48	.45	.44	.42	0.1124	0.0525
Essay length [a]	1.00	.60	.46	.37	.33	.27	.24	.20	.18	.16	.14	263.7	128.9

Note. AEL = average discourse element length; AWL = average word length.

[a] Scale of feature reversed by multiplying values by -1.

Table 4

Standardized Feature Weights (Expressed as Percent of Total Weights) From Program-Level Regression for Prediction of AHS

								GMAT	GMAT		
Feature	6th	7th	8th	9th	10th	11th	12th	arg.	issue	TOEFL	Average
Grammar	.08	.02	.06	.07	.09	.05	.06	.07	.06	.06	.06
Usage	.04	.06	.02	.03	.02	.03	.03	.03	.01	.01	.03
Mechanics	.11	.11	.08	.04	.10	.08	.08	.11	.13	.07	.09
Style	.08	.11	.10	.13	.10	.12	.06	.10	.12	.09	.10
Development	.28	.35	.26	.26	.23	.29	.27	.21	.22	.25	.26

(Table continues)

Table 4 (continued)

Feature	6th	7th	8th	9th	10th	11th	12th	GMAT arg.	GMAT issue	TOEFL	Average
AEL	.12	.18	.18	.11	.07	.17	.17	.14	.14	.17	.14
Type/token	.00	.03	.03	.08	.09	.04	.05	.08	.06	.05	.05
Vocabulary	.08	.09	.10	.14	.10	.05	.07	.05	.08	.08	.08
AWL	.12	.07	.15	.07	.11	.12	.12	.09	.08	.09	.10
Max. cos.	.03	.00	.00	.00	.06	.05	.03	.10	.06	.05	.04
Cos. w/6	.06	.00	.03	.07	.03	.00	.05	.02	.04	.08	.04

Note. AEL = average discourse element length; AWL = average word length.

GMAT Model Building Results

This section presents model-building results for e-rater v.2.0 and a comparison with e-rater v.1.3. The analyses were conducted on the GMAT data set presented previously, which included seven argument and nine issue prompts. The human score that was used in these analyses was the human resolved score (HRS), customarily used in this program for scoring essays. The HRS is the average of the first two human scores rounded up to the nearest whole score, unless the difference between the first two human scores is more than one score point; in this case, a third human score is obtained, and the HRS is based on the third score and the score most similar to this third score. Three types of e-rater v.2.0 models were used in these analyses. In addition to prompt-specific models, results are shown for generic models with and without the two prompt-specific vocabulary usage features (max. cos. and cos. w/6).

All analyses presented in this section are based on separate training and cross-validation samples, and results are always based on the cross-validation sample. The cross-validation data are composed of an independent sample of essay responses that have not been used for model building. Performance on the cross-validation set reveals what kind of system performance can be expected in the field. For prompt-specific models (both v.1.3 and v.2.0), a two-fold cross-validation approach was used. In this approach, the data were randomly divided into two (approximately) equal data sets. First, one half of the data were used for model building and the second half were used for cross-validation. This procedure was then repeated, but the set used for cross-validation in the previous run was now used for model building, and the one used for model building was used for cross-validation.

For model building and evaluation of the generic models, an *n*-fold cross-validation procedure was used, where *n* is equal to the number of prompts: 7 for argument and 9 for issue. For each run, *n* - 1 prompts were used for model building, and the *n*th prompt was held out to evaluate (cross-validate) the model built in each fold. The procedure was repeated *n* times.

Table 5 presents average kappa results for the three e-rater v.2.0 model building approaches and for several predetermined weights for essay length. Because of its high correlation with human score (see Table 2), the effect of running a free regression model with essay length as one of the features is to assign a large weight to this feature. On the other hand, building an optimal model from all other features and adding essay length with a predetermined weight has a very small effect on performance. The weights in Table 5 are expressed as percentages of total standardized weights for all features in the model. One can see that in the case of the argument prompts there is a significant increase in kappas when the essay length weight is increased from .0 to .1 and smaller increases up to when the weight is in the range .3-.4. For the issue prompts, there is a noticeable increase from .0 to .1 and a smaller increase from .1 to .2. In the case of the argument prompts, performance decreases when the essay length weight is raised to .5 from .4. The table also shows very similar results between the generic models, in particular the Generic12 model, and prompt-specific models with a slight advantage to the generic models.

