Research Report

On-the-Fly Customization of Automated Essay Scoring

Yigal Attali

Research & Development

December 2007
RR-07-42

On-the-Fly Customization of Automated Essay Scoring

Yigal Attali
ETS, Princeton, NJ

December 2007

As part of its educational and social mission and in fulfilling the organization's nonprofit charter and bylaws, ETS has and continues to learn from and also to lead research that furthers educational and measurement research to advance quality and equity in education and assessment for all users of the organization's products and services.

ETS Research Reports provide preliminary and limited dissemination of ETS research prior to publication. To obtain a PDF or a print copy of a report, please visit:

http://www.ets.org/research/contact.html

Abstract

Because there is no commonly accepted view of what makes for good writing, automated essay scoring (AES) ideally should be able to accommodate different theoretical positions, certainly at the level of state standards but also perhaps among teachers at the classroom level. This paper presents a practical approach and an interactive computer program for judgment-based customization.

This approach is based on the AES system, *e-rater*®. Through this new approach, a user can gain easy accessibility to system components, flexibility in adjusting scoring parameters, and procedures for making scoring adjustments that can be based on only a few benchmark essays. The interactive prototype program that implements this approach allows the user to customize *e-rater* and watch the effect on benchmark essay scores as well as on score distributions for a reference testing program of the user's choice. The paper presents results for the use of this approach in customizing *e-rater* to the standards of different assessments.

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

As early as 1966, Page developed an automated essay scoring (AES) system and showed that an automated rater is indistinguishable from human raters (Page, 1966). In the 1990s, more systems were developed; the most prominent systems are the Intelligent Essay Assessor (Landauer, Foltz, & Laham, 1998), Intellimetric (Elliot, 2001), a new version of the Project Essay Grade (PEG; Page, 1994), and *e-rater*® (Burstein, Kukich, Wolff, Lu, & Chodorow, 1998).

With all of the AES systems mentioned above, a scoring scheme is developed by analyzing a set of typically a few hundred essays written on a specific prompt and prescored by as many human raters as possible. In this analysis, the most useful variables (or features) for predicting the human scores, out of those that are available to the system, are identified. Then, a statistical modeling procedure is used to combine these features and produce a final machine-generated score of the essay.

As a consequence of this data-driven approach of AES, whose aim is to best predict a particular set of human scores, both what is measured and how it is measured may change frequently in different contexts and for different prompts. This approach makes it more difficult to discuss the meaningfulness of scores and scoring procedures.

e-rater Version 2 (V.2) presents a new approach in AES (Attali & Burstein, 2006). This new system differs from the previous version of *e-rater* and from other systems in several important ways that contribute to its validity. The feature set used for scoring is small, and the features are intimately related to meaningful dimensions of writing. Consequently, the same features are used for different scoring models. In addition, the procedures for combining the features into an essay score are simple and can be based on expert judgment. Finally, scoring procedures can be applied successfully to data from several essay prompts of the same assessment. This means that a single scoring model is developed for a writing assessment, consistent with the human rubric that is usually the same for all assessment prompts in the same mode of writing. In *e-rater* V.2, the whole notion of training and data-driven modeling is considerably weakened.

This paper presents a radical implementation of the score modeling principles of *e-rater* V.2, which allows a user to construct a scoring model with only a few benchmark essays of his or her choice. This can be achieved through a Web-based application that provides complete control over the modeling process.

The paper describes the statistical approach that allows modeling on the basis of a small set of essays and presents experiments for validating the approach. The success of the procedure was investigated in three experiments: (a) a simulation study based on essays written by students in Grades 6–12, (b) an experiment using state assessment essays and teachers, and (c) an experiment with GRE® essays and raters.

Description of *e-rater* Scoring and the On-the-Fly Application

The on-the-fly approach rests on an adaptation of the three scoring elements that are regularly used for *e-rater* V.2 scoring. In its regular implementation, *e-rater* scoring is based on a large set of analyzed essays in order to estimate parameters necessary for scoring. On the other hand, in the on-the-fly implementation, previously collected data and results are used as the source of parameters. The regular approach is termed here *estimated-parameter* (EP) scoring, whereas the on-the-fly approach is termed *predetermined-parameter* (PP) scoring.

In short, scoring with *e-rater* V.2 proceeds (both in EP and PP scoring) by first computing a set of measures of writing quality from the essay text. These measures have to be standardized in order to combine them into an overall score. The standardized measures are combined by calculating a weighted average of the standardized values of the measures. Finally, this weighted average is transformed to a desired scale, usually a 1–6 scale.

