Automated Essay Scoring With E-rater® V.2.0

Yigal Attali Jill Burstein

Paper to be presented at the Conference of the International Association for Educational Assessment (IAEA)
held between
June 13 to 18, 2004, in Philadelphia, PA.

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 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 the new version.

Key words: Automated essay scoring, e-rater, *Criterion*℠

E-rater has been used by Educational Testing Service (ETS) for automated essay scoring since February 1999. Burstein, Chodorow and Leacock (2003) describes the operational system put in use for scoring the Graduate Management Admissions Test Analytical Writing Assessment (GMAT AWA), and for essays submitted to ETS's writing instruction application, *Criterion*^{SM} Online Essay Evaluation Service. We will refer to the operational system as e-rater Version 1.3. 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, 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 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 eight 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 non-monotonic relationship with the human score. Features in e-rater V.2.0 were created by standardizing some features with regard to essay length, altering the definition of others to take into account the non-monotonic relationship with the human score, and also creating new features. Below is a description of this new feature set.

Errors in Grammar, Usage, Mechanics, and Style

Feedback about a total of 33 errors in grammar, usage, and mechanics, and comments about style are output from *Criterion*. Most of these features are identified using natural language processing (Burstein, Chodorow, & Leacock, 2003). These counts formed the basis of four features in e-rater V.2.0. These are the rates of errors in the four categories and are calculated by counting the total number of errors in each category and dividing this by the total number of words in an essay.

Organization and Development

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 methods (Burstein, Marcu, & Knight, 2003). Two features were derived from this feedback information.

An overall development score 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. An exception to these counts are made in the case of supporting ideas elements that are counted only when they immediately follow a main point element, and main point elements that 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 (a pair of main point and supporting idea paragraphs), and a concluding paragraph.

In e-rater V.2.0 the development score is defined as 8 minus the above sum. This development score may be interpreted as the difference or discrepancy between the 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.

Lexical Complexity

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. For example, in "This essay is a long, long, long essay." there are 5 word types (*this*, *essay*, is, *a*, and *long*), and 8 tokens (*this, essay, is, a, long, long, long,* and *essay*). So the type/token ratio is 6/8, or 0.75. 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. Each word in the essay is assigned a vocabulary level value based on Breland's Standardized Frequency Index 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.

Prompt-Specific Vocabulary Usage

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. 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 essay and, in addition, a set of training essays from each score point, is 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 the six 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 in 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. 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 (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

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 & Scoring

In e-rater V.1.3 models have always been prompt-specific. That is, models were built specifically for each topic, using data from essays written to a particular topic 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 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 standardized 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 standardization is the

reduced feature set. Because the number of features is small and each one of them significantly contributes to the goal of predicting 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 we know what features will be in a model, it is possible to specify the weight of some or all 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 discussion below 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 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 k standardized weights for these features (labeled s_1 - s_k).
4. The fixed standardized weights of the last $n - k$ features should be expressed as percentages of the sum of standardized weights for all features (labeled p_{k+1} - p_n). For example, if there are two fixed weights in a set of 12 features then p_{11} and p_{12} could be .1 and .2, respectively, which means that s_{11} will be equal to 10% of the sum of s_1 - s_{12}, s_{12} will be equal to 20% of s_1 - s_{12}, and the sum of s_1 - s_{10} will account for the remaining 70% of the standardized weights.
5. Find the fixed standardized weights by applying the following formula to the last $n - k$ features:

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

6. To find the un-standardized weights (labeled w_1 - w_n), multiply s_i by the ratio of the standard deviation for human score to standard deviation for the feature.
7. Compute an interim predicted score as the sum of the product of feature values and weights w_1 - w_n.
8. 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.
9. The final un-standardized weights are given by multiplying a by w_i ($1 \le i \le n$).

The combined use of a small feature set and more standardized modeling procedures may have a beneficial effect on the interpretability, reliability, and validity of automated scores.

