Measuring the Power of Learning.™

Research Report
ETS RR–16-04

Building *e-rater*® Scoring Models Using Machine Learning Methods

Jing Chen

James H. Fife

Isaac I. Bejar

André A. Rupp

February 2016

ETS Research Report Series

Since its 1947 founding, ETS has conducted and disseminated scientific research to support its products and services, and to advance the measurement and education fields. In keeping with these goals, ETS is committed to making its research freely available to the professional community and to the general public. Published accounts of ETS research, including papers in the ETS Research Report series, undergo a formal peer-review process by ETS staff to ensure that they meet established scientific and professional standards. All such ETS-conducted peer reviews are in addition to any reviews that outside organizations may provide as part of their own publication processes. Peer review notwithstanding, the positions expressed in the ETS Research Report series and other published accounts of ETS research are those of the authors and not necessarily those of the Officers and Trustees of Educational Testing Service.

The Daniel Eignor Editorship is named in honor of Dr. Daniel R. Eignor, who from 2001 until 2011 served the Research and Development division as Editor for the ETS Research Report series. The Eignor Editorship has been created to recognize the pivotal leadership role that Dr. Eignor played in the research publication process at ETS.

ETS Research Report Series ISSN 2330-8516

RESEARCH REPORT

Building *e-rater*® Scoring Models Using Machine Learning Methods

Jing Chen, James H. Fife, Isaac I. Bejar, & André A. Rupp

Educational Testing Service, Princeton, NJ

The *e-rater*® automated scoring engine used at Educational Testing Service (ETS) scores the writing quality of essays. In the current practice, e-rater scores are generated via a multiple linear regression (MLR) model as a linear combination of various features evaluated for each essay and human scores as the outcome variable. This study evaluates alternative scoring models based on several additional machine learning algorithms, including support vector machines (SVM), random forests (RF), and *k*-nearest neighbor regression (*k*-NN). The results suggest that models based on the SVM algorithm outperform MLR models in predicting human scores. Specifically, SVM-based models yielded the highest agreement between human and e-rater scores. Furthermore, compared with MLR, SVM-based models improved the agreement between human and e-rater scores at the ends of the score scale. In addition, the high correlation between SVM-based e-rater scores with external measures such as examinee's scores on the other parts of the test provided some validity evidence for SVM-based e-rater scores. Future research is encouraged to explore the generalizability of these findings.

Keywords machine learning; automated essay scoring; multiple linear regression; support vector machines; statistical modeling

doi:10.1002/ets2.12094

Automated scoring is widely used to score constructed-response items in educational tests. Systems such as the *e-rater*® automated scoring system developed by Educational Testing Service (ETS; Burstein, Tetreault, & Madnani, 2013) score essays automatically. Specifically, the scores generated by e-rater are predicted human scores where the prediction equation is obtained by a multiple linear regression (MLR) model, with main effects only, that regresses human scores on a set of computer generated features of the essays (for a deeper discussion of e-rater, see Attali & Burstein, 2006, or Quinlan, Higgins, & Wolff, 2009).

The current MLR model used to generate e-rater scores considers only the first order of the feature variables, and no interaction terms of feature combinations are included. Although this model is simple, intuitive, and easily interpretable, there is no a priori reason to assume a strictly linear relationship between the features and the scores. The nonlinear part of the relationship between the response features and the overall score, if any, will not be captured by the MLR models in the current settings where no polynomial terms for individual features and no interaction terms for feature combinations are used. Furthermore, previous studies suggest that under the current MLR model, the discrepancy between e-rater scores and human scores is larger at the extreme ends of the score scale than in the middle (Burstein, Kukich, Wolff, Lu, & Chodorow, 1998; J. Chen & Ramineni, 2012).

Algorithms such as support vector machines (SVM), random forests (RF) and *k*-nearest neighbor regression (*k*-NN) provide a wide range of mappings between dependent variables and independent variables and are good candidates for capturing more complex relationships between the human scores and the essay features. In this study, we applied SVM, RF, and *k*-NN to build e-rater models to see whether they can outperform MLR models in predicting human scores.

The results of this study have two practical implications. First, if applying more sophisticated machine learning models to the current features can significantly improve e-rater performance, we can build a better empirical representation of the relationship between the feature values and the human scores. Second, given the large number of ETS test programs, such as the *TOEFL*® products and the *GRE*® general test, that currently use e-rater and the associated high volumes of responses across programs, even a small improvement in the agreement between human scores and automated scores will save substantial scoring time and expenses associated with using a second human rater.

Corresponding author: J. Chen, E-mail: jchen003@ets.org

This paper is divided into five sections. In the first section, we review studies that apply machine learning in automated scoring and studies that compare the performance of scoring models built by different machine learning algorithms. Then, we briefly introduce the e-rater scoring engine in the second section and introduce our research methods in the third section. In the fourth section, we present the results, and in the final section, we discuss findings and implications of this study.

Literature Review

Samuel (1959) defined machine learning as a "field of study that gives computers the ability to learn without being explicitly programmed" (p. 89). The key feature of machine learning is to predict an unknown quantity based on knowledge of existing data (or training data). In this study, we compare four machine learning models: SVM (Cristianini & Shawe-Taylor, 2000; Vapnik, 1995), RF (Breiman, 2001), *k*-NN (Skellam, 1952), and MLR.