The feature set used with *e-rater* includes eight measures: grammar, usage, mechanics, style, organization, development, vocabulary, and word length. Attali and Burstein (2006) provided a detailed discussion of these measures. In addition, two prompt-specific vocabulary usage features are sometimes used. However, in contrast to the standard eight features, the prompt-specific vocabulary features require a large sample of prompt-specific essays in order to calculate their values. The other features require essay data only to interpret the values in the context of producing an overall score. This data requirement for the prompt-specific vocabulary features is prohibitive for their use in on-the-fly scoring. Attali and Burstein also showed that these features' contribution to scoring in many types of prompts is small and that their reliability is low compared to the other features.

Scoring Example

Table 1 shows a simplified scenario that exemplifies the scoring process for a single essay and introduces the parameters necessary for scoring. This example has only two features, A and B. In order to score essays, the means, SDs, and relative weights of features are needed, in addition to the correlations between features and final scaling parameters. The means, SDs, feature correlations, and weights that are used in scoring are presented in the first two rows of the table. These can be obtained in different ways under EP or PP scoring, as is discussed below. The raw feature values for the example essay are 110 and .35, and the standardized feature values are 1.0 and 0.5.

Table 1

Scoring Example

	M	SD	R with other feature	Relative weight	Example raw value	Example scaled value
Feature A	100.00	10.00	0.5	70%	110.00	1.00
Feature B	0.30	0.10	0.5	30%	0.35	0.50
Standardized weighted score, *Z*	0.00	0.89[a]				0.85[b]
Final score, *E*	3.5	1.2				4.65

[a] Based on a .5 correlation between two features. [b] Weighted average of standardized feature values.

The third row in Table 1 presents the distribution parameters and example value of the standardized weighted scores. These scores are computed as the sum product of standardized feature values and their weights, which for the example essay is equal to 0.85 (1.0 x 70% + 0.5 x 30%). The mean of this distribution is equal to 0 by definition. The SD of the standardized weighted scores depends on the intercorrelations between features. In this example there is only one such correlation (between A and B), which is assumed to be .5. To compute the variance of the standardized weighted scores, the formula in Equation 1should be used:

$$\sum_i w_i^2 + 2\sum_{i<j} w_i w_j r_{ij} = 0.7^2 + 0.3^2 + 2 \cdot [0.7 \cdot 0.3 \cdot 0.5] = 0.79 \quad (1)$$

Where w_i is the feature weight, r_{ij} is the intercorrelation of features, and the standardized feature SDs are equal to 1. Thus, the SD of standardized weighted scores should be .89 (the square-root of .79).

The fourth row in Table 1 shows possible (human) criterion scaling parameters that the final scores should be scaled to, in this case with a mean of 3.5 and SD of 1.2. When the standardized weighted score value of .85 is scaled according to these parameters, the resulting final score is 4.65.

To summarize, *e-rater* scores are calculated as a weighted average of the standardized feature values, followed by applying a linear transformation to achieve a desired scale. The following sections outline how this procedure can be implemented with a very small set of essays: on-the-fly.

Determining Feature Weights On-the-fly

The first element in the scoring process is identifying the relative feature weights (expressed as percentages). Although relative weights could (in the EP approach), be based on statistical optimization methods, like multiple regression, Attali and Burstein (2006) suggested that nonoptimal weights do not necessarily lower the agreement of machine scores with human scores. Specifically, they argued that a single program-level model should be preferred over the traditional prompt-level models on theoretical grounds, although they are nonoptimal for each individual prompt. In addition, an analysis of a wide range of scoring models (from sixth graders to college students and English-as-a-second-language learners) showed that the statistically optimal weights of these diverse models were remarkably similar (Attali & Burstein, 2006). Finally, Ben-Simon and Bennett (2006) studied the effect of setting weights in *e-rater* on the basis of judgments by content experts with good results. To summarize, PP alternatives in setting relative weights can be based on either content expert judgments or previous models of similar assessments.

Determining Feature Distributions On-the-Fly

The second element in the scoring process is identifying the means and SDs to be used in standardizing each feature values, and the correlations between features to be used for calculating the variance of the standardized weighted scores. Obviously, many essays (and their corresponding feature values) are needed to obtain an accurate estimate of the feature means, SDs, and intercorrelations for a relevant population of essays. However, PP scoring requires an alternative approach. Instead of estimating feature distributions and intercorrelations every time a scoring model is developed, typical estimates from previous assessments can be used. These typical values may not be accurate for a particular assessment, but results in this paper suggest that it is possible to use them without compromising the quality of scores.