Generic Model Building

The combination of predetermined and optimal weights in regression modeling can be applied to both prompt-specific and generic modeling. 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 for each topic a sample of scores from human-rated essays on this 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 are used for model building and the remaining set is used for cross-validating the model. In addition, such prompt-specific models will obviously be different for different prompts of the same program or grade. The relative weights of features will vary between prompts and even the sign of weights might be reversed for some prompts. This 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 useful. A single regression model that is used with all

prompts from one program or grade 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 than 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 are treated in the same way in modeling. First, prompts may have different difficulty levels, since only the *interpretation* of levels of performance for individual features are the same across prompts, not the levels of performance. Second, e-rater generic models can still incorporate prompt-specific vocabulary usage information through the two vocabulary-based features. Again, it is important to distinguish between the *computation* of these two features, which must be based on 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 simulating the human score. We have found that different systems and datasets have different "natural" sets of cutoffs. The factors that influence the values of the best cutoffs are, among others, the distributions of the features used in modeling and the distribution of human scores.

To find a suitable set of cutoffs for a system, a search through a large set of possible cutoff sets is performed to find the set that maximizes overall exact agreement and minimum exact agreement across all score points. These two criteria are weighted to produce ratings of sets of cutoffs. This search process is performed "generically" on pooled data across a number of prompts to produce an appropriate set of cutoffs for a program.

The following is an example of how the cutoffs are determined. Each set of cutoffs is assigned a score, which is computed as 80% of the overall exact agreement for that set of cutoffs and 20% of the minimum exact agreement across all score points. The set of cutoffs with the highest score is selected for use. For example, given a specific set of cutoffs with the exact agreement values: .30, .35, .40, .50, .45, and .40 for the six score points 1 to 6, respectively, and an overall exact agreement value of .40, the score assigned to this set would be 20% of .30 (the minimum exact agreement, achieved for score 1) plus 80% of .40 (the overall exact agreement), or .38.

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 sixth through twelfth grade *Criterion* user data, and GMAT and TOEFL® (Test of English as Foreign Language) human-scored essay data. The sixth through twelfth 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.

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 there were 64 different prompts and almost 18,000 essays analyzed.

Table 1

Descriptive Statistics on Essays and Average Human Score (AHS)

Program	Prompts	Mean # 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 average AHS for most programs is around 3.5, except for somewhat lower scores for sixth and seventh grade and higher scores for eleventh grade. The standard deviations 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 (tenth feature in Table 2) with lower correlations in lower grades.

Table 2

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

Feature	6th	7th	8th	9th	10th	11th	12th	GMAT argument	GMAT 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
EL	0.74	0.82	0.79	0.81	0.79	0.74	0.83	0.71	0.79	0.82

Table 3 presents the average feature values for each AHS from 1.0 to 6.0. The average 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 3

Average Feature Values (Relative to the Average of AHS of 1.0) per AHS, and Overall Mean and Std

	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	.19	.15	0.0863	0.0815
Development	1.00	.87	.76	.62	.53	.41	.33	.24	.21	.17	.15	-3.0	2.4
AEL*	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*	1.00	1.00	.99	.98	.97	.96	.95	.94	.94	.94	.92	4.5	0.5
Max. Cos.*	1.00	1.00	.95	.91	.88	.83	.82	.78	.75	.73	.71	4.1	1.2
Cos. w/6*	1.00	.76	.65	.60	.54	.52	.49	.48	.45	.44	.42	0.1124	0.0525
EL*	1.00	.60	.46	.37	.33	.27	.24	.20	.18	.16	.14	263.7	128.9

* Scale of feature reversed by multiplying values by -1.

To give a sense of the relative importance of the different features in the regression models Table 4 presents the relative standardized weights of the first 11 features when a regression analysis for prediction of AHS was for each program separately. The Table shows similar weights across programs with no significant developmental trends 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 4

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

Feature	6th	7th	8th	9th	10th	11th	12th	GMAT arg.	GMAT 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
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

GMAT Model Building Results

In this section, we will present 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 above that 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 which 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.