Machine learning methods have been applied to the automated scoring of various types of tasks. For example, Nehm, Ha, and Mayfield (2012) explored the use of machine learning to automatically evaluate the accuracy of students' written explanations of evolutionary change. Their scoring program was found to be a powerful and cost-effective tool for assessing student knowledge and performance in a complex science domain. An automated scoring system was developed at ETS in 2012 to score written short responses (Heilman & Madnani, 2013). It uses a machine learning algorithm to create scoring models to map test takers' responses to scores. Yannakoudakis, Briscoe, and Medlock (2011) applied machine learning techniques to score English as a second or other language (ESOL) examination scripts. Some other studies (M. Chen & Zechner, 2011; Zechner & Bejar, 2006) used machine learning approaches to investigate the feasibility of the automated scoring of spoken English proficiency of non-native speakers.

Studies have been conducted to compare the performance of different machine learning methods in automated scoring. H. Chen, He, Luo, and Li (2012) applied a machine learning method called ranking SVM[1] (Joachims, 2006) to automated essay scoring and found that ranking SVM outperforms *k*-NN and MLR in the automated scoring of essays. Santos, Verspoor, and Nerbonne (2012) found that among 11 machine learning algorithms, including RF, naïve Bayes, and simple cart, logistic model trees (LMT), which use logistic regression, achieved the best classification accuracy in terms of predicting the English proficiency level of the essays. Zechner and Bejar's (2006) study applied both SVM and classification and regression trees (CART; Breiman, Friedman, Olshen, & Stone, 1984) to score spoken responses. They used SVM models for score prediction and CART models for uncovering the role of different features and feature classes in classifying spoken responses. The results suggested that scoring based on SVM yields machine–human agreement that approaches human–human agreement in some cases. M. Chen and Zechner (2011) experimented with two algorithms, MLR and classification tree, to build scoring models for automatic scoring of speech responses. The results showed that MLR models consistently outperformed decision tree models.

The studies reviewed in the preceding paragraphs compared the performance of different machine learning methods in automated scoring of various types of tasks. This study compares the performance of four machine learning methods in building e-rater scoring models to predict human raters' scores. Previous studies also suggested different machine learning methods that show strengths in different applications. The performance of machine learning methods varies significantly across different problems, different evaluation metrics, and different datasets (Caruana & Niculescu-Mizil, 2006). Thus, in this study, we compare four machine learning algorithms to discover the most effective algorithm to build scoring models for our particular datasets.

The research questions of this study are the following:

- Can e-rater models calibrated by SVM, RF, or *k*-NN methods outperform MLR-based e-rater models in predicting human raters' scores?
- Which machine learning algorithm predicts the human raters' scores the best?

Automated Scoring With E-rater

E-rater has been used by ETS for automated essay scoring since 1999. It was initially developed (Burstein et al., 1998) to score GMAT essays automatically. Since then the engine has been periodically updated and enhanced with newer versions. This study employed e-rater version 12.1, which was used in the operational scoring from June 2012 to July 2013.

Table 1 Features of E-rater Engine Version 12.1

Feature	Descriptions
1. Grammar	A statistic based on the number of errors related to pronouns, run-ons, missing possessives, etc.
2. Mechanics	A statistic based on the number of errors related to capitalization, punctuation, commas, hyphens, etc.
3. Style	A statistic based on the number of errors related to repetition of words, inappropriate words, etc.
4. Usage	A statistic based on the number of errors related to in missing/wrong articles, nonstandard verbs, etc.
5. Organization	Logarithm of the number of discourse units
6. Development	Logarithm of the mean length of discourse units
7. Word choice	Word frequency measured by the Standard Frequency Index (SFI; Breland, 1996)
8. Word length	Average word length in an essay
9. Positive features	Correct use of collocations and prepositions, and sentence variety feature
10. Differential word use	The frequency of individual word use in an essay based on the word's relative use in high and/or low scoring essay responses

E-rater version 12.1 uses 10 response features to evaluate the quality of each essay (see Table 1 for the 10 features used in version 12.1). Most of these 10 features are composed of a set of subfeatures computed using natural language processing (NLP) techniques. In addition, some of the subfeatures are composed of a set of microfeatures that are combined to produce the subfeature values.

A set of advisory flags has been developed to indicate when a specific essay is not suitable for e-rater scoring (for more details on advisory flags see Ramineni, Trapani, Williamson, Davey, & Bridgeman, 2012). Each advisory flag marks a different problem, such as an essay containing excessive repetition of words, an essay not relevant to the assigned topic, or an essay that is too brief to be evaluated. After filtering out the inappropriate essays, e-rater uses the 10 response features in Table 1 to evaluate the quality of the remaining essays.

Usually, the data are randomly split into a training dataset and a validation dataset. Both datasets should be representative of the population for which e-rater is intended for use. The training dataset is used to build scoring models, and the evaluation dataset is used to evaluate the scoring models. As noted earlier, e-rater scoring models are built using an MLR approach, in which the 10 features listed in Table 1 are the independent variables and the human score is the dependent variable. The weights of the features are thus estimated to maximize the overall agreement with human scores on the basis of a least-squares estimation approach and removal of features with negative weights from the model. The final e-rater scores are scaled to match the distributional mean and standard deviation of e-rater scores to those of the human scores.

Two primary variations of the MLR models are built for e-rater essay scoring: generic models and prompt-specific models. Generic models are built using the pooled data from a group of related prompts and do not contain prompt-specific content features. These models thus result in a single set of regression intercept and feature weights that are suitable for all prompts that are designed to the same general task design specifications. The prompt-specific models are custom-built for each prompt using data from that prompt and include two content features measuring prompt-specific vocabulary usage. These models thus contain regression intercepts and feature weights that are specific to each prompt. A majority of the current ETS test programs that use e-rater for operational scoring use generic models because there is no substantial model performance variation across prompts and generic models are easier to implement. In this study, we built generic models across all statistical modeling approaches we utilized.