Determining Final Scaling Parameters

The last step in scoring requires scaling the standardized weighted scores to final scores. This step should be based on a paired set of parameters: the mean and SD of the standardized weighted scores (in the third row of Table 1) and of corresponding human scores (in the fourth row of Table 1).

In the usual EP scenario, where a scoring model is developed based on a large set of training essays with associated human scores, these paired sets of parameters are developed based on the same training sample. The mean and SD of standardized weighted scores are based on feature parameters and intercorrelations (as in the example above), and the final scaling parameters are equal to the mean and SD of the corresponding human scores for the training sample essays.

Final scaling in PP scoring is similar, in that a training set of human-scored essays is still used to estimate the two sets of scaling parameters. However, in PP scoring the training set is used *only* for scaling. Feature standardization and feature weights are not based on this training sample, but on past results. Therefore, the training sample in PP scoring is termed the *scaling sample*.

In PP scoring, standardized weighted scores are developed for the scaling sample, based on the predetermined parameters. Similarly to the EP scenario, the mean and SD of the standardized weighted scores for the scaling sample (labeled M_Z and S_Z) as well as their corresponding human scores (labeled M_H and S_H) can be computed. However, it is important to note that M_Z and S_Z are not necessarily equal to the original values that were obtained in

developing the scoring parameters that were reproduced for PP scoring. For example, in PP scoring, M_Z is not necessarily equal to 0. However, in PP scoring as in EP scoring, the relation between M_Z-S_Z and M_H-S_H determines the final scaling of scores. Scaling of a standardized weighted scores (Z) to final *e-rater* scores (E) is done by matching the mean and SD of the scaling sample *e-rater* scores to the human mean and SD scores in the scaling sample. This is accomplished through Equation 2, applied on any essay, for either a scaling sample essay or a new essay:

$$E = \frac{S_H}{S_Z}(Z - M_Z) + M_H \tag{2}$$

From Equation 2 the scaling parameters can be extracted. The slope and intercept of the linear transformation are shown in Equation 3:

$$E = \left[\frac{S_H}{S_Z}\right] Z + \left[M_H - \frac{S_H}{S_Z} M_Z\right] \tag{3}$$

After applying this formula to the essays in the scaling sample, the mean and SD of *e-rater* scores in the scaling sample will be the same as the human scores.

Statistical Issues

In the previous section, PP scoring was described in relation to regular EP scoring. The PP approach is based on borrowing parameters from previously developed scoring models. In this section, the effects of adopting incorrect parameters and the influence of essay training sample size are explored from a statistical point of view.

Expected Magnitude of Errors in Predetermined Parameters

PP scoring is based on previous estimates of feature distributions obtained from an independent set of essays. The assumed feature distributions (those adopted from previous results) may be different from the actual feature distributions in the population of essays for which the new PP scoring is developed. It is important to evaluate the effect of discrepancies between the assumed and actual feature distributions on the quality of scoring.

Discrepancies are possible in means and in SDs of features. Discrepancies in feature SDs will affect the actual weight that features will have in the final *e-rater* score. In general, when the

actual SD of a feature is relatively larger than its assumed SD, it will have a larger influence in the final score than its assumed weight. The effect is relative to the actual-to-assumed SD ratio for other features. That is, if all actual SDs are larger (to the same degree) than assumed, the actual weights will correspond to the assumed weights. Discrepancies in feature means will not have an effect on relative weights and should not have an effect on scores, since the final scaling is based on essay scores in the training sample.

Therefore, in this section an estimate of the possible magnitude of discrepancies in feature standard errors (that is, in sample SDs) is computed. In the following section, the effect of these possible discrepancies on relative weights is estimated.

In order to evaluate the magnitude of possible discrepancies in feature standard errors, a large dataset of actual essays was analyzed. It includes essays of students in Grades 6–12 that were submitted to an online writing instruction application, *Criterion*℠, developed by ETS. In addition, the dataset includes GMAT essays written in response to issue and argument prompts and *Test of English as Foreign Language*™ (TOEFL®) essays. Overall, 64 prompts are included, with an average of 400 essays per prompt. Table 2 shows the mean and variability in the sample SD of *e-rater* feature values across prompts. Also shown is the coefficient of variation (CV) for this same statistic, a measure of relative variability of scores. CV is computed as the ratio of the SD of a variable (in this case the variable is the sample SDs) to the mean of the variable and is expressed in percentages. Table 2 shows that, except for one higher *CV* of 26%, all *CV*s are between 11% and 15%. This result is based on an average sample size of 400 essays.