Table 5

Average Kappas for E-rater v.2.0

		Essay length weight					
System	GMAT program	0.0	0.1	0.2	0.3	0.4	0.5
Generic10[a]	Argument	0.32	0.34	0.35	0.35	0.36	0.35
	Issue	0.38	0.41	0.42	0.42	0.42	0.42
Generic12[b]	Argument	0.34	0.37	0.38	0.39	0.39	0.38
	Issue	0.42	0.44	0.46	0.44	0.44	0.44
Specific	Argument	0.34	0.37	0.38	0.38	0.39	0.39
	Issue	0.41	0.44	0.44	0.44	0.44	0.43

[a] Generic model without vocabulary usage features—10 features. [b] Generic model with vocabulary usage features—12 features.

Table 6 presents the results from comparing the three v.2.0 models with the v.1.3 model for each prompt type. The table shows that the v.2.0 prompt-specific and Generic12 models outperformed the v.1.3 models, and that in the case of the issue prompts, even the Generic10 models performed better than the v.1.3 model (mean kappa of .42 compared to .40).

Table 6

Kappas and Rates of Exact Agreement for Different Systems

System	ELW	N	Mean kappa	STD kappa	Exact agreement
GMAT argument					
Specific	.2	7	.38	.06	.52
	.3	7	.38	.07	.52
Generic10	.2	7	.35	.08	.50
	.3	7	.35	.08	.50
Generic12	.2	7	.38	.06	.52
	.3	7	.39	.07	.52
v.1.3	-	7	.36	.07	.51
GMAT issue					
Specific	.2	9	.44	.03	.57
	.3	9	.44	.03	.57
Generic10	.2	9	.42	.05	.56
	.3	9	.42	.04	.56
Generic12	.2	9	.46	.05	.58
	.3	9	.44	.04	.57
v.1.3	-	9	.40	.05	.54

Note. ELW = essay length weight.

Reliability of E-rater v.2.0

Evaluations of automated essay scoring systems are usually based on single-essay scores. In these evaluations, the relations 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.2.0, the sense in which machine scores imitate human scores is very limited. The e-rater score is composed of a fixed set of features of writing that are not derived from the human holistic scores. As this paper showed, the combination of the features is not necessarily based on optimal regression weights for the prediction of the human scores, and the difference in performance (relation with human score) between "optimal" and predetermined weights is very small. This means that the machine scores are not dependent on human scores: They 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. But 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. In addition, in this kind of analysis, the machine-human relation would never be stronger than the human-human relation, even if the machine reliability were perfect. This is because the relation between the scores of two human raters to essays written to one 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, interrater inconsistencies within one essay, and not the intertask inconsistencies. The machine scores, on the other hand, have perfect interrater 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 come from the Criterion essays that were analyzed in previous sections. 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 test-retest reliability of human and automated scoring. The computation of automated scores was based, in this analysis, on the average relative weights across programs from Table 4. This was done to avoid overfitting as much as possible. Note that the weights chosen are not only suboptimal on the prompt level,; they are not even the best weights at the grade level. The essay length weight was set to 20%, and since the results in this section are based on correlations, no scaling of scores was performed (since scaling would not change the results).

Table 7 presents the test-retest reliabilities of the automated scores, single human scores, and AHS, for each grade and overall. The table shows that the e-rater score has higher reliabilities than the single human rater (in five out of seven grades) and fairly equivalent reliabilities to the average of two human raters, with overall reliability of .60, higher than that of the AHS (.58).

Table 7

Test-Retest Reliabilities

Grade	N	E-rater	Single human rater	AHS
Criterion 6th grade	285	.61	.48	.65
Criterion 7th grade	231	.63	.52	.59
Criterion 8th grade	334	.54	.49	.58
Criterion 9th grade	280	.40	.45	.41
Criterion 10th grade	352	.52	.52	.57
Criterion 11th grade	280	.44	.33	.44
Criterion 12th grade	225	.76	.63	.74
Overall	1,987	.60	.50	.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 multimethod analysis. Table 8 presents a typical multimethod 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 the single human rater and average of two human scores. The reliabilities of the scores are presented on the diagonal.

Table 8

Multimethod Correlations Across Different Prompts

Score	E-rater	Single human rater	AHS
E-rater	.60	.93	.93
Single human rater	.51	.50	–
AHS	.55	–	.58

Note. Diagonal values are test-retest reliabilities. Values above diagonal are corrected for unreliability of scores.

The main finding presented in Table 8 is the high corrected correlation (or true-score correlation) of .93 between human and machine scores. This high correlation is evidence that e-rater scores, as an alternative method for measuring writing ability, are measuring a construct that is very similar to the human scoring 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).