E-rater scores produced from MLR models are real-number scores, unlike human scores, which are restricted to integer values. For comparison with human scores, e-rater scores are truncated into the range of the score scale and then rounded to the nearest integer, using normal rounding rules. Consequently, there are three types of e-rater scores: (a) unbounded/raw e-rater scores (erater_raw), (b) bounded but unrounded e-rater scores (erater_bound), and (c) rounded e-rater scores (erater_round). Unbounded e-rater scores (erater_raw) can theoretically range from $-\infty$ to $+\infty$ and are used during human scoring for adjudication purposes.[2] Bounded but unrounded scores (erater_bound) are truncated e-rater scores with a score scale that matches the score scale of human ratings. Finally, rounded e-rater scores (erater_round) are computed by rounding the bounded score to the nearest integer value to align with the measurement scale of the human scores. We used both erater_bound and erater_round for different statistical analyses depending on the required score scale of the associated statistics.

Furthermore, e-rater scores are mainly used in two ways at ETS in operational scoring: as contributory scores or as confirmatory scores. In this study, e-rater scores were used as contributory scores for one test assessment and confirmatory

scores for the other test assessment from which we collected data. When e-rater scores are used as confirmatory scores, they are only used as machine-based validations of the human scores; additional human ratings are obtained if the e-rater scores are discrepant from the human score by a certain amount. Under this model, the final score for the examinee comes from human ratings only. When e-rater scores are used as contributory scores, they are used in conjunction with one human rating in determining the final score for a writing task. That is, a weighted combination of the e-rater score and the human score yields the final score.

Methods

Datasets

The data for this study came from the writing tasks of two large-scale college level assessments. The test takers in the first assessment (Assessment I) formed a mix of native and non-native English speakers, and the test takers in the second assessment (Assessment II) were all non-native English speakers. The essays from Assessment I were scored on a 6-point holistic scale, and the essays from Assessment II were scored on a 5-point holistic scale. These scales reflected the overall quality of an essay in response to the assigned task. E-rater scores were used as confirmatory scores for the essay scoring of Assessment I and were used as contributory scores for the essay scoring of Assessment II. The writing tasks of Assessment I included two task types: Task A and Task B. Task A required examinees to critique an argument. Task B required examinees to articulate an opinion and support their opinions by using examples or relevant reasoning. Similar to the writing tasks of Assessment I, the writing tasks of Assessment II also included two task types: Task C and Task D. Task C required test takers to read, listen, and then respond in writing by synthesizing the information that they had read with the information they had heard. Task D required test takers to articulate and support an opinion on a topic. Examinees of the first assessment were given 30 minutes to complete each of the writing tasks; examinees of the second assessment were given 25 minutes to complete each of their writing tasks.

The data we used came from both writing tasks of the two assessments. They were used to build and validate scoring models of e-rater engine 12.1 for these two assessments. For each test, essays from a certain period of administration time were chosen as a representative sample to build and validate e-rater scoring models. We used these model building and validation datasets. Responses with so-called fatal advisory flags, which indicate that the responses were unsuitable for automated scoring, were excluded from our analysis. The proportion of the responses with fatal flags was less than 5% of the total responses. In operational scoring, the MLR model was trained using a randomly selected training dataset and validated on a randomly selected validation dataset. The sample sizes of the training and the validation datasets are listed in Table 2. To compare with the results from the MLR model used in operational scoring, we used the same training and validation datasets to train and validate all the models. The variables included in each dataset were human raters' scores, feature scores, and examinees' scores on the other sections of the tests.

Machine Learning Methods

As discussed in the literature review section, each machine learning method may work well for some cases but not for others. By comparing the results from different methods, we can identify the method that works best for our specific datasets. As noted above, in addition to MLR, in this study we used SVM, RF, and *k*-NN.

Each of these three algorithms can be applied to solve both classification and regression problems. The difference between classification and regression problems is the property of the output variable that a model is built to predict. The output variable of a classification problem is a categorical variable, whereas the output variable of a regression problem is a continuous variable. As the current output from the MLR model is a continuous variable (i.e., erater_raw), we use the regression methods of each of the machine learning algorithms in our analysis so that we can directly compare the results

Table 2 Total Sample Size and Sample Sizes of Model Building and Validation Datasets

	Assessment I Task A	Assessment I Task B	Assessment II Task C	Assessment II Task D
Total sample size	71,441	74,295	94,106	93,786
Model building (*N*)	14,398 (20.2%)	14,997(20.2%)	47,053 (50.0%)	46,884 (50.0%)
Validation (*N*)	57,043 (79.8%)	59,298(79.8%)	47,053 (50.0%)	46,902(50.0%)

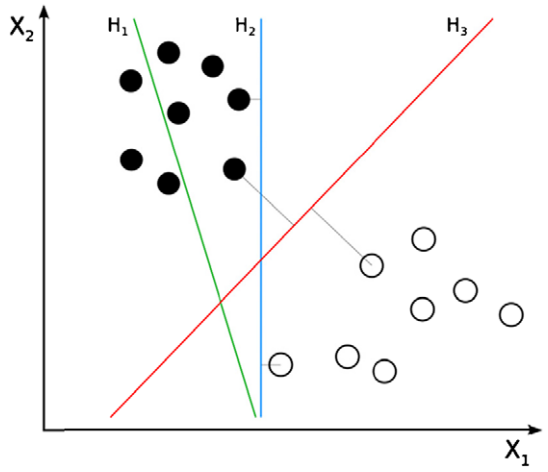

Figure 1 A graphical illustration of support vector machine method.

with the MLR-based output. Each method produces continuous e-rater scores that are then discretized into categorical scores in the same way as in the MLR model.

We conducted our analysis in the statistical programming language R. The packages for SVM, RF, and *k*-NNs are e1071, randomForest, and kknn, respectively. Detailed descriptions of the algorithms are beyond the scope of this paper. In the following, we briefly introduce the basic ideas of each algorithm and provide the relevant literature for interested readers to check the details.