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 is composed of an independent sample of essay responses that have not be used for model building. Performance on the cross-validation set tells us what kind of system performance we can expect in the field. For prompt-specific models (both V.1.3 and V2.0) 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 was 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. As was discussed above, because of its large correlation with human score the effect of running a free regression model with essay length as one of the features is a large weight for 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 percents of total standardized weights for all features in model. One can see that in the case of the argument prompts there is a significant increase in Kappas from .0 to .1 weight and a smaller increase up to a weight of .3-.4. For the issue prompts, we find a significant increase from .0 to .1 and a smaller increase from .1 to .2. In the case of the argument prompts we see a decrease in performance when weight is raised to .5 from .4. The Table also shows very similar results between the Generic12 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	Program	0.0	0.1	0.2	0.3	0.4	0.5
Generic10[1]	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[2]	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

[1] Generic model without vocabulary usage features—10 features

[2] Generic model with vocabulary usage features—12 features

Table 6 presents the results for comparing the three V.2.0 models with the V.1.3 models. The Table shows that the prompt-specific and Generic12 models outperformed V.1.3 models and that in the case of the issue prompts even the Generic10 models performed better than the V.1.3 models.

Table 6

Kappas and Rates of Exact Agreement for Different Systems

System	ELW[1]	N	Mean Kappa	STD Kappa	Exact agreement
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
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

1. Essay Length Weight.

Reliability of E-rater V.2.0

Evaluations of automated essay scoring systems are usually based on single-essay scores and on a comparison between the relation of two human rater scores and the relation of machine-human scores. 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 to 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, similarly 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 would be perfect, simply because no measure can have a stronger relation to a human score than that of two human scores have between each other. Finally, this analysis takes into account only one kind of inconsistency between human scores, inter-rater inconsistencies within one essay, and not the 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* essays that were analyzed in previous sections. These essays were chosen from the *Criterion* database to include as many multiple essays per student as possible. Consequently it was possible to extract, out of the 7,575 essays from sixth and twelfth grade students in the sample, almost 2,000 students who submitted at least two different essays. These essays 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 over-fitting as much as possible. Note that the weights chosen are not only sub-optimal on the prompt level, but 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 six out of seven grades) and equivalent reliabilities to the average of two human raters, with overall reliability of .60, higher than that of the AHS.

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	1987	.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 multi-method analysis. Table 8 presents a typical multi-method correlation table. The two correlations above the main diagonal are equal to the average of the correlations between the first e-rater and second human score (either single or average of two), and between the second e-rater and first human score (both pairs of correlations were almost identical). The correlations below 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

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

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

The main finding presented in Table 8 is the high corrected-correlation (or true-score correlation) between human and machine scores—.93. This high correlation is evidence that e-rater scores, as an alternative method for measuring writing ability, is 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). In an unpublished statistical report of the New SAT field trial (Liu & Feigenbaum, 2003) the estimate of the test-retest reliability (alternate-form Pearson correlation) of a single essay test, the SAT II: Writing test was .59 (Breland et al., 2004); the KR20 reliability estimate for the multiple-choice New SAT Writing test was .86; and the raw correlation between the multiple-choice and essay writing tests was .49. Finally, the estimate for the true-score correlation between the essay and multiple-choice test was .69, much lower than the .93 estimate for the true-score correlation between human and e-rater scoring of essays.

Table 9 shows the results from another interesting analysis that is made possible with the multiple-essay data, 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

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
EL	0.56	0.20	0.78

The table above shows that the essay length feature has the highest reliability (.56), higher then 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 average word length (AWL) features are in the 40s; the reliabilities of the mechanics, average element length (AEL), and the 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.

The comparison between the three columns of Table 9 show 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.

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 these advantages 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 sixth- to twelfth-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 these discrepant essays.

References

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., Kubota, M., Nickerson, K., Trapani, C., & Walker, M. (2004). *New SAT writing prompt study: Analyses of group impact and reliability* (College Board Report No. 04-1). New York: College Board.

Breland, Jones, & Jenkins, 1994

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., Chodorow, M., & Leacock, C. (2003). *Criterion*^{SM}: Online essay evaluation: An application for automated evaluation of student essays. In *Proceedings of the Fifteenth Annual Conference on Innovative Applications of Artificial Intelligence*.

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-??). Princeton, NJ: Educational Testing Service.

Liu, J., & Feigenbaum, M. (2003). *Prototype analyses of Spring 2003 New SAT field trial* (unpublished statistical report). Princeton, NJ: Educational Testing Service.

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