On three different occasions, Breland and Gaynor (1979) studied the relationship between students' performance on the Test of Standard Written English (TSWE), a multiple-choice test, and their performance on three different open-ended writing tasks. A total of 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 test-retest reliability of each feature alongside the overall correlation with AHS and the relative weights used in this section.

Table 9 shows that the essay length feature has the highest reliability (.56), higher than the reliability of a single human rater and almost as high as the reliability of the entire e-rater score. The reliabilities of the style, development, and AWL features are in the 40s; the reliabilities of the mechanics, AEL, and type/token ratio features are in the 30s; the reliabilities of the vocabulary and cosine 6 correlation features are in the 20s; and finally, the reliabilities of the grammar, usage, and max cosine value features are .16 and lower.

Table 9

Test-Retest Reliabilities of Individual Features

Feature	Test-retest reliability	Weight	Overall correlation with AHS
Grammar	0.07	0.05	0.16
Usage	0.16	0.02	0.20
Mechanics	0.36	0.07	0.34
Style	0.43	0.08	0.55
Development	0.48	0.21	0.65
AEL	0.32	0.12	0.17
Type/token	0.38	0.04	0.44
Vocabulary	0.24	0.07	0.50
AWL	0.47	0.08	0.32
Max. cos.	0.11	0.03	0.22
Cos. w/6	0.25	0.03	0.32
Essay length	0.56	0.20	0.78

Note. AEL = average discourse element length; AWL = average word length.

The comparison between the three columns of Table 9 shows that there is a relatively high positive correlation between all three measures of feature performance: feature reliability, contribution in regression analysis, and simple correlations with AHS. The rank-order correlation between feature reliability and the other two measures is .78 in both cases.

Summary and Future Directions

E-rater v.2.0 uses a small and fixed set of features that are also meaningfully related to human rubrics for scoring essays. This paper showed that this could be exploited to create automated essay scores that are standardized across different prompts without loss in performance. The creation of grade- or program-level models also contributes to the transparency and interpretability of automated scores. Last but not least, standard grade-level models provide an opportunity to interpret automated writing scores in a cross-grade perspective—to compare automated essay scores from different grades on the same cross-grade scale. This was impossible with previous e-rater versions or with human holistic rubrics.

The paper showed that e-rater v.2.0 scores have higher agreement rates with human scores than e-rater v.1.3 scores have. The test-retest reliability of e-rater scores (for a single essay) in a 6th- to 12th-grade population (.60) was higher than the test-retest reliability of a single human rater (.50) and was comparable to the average of two human raters (.58). The true-score correlation between the e-rater and human scores was very high (.93).

There are three main directions for improvements to the current version of e-rater. One line of research that should be pursued is concerned with modifications and enhancements to the set of features used for modeling writing ability. By employing natural language processing and other techniques, it should be possible to capture more of the different writing aspects that are deemed important by theories of writing.

A second line of research is related to modifications and improvements of the modeling process. Haberman (2004) explored statistical transformations of the feature values that might be beneficial for the measurement properties of e-rater. In another direction, the use of regression for combining the features into a single automated score may not be optimal. Since the scoring rubric that is usually used in writing assessments is discrete with typically six (or fewer) levels of performance, the regression score must be rounded to provide the final automated score, and the method of rounding has a substantial effect on the performance of the automated scores. Alternative methods for scaling e-rater scores should be investigated.

The last line of development suggested here is concerned with the identification of discrepant essays that should not be scored with the regular model. Improving this capability is important for establishing the validity of the system for use in high-stakes testing. Although the use of a weighted average of writing features to score the vast majority of essays is adequate, it is likely that a more rule-based approach should be employed to identify discrepant essays.

References

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

Breland, H. M., Jones, R. J., & Jenkins, L. (1994). *The College Board vocabulary study* (College Board Rep. No. 94-4; ETS RR-94-26). New York: College Entrance Examination Board.

Burstein, J., Chodorow, M., & Leacock, C. (2003, August). Criterion^{SM}: Online essay evaluation: An application for automated evaluation of student essays. In J. Riedl & R. Hill (Eds.), *Proceedings of the Fifteenth Annual Conference on Innovative Applications of Artificial Intelligence,* Acapulco, Mexico (pp. 3-10). Menlo Park, CA: AAAI Press.

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.

Haberman, S. (2004). *Statistical and measurement properties of features used in essay assessment* (ETS RR-04-21). Princeton, NJ: ETS.

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

Swets, J. A. (1996). Signal detection theory and ROC analysis in psychology and diagnostics: collected papers. Mahwah, NJ: Lawrence Erlbaum Associates.