1. SVM: Each response can be considered as a point in a high-dimensional space with each feature as one dimension. Based on the training dataset, the SVM algorithm uses decision surfaces (see H_1, H_2, H_3 in Figure 1) in the space so as to separate the responses optimally into different score categories. The decision surfaces are chosen to maximize the average distance between the decision surface and the responses belonging to different score categories. For example, in Figure 1, the decision surface H_3 is chosen because it maximizes the distance between the surface and the dots belonging to different categories. This is the basic idea of the SVM-based classification algorithm. In regression analysis, things are similar except that one is finding a function f(x) (here x represents the feature variables) that has at most ϵ deviation from the target labels for all training data, and at the same time is as flat as possible (Vapnik, 1995). More details of the algorithm can be found in (Smola & Schölkopf, 2004).
2. RF: An RF classifier uses a number of decision trees to predict the value of a dependent variable based on the independent variables. In the tree structures, nodes are the splitting criteria for independent variables, leaves represent class labels to which the dependent variable values are assigned, and branches represent conjunctions of features that lead to those class labels. This process is repeated in a recursive manner until the subset at a node has all the same values of the predicted variable or when splitting no longer adds value to the predictions. Based on the training dataset, each tree is learned by splitting the data into subsets based on the input features. After training, each decision tree will make a prediction of the output variable and the final prediction is based on the votes across all the trees. The regression version of the RF differs from the classification version at the tree level. For each tree, the feature space is portioned into disjoint regions and a fitting function is used in each region. For more details, readers can refer to Breiman et al. (1984).
3. *k*-NN: As with SVM, each essay response in the training set can be considered as a point in a high-dimensional Euclidean space, with each feature as one dimension. By viewing an unscored response as a point in this space also, the score for this response can be predicted using the scores of the *k* training essays nearest to the unscored essay in this space (using the Euclidean distance function or some other metric). For *k*-NN regression, the score for an unscored response is predicted by the average of the scores of the *k* training essays nearest to the unscored response. Alternatively, the contribution to the prediction from each neighbor can be weighted based on the distance between the neighbor and the unscored essay. For example, a common weighting scheme consists in giving each neighbor a weight of $1/d$, where d is the distance to the neighbor.

It is worth noting that each of these three machine learning algorithms has some hyperparameters that can be tuned to improve the performance. For example, SVM has four hyperparameters that can be adjusted. These hyperparameters are kernel type, degree (when using polynomial kernel), gamma, and cost. The default hyperparameters are generally considered as a starting point and one needs to fine-tune them to find out the best hyperparameters for a particular problem.

In our analysis, we started with these default hyperparameters and tuned them through a grid search to optimize the agreement statistics. To perform a grid search, some trial values for each hyperparameter were selected. For example, for the SVM method, two common kernel types, polynomial kernel and radial basis function, and some trial values of gamma and cost were selected. The Cartesian product of these sets of values formed the tuning grid of hyperparameter values. Using the model building dataset, we tested all the possible combinations in the grid to find a combination of hyperparameters that optimized the agreement statistics. The model, with these optimized hyperparameters, was evaluated using the validation data.

Data Analysis

E-rater scores generated from the machine learning models were compared with e-rater scores[3] generated from the MLR models to see whether machine learning models outperform MLR models in predicting human scores. The performance of e-rater models was evaluated according to several quality-control/agreement statistics that ETS uses to gauge alignment between human raters and automated scoring engines (Williamson, Xi, & Breyer, 2012): quadratic-weighted kappa (QWK), percentage of exact agreement (% exact agreement), standardized mean difference (SMD) between human and rounded e-rater scores, and Pearson correlation (r) between human and unbounded/raw e-rater scores. In addition, we analyzed whether automated scores generated from machine learning methods were related to scores on other sections of the test in similar ways when compared to human scores. All these statistics were calculated using the validation datasets.

Typically, the QWK between rounded e-rater scores and human scores must be at least 0.70 and the Pearson correlation between unrounded e-rater scores and human scores must be at least 0.70 as well. This 0.70 value was selected on the basis that it nearly reaches the tipping point at which signal outweighs noise in the prediction (the true such point is 0.7071, the square root of 0.50) and so nearly half the variance is accounted for in the prediction (Ramineni & Williamson, 2013). Similarly, the absolute value of the SMD is not recommended to exceed 0.15. This standard ensures that the distribution of the e-rater scores is centered on a point that is close to the center of the human score distribution. To evaluate the fairness of automated scores for different subgroups of examinees, a more stringent flagging criterion of standardized mean score differences is set at 0.10 for each subgroup. This flagging criterion is applied to all subgroups of examinees to identify patterns of systematic differences in the distribution of scores between human scoring and automated scoring for subgroups (see Ramineni & Williamson, 2013, and Williamson et al., 2012, for a more detailed discussion on guidelines).

Some studies (Bridgeman, Trapani, & Attali, 2012; J. Chen & Ramineni, 2012) revealed unacceptable SMDs between machines and human scores for some demographic subgroups. For example, Bridgeman et al. (2012) found Chinese examinees received significantly higher scores from e-rater than from human raters in GRE writing. Therefore, in this study, we also compared the fairness of the automated scores generated from different models for different subgroups of examinees.

Previous studies investigated not only the consistency between automated scores and human scores, but also the relationship of automated scores with external criteria, both in an absolute sense and relative to the relationship based on human scores (Attali, Bridgeman, & Trapani, 2010; Petersen, 1997; Powers, Burstein, Chodorow, Fowles, & Kukich, 2002; Weigle, 2010). The assumption is that if human and automated scores reflect similar constructs, they are expected to relate to other measures of constructs in similar ways and should thus show similar correlation patterns.

Results

In Tables 3–6, we present the results from MLR, SVM, RF, and k-NN models in terms of the four evaluation metrics discussed in the previous section: QWK, % exact agreement, SMD between human and rounded e-rater scores, and r between human scores and bounded e-rater scores. These tables present the results from each of the four writing tasks. In each table, the word *default* after an algorithm name means that the default hyperparameters for that algorithm were used, whereas *after tuning hyperparameters* means that the hyperparameters were used that yielded the best results after many trials. In each table, the smallest SMD and the highest QWK, % exact agreement, and r are bolded.

Table 3 Agreement Statistics for Task A, Assessment I ($N = 57{,}043$)

Method	SMD	QWK	% exact agreement	*r*
MLR	−0.029	0.683	59.900	0.751
SVM (default)	−0.038	0.705	61.874	0.765
SVM (after tuning hyperparameter)	−0.044	**0.713**	**62.224**	**0.771**
RF (default)	**−0.014**	0.681	60.833	0.760
RF (after tuning hyperparameter)	**−0.014**	0.688	60.989	0.761
k-NN (default)	−0.073	0.598	53.395	0.652
k-NN (after tuning hyperparameter)	−0.094	0.611	56.796	0.716

Note. MLR = multiple linear regression; SVM = support vector machine; RF = random forest; *k*-NN = *k*-nearest neighbor regression; SMD = standardized mean difference; QWK = quadratic-weighted kappa. *Default* means that the default hyperparameters for that algorithm were used; *after tuning hyperparameters* means that the hyperparameters were used that yielded the best results after many trials. The smallest SMD and the highest QWK, % exact agreement, and *r* are bolded.

Table 4 Agreement Statistics for Task B, Assessment I ($N = 59{,}298$)

Method	SMD	QWK	% exact agreement	*r*
MLR	−0.029	0.740	67.124	0.803
SVM (default)	−0.014	0.757	68.621	0.817
SVM (after tuning hyperparameter)	−0.021	**0.768**	**69.331**	**0.823**
RF (default)	**−0.005**	0.743	67.955	0.813
RF (after tuning hyperparameter)	−0.007	0.754	68.441	0.816
k-NN (default)	−0.071	0.670	60.569	0.722
k-NN (after tuning hyperparameter)	−0.105	0.674	63.159	0.774

Note. MLR = multiple linear regression; SVM = support vector machine; RF = random forest; *k*-NN = *k*-nearest neighbor regression; SMD = standardized mean difference; QWK = quadratic-weighted kappa. *Default* means that the default hyperparameters for that algorithm were used; *after tuning hyperparameters* means that the hyperparameters were used that yielded the best results after many trials. The smallest SMD and the highest QWK, % exact agreement, and *r* are bolded.

Table 5 Agreement Statistics for Task C, Assessment II ($N = 47{,}053$)

Method	SMD	QWK	% exact agreement	*r*
MLR	**0.000**	0.649	61.599	0.728
SVM (default)	0.043	0.633	61.837	0.726
SVM (after tuning hyperparameter)	0.020	**0.650**	**62.479**	**0.737**
RF (default)	0.013	0.619	60.614	0.717
RF (after tuning hyperparameter)	0.013	0.625	60.675	0.717
k-NN (default)	−0.026	0.533	53.938	0.593
k-NN (after tuning hyperparameter)	−0.104	0.550	56.475	0.659

Note. MLR = multiple linear regression; SVM = support vector machine; RF = random forest; *k*-NN = *k*-nearest neighbor regression; SMD = standardized mean difference; QWK = quadratic-weighted kappa. *Default* means that the default hyperparameters for that algorithm were used; *after tuning hyperparameters* means that the hyperparameters were used that yielded the best results after many trials. The smallest SMD and the highest QWK, % exact agreement, and *r* are bolded.

The results show that SVM gives the best results based on most of the evaluation metrics across all four writing tasks because the SVM-based models with tuned hyperparameters have the highest QWK, percent exact agreement, and *r* across all four datasets and the SMD is less than 0.15 across all models. Radial basis function was used as the kernel type. The results of all four evaluation metrics from the RF-based models with tuned hyperparameters are better than those from MLR models for essays of Assessment I but not for essays of Assessment II. According to all four evaluation metrics, *k*-NN performs worst among the three machine learning algorithms and does not predict human scores as well as MLR models across all four datasets.

We note that the values of the agreement statistics of the SVM models are only slightly larger than those of the MLR model. However, given the large test volumes, these slight increases in agreement of e-rater scores with human scores can

Table 6 Agreement Statistics for Task D, Assessment II ($N = 46{,}902$)

Method	SMD	QWK	% exact agreement	*r*
MLR	−0.003	0.526	41.296	0.620
SVM (default)	−0.020	0.554	41.804	0.627
SVM (after tuning hyperparameter)	−0.025	**0.556**	**41.906**	**0.634**
RF (default)	−0.005	0.513	40.476	0.617
RF (after tuning hyperparameter)	**−0.002**	0.519	40.612	0.613
k-NN (default)	−0.033	0.442	36.212	0.484
k-NN (after turning hyperparameter)	−0.048	0.473	39.509	0.574

Note. MLR = multiple linear regression; SVM = support vector machine; RF = random forest; *k*-NN = *k*-nearest neighbor regression; SMD = standardized mean difference; QWK = quadratic-weighted kappa. *Default* means that the default hyperparameters for that algorithm were used; *after tuning hyperparameters* means that the hyperparameters were used that yielded the best results after many trials. The smallest SMD and the highest QWK, % exact agreement, and *r* are bolded.

Table 7 Agreement Statistics of Tasks A and B of Assessment I for Subgroups of Examinees

Country of subgroup	Task type	Method	SMD	QWK	% exact agreement	*r*
United States	Task A	MLR	0.028	0.681	60.223	0.755
		SVM	−0.011	0.711	62.020	0.771
	Task B	MLR	0.047	0.753	69.045	0.817
		SVM	0.028	0.777	70.510	0.830
China	Task A	MLR	−0.340	0.468	57.770	0.561
		SVM	−0.249	0.474	62.615	**0.555**
	Task B	MLR	−0.452	0.494	60.022	0.612
		SVM	−0.313	0.511	65.882	**0.609**
India	Task A	MLR	−0.194	0.646	59.567	0.723
		SVM	−0.139	0.668	61.600	0.740
	Task B	MLR	−0.161	0.660	63.245	0.732
		SVM	−0.109	0.675	65.135	0.752
Pakistan	Task A	MLR	−0.135	0.604	55.654	0.703
		SVM	−0.119	0.665	59.717	0.746
	Task B	MLR	−0.052	0.643	63.176	0.742
		SVM	0.036	0.693	66.892	0.763

Note. In this table, the agreement statistics from the SVM models that are worse than those from the MLR models are bolded. SMD = standardized mean difference; QWK = quadratic-weighted kappa; MLR = multiple linear regression; SVM = support vector machine.

result in a significant reduction in the costs of human scoring due to the reduced number of human adjudication scores required. In the discussion section that follows, we provide a rough estimate of the annual reduction in the costs of human scoring that ETS could realize if MLR-based models were replaced with SVM-based models.

Our results suggest that SVM-based scores improve the agreement between humans and e-rater for all major subgroups of examinees as well. The four largest subgroups in the population of Assessment I are American, Chinese, Indian, and Pakistani examinees, which constitute over 90% of the test-taker population. The four largest examinee subgroups for the population of Assessment II are Chinese, Korean, Japanese, and Indian examinees, which comprise more than 95% of the population.

Tables 7 and 8 present the agreement statistics between human scores and automated scores using different models for these particular subgroups for both assessments. The agreement statistics from SVM-based scores with tuned parameters (radial basis function is the kernel type) are better than those from MLR models except for the five cases that are bolded.

It is worth noting that SVM-based scores can reduce the discrepancy between human and automated scores for some particular subgroups of interest. For instance, a previous study (Bridgeman et al., 2012) revealed unacceptable differences between the human and the e-rater scores that Chinese examinees received in writing. We found that by using SVM-based models, the SMD between human and automated scores for this subgroup could be reduced from −0.340 to −0.249 for Task A and from −0.452 to −0.313 for Task B. Although these SMD values still exceed the flagging criterion of 0.10 in absolute value, they suggest that SVM-based scores do alleviate the problem.

Table 8 Agreement Statistics of Tasks C and Task D of Assessment II for Subgroups of Examinees

Country of subgroup	Task type	Method	SMD	QWK	% exact agreement	*r*
China	Task C	MLR	−0.062	0.497	42.024	0.599
		SVM	**−0.083**	0.529	42.481	0.609
	Task D	MLR	−0.181	0.604	62.299	0.694
		SVM	−0.141	**0.592**	63.227	0.697
Korea	Task C	MLR	−0.037	0.518	40.663	0.623
		SVM	**−0.061**	0.556	41.326	0.641
	Task D	MLR	−0.090	0.675	62.437	0.746
		SVM	−0.070	0.677	63.181	0.753
Japan	Task C	MLR	−0.162	0.556	43.013	0.618
		SVM	−0.124	0.585	43.936	0.643
	Task D	MLR	−0.093	0.677	64.341	0.746
		SVM	−0.046	0.683	66.085	0.756
India	Task C	MLR	0.238	0.442	39.049	0.569
		SVM	0.192	0.469	39.384	0.579
	Task D	MLR	0.226	0.571	56.689	0.701
		SVM	0.213	0.589	57.601	0.703

Note. In this table, the agreement statistics from the SVM models that are worse than those from the MLR models are bolded. SMD = standardized mean difference; QWK = quadratic-weighted kappa; MLR = multiple linear regression; SVM = support vector machine.

Table 9 Correlations Between Human Scores and External Variables and Correlations Between E-rater Scores and External Variables

		Score from human rater	*e-rater* score SVM	*e-rater* score MLR
Task A	Verbal	0.595	0.604	0.597
Task B	Verbal	0.552	0.551	0.551
Task C	Speaking (S)	0.598	0.559	0.543
	Listening (L)	0.674	0.569	0.550
	Reading (R)	0.636	0.587	0.570
	Sum of S, L, and R	0.745	0.666	0.646
Task D	Speaking (S)	0.593	0.562	0.562
	Listening (L)	0.561	0.531	0.540
	Reading (R)	0.545	0.563	0.576
	Sum of S, L, and R	0.654	0.641	0.652

Note. MLR = multiple linear regression; SVM = support vector machine.

Finally, we compared the correlations between the human scores of writing and the same examinees' scores on other sections of the test with the correlations between e-rater scores and the same examinees' scores on the other sections using both MLR and SVM models with tuned parameters. Table 9 presents the correlations. For Assessment I, the verbal score measures examinees' reading ability and reasoning skills. We compared the correlations between students' verbal scores and essay scores generated from different scoring methods (i.e., human scoring, e-rater scoring using MLR model, and e-rater scoring using SVM model with tuned parameters).

For Assessment II, the reading score measures examinees' ability to read academic texts; the listening score measures examinees' listening comprehension of lectures, classroom discussions, and conversations; and the speaking score measures examinees' ability to express an opinion on a familiar topic or to speak based on reading and listening tasks. We compared the correlations between students' essay scores generated from different scoring methods (i.e., human scoring, e-rater scoring using the MLR model, and e-rater scoring using the SVM model with tuned parameters) and their scores on the speaking, listening, and reading sections and the total of the speaking, listening, and reading scores of Assessment II.

If the human and the automated scores reflect similar constructs, they should relate to examinees' scores on the other sections of the test in similar ways; therefore, the correlations between automated scores and examinees' scores on the other sections of the tests should be similar to the correlations between human scores and the same scores on the other sections. If so, this similarity provides validity evidence for the automated scores.

Overall, the correlations between SVM-based e-rater scores and the scores from the other sections of the test are close to those between human scores and the scores from the other sections of the test. Furthermore, the correlations between SVM-based e-rater scores and the scores from the other sections of the test are comparable with those from the linear regression-based e-rater scores. These results suggest that SVM-based scores and human scores are related to examinees' scores on the other sections of the test to a similar extent, providing validity evidence for SVM-based e-rater scores.

Conclusions and Implications

The results from this study suggest that the SVM algorithm outperforms MLR models in predicting human scores. Overall, SVM models yielded the highest agreement between human and e-rater scores and improved the agreement between human and e-rater scores for subgroups of examinees. In addition, SVM-based e-rater scores and human scores related to students' scores on the other sections of the tests in similar ways, which provided validity evidence for SVM-based e-rater scores. In contrast, *k*-NN models did not predict human scores as well as MLR models, and RF models predicted human scores better than MLR models under some circumstances but not others. These findings indicate that the MLR models do not fully employ the useful information contained in the feature variables for predicting human scores. More sophisticated models need to be developed to improve e-rater's scoring performance.

At the same time, the interpretation of results from machine learning methods may not be as straightforward as the MLR models. The advantage of the MLR modeling approach is that the basis for the score is explicitly seen in the weight that each feature receives in predicting human scores. By contrast, methods like SVM with radial basis function kernel are opaque in the sense that the basis for the scores is not immediately apparent. This trait has limited their implementation in the current automated scoring system at ETS, especially in cases when the e-rater score is used as a contributory score. However, when e-rater scores are used as confirmatory scores, the loss of interpretability may not be as serious an issue in that case.

If the results from this study can be replicated in an operational context, a potential benefit of using SVM rather than MLR for predicting human scores is the potential to reduce the total cost of human scoring. The higher agreement of SVM-based scores and human scores translates into a reduction of the number of scores that exceed the allowable discrepancy threshold; in other words, a second rater would be required less frequently when using SVM-based prediction. This could result in a substantial reduction in the cost of human scoring. These savings in human scoring cost would not entail a sacrifice in the validity of the scores according to our analyses of the correlations between machine scores and scores on the other sections of the test.

In conclusion, the superior performance of SVM and RF-based models is not surprising because these models are statistically more sophisticated than MLR models. They can capture additional nuances of relationships between the feature scores and human scores and thus provide a better mapping between them. Different machine learning algorithms make use of different strategies to arrive at their optimal classification. The SVM algorithm, which employs a set of decision surfaces to separate the essays optimally, works best for our particular datasets.

Our findings need to be tested in future studies to see how generalizable they are across different test programs and across essays collected from different administration years. If SVM-based models consistently outperform MLR models, they have the potential to be implemented in automated essay scoring to increase the fairness of the scores and decrease the need and expense of human labor.

In future studies, we will try to use the classification methods of each algorithm (e.g., SVM-rank) to directly predict the categorical scores that human raters give to the essays. Estimation based on a model that assumes that the response is categorical will be more accurate than linear regression that assumes the response is continuous and may involve inefficient approximations to essay scores obtained by raters (Haberman & Sinharay, 2010). By applying the classification method of each machine algorithm, we expect to further improve the performance of e-rater scoring models in predicting human scores.

Notes

1 Ranking SVM is a pairwise ranking algorithm based on conventional SVM. The main difference is that the constraints in ranking SVM are defined on partial-order relationships within document pairs.

2 Starting with version 14.1, bounded scores will be used for adjudication to reduce the number of second human read rates.
3 In this study, the e-rater scores generated from linear regression were slightly different from the e-rater scores generated in operational scoring. In operational scoring, the e-rater scores generated from MLR models were adjusted so that the distribution of e-rater scores (e.g. mean and standard deviation) matched with the score distribution of human raters' scores. In this study, we did not include this adjustment step so that the results from all different models are comparable.
4 This is estimated from the rater's hourly pay rate and the average time required to score one essay.

References

Attali, Y., Bridgeman, B., & Trapani, C. (2010). Performance of a generic approach in automated essay scoring. *The Journal of Technology, Learning, and Assessment, 10*(3), 1–15. Retrieved from http://www.jtla.org

Attali, Y., & Burstein, J. (2006). Automated essay scoring with e-rater® V.2. *Journal of Technology, Learning, and Assessment, 4*(3). Retrieved from http://www.jtla.org

Breiman, L. (2001). Random forests. *Machine Learning, 45*(1), 5–32.

Breiman, L., Friedman, J., Olshen, R., & Stone, C. (1984). *Classification and regression trees*. Belmont, CA: Wadsworth Int. Group.

Breland, H. M. (1996). Word frequency and word difficulty: A comparison of counts in four corpora. *Psychological Science, 7*, 96–99.

Bridgeman, B., Trapani, C., & Attali, Y. (2012). Comparison of human and machine scoring of essays: Differences by gender, ethnicity, and country. *Applied Measurement in Education, 25*(1), 27–40.

Burstein, J., 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.

Burstein, J., Tetreault, J., & Madnani, N. (2013). The e-rater® automated essay scoring system. In M. D. Shermis & J. Burstein (Eds.), *Handbook of automated essay scoring: Current applications and future directions* (pp. 55–67). New York, NY: Routledge.

Caruana, R., & Niculescu-Mizil, A. (2006). *An empirical comparison of supervised learning algorithms*. Proceedings of the 23rd International Conference on Machine Learning, Pittsburgh, PA.

Chen, H., He, B., Luo, T., & Li, B. (2012). A ranked-based learning approach to automated essay scoring. In *Cloud and green computing (CGC) 2012 Second International Conference* (pp. 448–455). New York, NY: IEEE.

Chen, J., & Ramineni, C. (2012). *Identifying sources of discrepancy between human and automated scores*. (Unpublished manuscript).

Chen, M., & Zechner, K. (2011). Computing and evaluating syntactic complexity features for automated scoring of spontaneous non-native speech. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics 2011* (pp. 722–731). Retrieved from http://dl.acm.org/citation.cfm?id=2002564

Cristianini, N., & Shawe-Taylor, J. (2000). *An introduction to support vector machines and other kernel-based learning methods*. New York, NY: Cambridge University Press.

Haberman, S. J., & Sinharay, S. (2010). The application of the cumulative logistic regression model to automated essay scoring. *Journal of Educational and Behavioral Statistics, 35*, 586–602.

Heilman, M., & Madnani, N. (2013). ETS: Domain adaptation and stacking for short answer scoring. In *Proceedings of the Second Joint Conference on Lexical and Computational Semantics (*SEM): Vol. 2. Proceedings of the Seventh International Workshop on Semantic Evaluation (SemEval)* (pp. 275–279). Retrieved from http://www.aclweb.org/anthology/S13-2046

Joachims, T. (2006). Training linear SVM in linear time. In *Proceedings of the ACM Conference on Knowledge Discovery and Data Mining (KDD)* (pp. 217–226). New York, NY: ACM. Retrieved from http://dl.acm.org/citation.cfm?id=1150429

Nehm, R. H., Ha, M., & Mayfield, E. (2012). Transforming biology assessment with machine learning: Automated scoring of written evolutionary explanations. *Journal of Science Education and Technology, 21*(1), 183–196.

Petersen, N. S. (1997, March). *Automated scoring of writing essays: Can such scores be valid?* Paper presented at meeting of the National Council on Education, Chicago, IL.

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

Quinlan, T., Higgins, D., & Wolff, S. (2009). *Evaluating the construct-coverage of e-rater®* (Research Report No. RR-09-01). Princeton, NJ: Educational Testing Service. http://dx.doi.org/10.1002/j.2333-8504.2009.tb02158.x

Ramineni, C., Trapani, C. S., Williamson, D. M., Davey, T., & Bridgeman, B. (2012). *Evaluation of the e-rater scoring engine for the GRE Issue and Argument prompts* (Research Report No. RR-12-02). Princeton, NJ: Educational Testing Service. http://dx.doi.org/10.1002/j.2333-8504.2012.tb02284.x

Ramineni, C., & Williamson, D. M. (2013). Automated essay scoring: Psychometric guidelines. *Assessing Writing, 18*, 25–39.

Samuel, A. (1959). Some studies in machine learning using the game of Checkers. *IBM Journal of Research and Development, 3*(3), 211–229

Santos, V. D. O., Verspoor, M., & Nerbonne, J. (2012). Identifying important factors in essay grading using machine learning. In D. Tsagari, S. Papadima-Sophocleous, & S. Ioannou-Georgiou (Eds.), *International experiences in language testing and assessment — Selected papers in memory of Pavlos Pavlou* (pp. 295 – 309). Frankfurt au Main, Germany: Peter Lang GmbH.

Skellam, J. (1952). Studies in statistical ecology. I. Spatial pattern. *Biometrica, 39*, 346 – 362.

Smola, A. J., & Schölkopf, B. (2004). A tutorial on support vector regression. *Statistics and Computing, 14*(3), 199 – 222.

Vapnik, V. N. (1995). *The nature of statistical learning theory*. New York, NY: Springer-Verlag.

Weigle, S. C. (2010). Validation of automated scores of TOEFL iBT tasks against non-test indicators of writing ability. *Language Testing, 27*(3), 335 – 353.

Williamson, D., Xi, X., & Breyer, F. (2012). A framework for evaluation and use of automated scoring. *Educational Measurement: Issues and Practice, 31*(1), 2 – 13.

Yannakoudakis, H., Briscoe, T., & Medlock, B. (2011). A new dataset and method for automatically grading ESOL Texts. In *Proceedings of the 49th annual meeting of the Association for Computational Linguistics* (pp. 180 – 189). Stroudsburg, PA: Association for Computational Linguistics. Retrieved from http://www.aclweb.org/anthology/P11-1019

Zechner, K., & Bejar, I. (2006). Towards automatic scoring of non-native spontaneous speech. In *Proceedings of the 2006 Conference on Human Language Technology and the North-American Association for Computational Linguistics* (pp. 216 – 223). Stroudsburg, PA: Association for Computational Linguistics. Retrieved from http://www.aclweb.org/anthology/N06-1028

Erratum

In Research Report no. RR-16-04, some content was included in error. Corrections were made to this report to remove that content. ETS apologizes for this error.

Suggested citation:

Chen, J., Fife, J. H., Bejar, I. I., & Rupp, A. A. (2016). *Building e-rater® scoring models using machine learning methods* (ETS Research Report No. RR-16-04). Princeton, NJ: Educational Testing Service. http://dx.doi.org/10.1002/ets2.12094

Action Editor: Keelan Evanini

Reviewers: Michael Heilman and Jiangang Hao