Through these CV values, it is possible to estimate the possible magnitude of discrepancies in feature SDs in a typical application of PP scoring. If the mean SD values were chosen as the assumed SDs of feature values, we could expect discrepancies between assumed and actual SDs of around 15%.

Effect of Errors in Feature SDs on Relative Weights

The purpose of this section is to provide an estimate, through a simulation, of the effect of different magnitudes of discrepancies in feature SDs on discrepancies between assumed and actual relative weights. In this simulation, 10 standard normal variables that simulated possible (standardized) essay features were generated for 1,000 essays. The number of features (10) chosen for the simulation was arbitrary; the purpose of the simulation was to demonstrate different degrees of discrepancy in feature SDs. The feature values were generated such that the

correlation between features was .35. This correlation was selected for two reasons: It is the median intercorrelation among *e-rater* features in the dataset analyzed in the previous section, and simulating different intercorrelations would be very difficult.

Table 2

Sample Distribution (Across 64 Prompts) of the Feature SD Statistic

Feature	M	SD	CV
Grammar	0.72	0.08	11%
Usage	0.65	0.10	15%
Mechanics	0.95	0.11	12%
Style	0.08	0.02	26%
Organization	0.53	0.06	12%
Development	0.44	0.06	14%
Vocabulary	5.18	0.80	15%
Word length	0.29	0.04	15%

Note. *CV* is coefficient of variation, the ratio of SD to mean score.

The main purpose of the simulation was to observe the effect of wrong assumptions about feature SDs in modeling. Therefore, the assumed SDs of the features varied, some smaller and some larger than actual SDs, which were always equal to 1 (assumed and actual SDs are presented in Table 3).

Equal weights (10%) were used in computing scores for each essay in order to simplify the comparison of discrepancy effects on the different features. Standardized weighted scores were computed in the prescribed manner by standardizing the features and then using equal weights to sum the feature values. The standardization was computed once with the actual SD values and once with the assumed values.

To evaluate the relative influence of each feature (and corresponding discrepancy) on the two kinds of standardized weighted scores, a multiple regression analysis of the composite scores on the features was performed, and the standardized parameter values for each feature were compared. These standardized parameter values are presented in Table 3. Obviously, the actual (or true) parameters are all equal to 0.1, because all simulated features have the same influence on the composite scores. However, Table 3 shows that when the assumed SDs were used in

standardization of features, features with smaller assumed SD resulted in larger observed influence on composite scores. The larger observed influence was proportional to the ratio of actual-to-assumed SD. For example, the assumed SD of Feature 7 was 15% larger than its actual SD. Consequently, when features were standardized based on their (erroneously) assumed SDs, the observable influence of this feature on composite scores was about 20% smaller than its true influence.

Table 3

Effects of Discrepancies Between Assumed-to-Actual Feature SDs on Standardized Betas

				Standardized betas based on		
Feature	Assumed SD	Actual SD	Inverse SD ratio	Assumed SD	Actual SD	Beta ratio
1	0.55	1.00	1.82	0.17	0.1	1.66
2	0.65	1.00	1.54	0.14	0.1	1.40
3	0.75	1.00	1.33	0.12	0.1	1.22
4	0.85	1.00	1.18	0.11	0.1	1.07
5	0.95	1.00	1.05	0.10	0.1	0.96
6	1.05	1.00	0.95	0.09	0.1	0.87
7	1.15	1.00	0.87	0.08	0.1	0.79
8	1.25	1.00	0.80	0.07	0.1	0.73
9	1.35	1.00	0.74	0.07	0.1	0.68
10	1.45	1.00	0.69	0.06	0.1	0.63

Beyond the effects on the relative influence of individual feature, it is interesting to see what the overall influence of the feature SD errors is on the overall composite scores. The correlation between the two composite scores in this simulation was practically perfect (.995). Considering the relatively large errors that were examined in this simulation and the relatively small fluctuations in feature SDs that can be expected in practice (see previous section), it seems that feature standardization would not constitute a detrimental factor on the quality of PP scoring.

Standard Error of Means for the Scaling Procedure

The final scaling of the standardized weighted scores is primarily based on the discrepancy between the mean of standardized weighted scores and human scores for a sample of benchmark essays. For a given sample of essays and of their corresponding initial *e-rater* scores, the sample mean of human scores is only an estimate of that value over all possible human raters and is subject to sampling error. In order to evaluate how small that sample can be, it is important to estimate the SD of the sample mean, the standard error of the means (σ_M).

The value of σ_M can be estimated from a single sample by the formula in Equation 4:

$$\sigma_M = \frac{\sigma_H}{\sqrt{n}} \tag{4}$$

Where σ_H is the SD of the human scores (each score is the average of all its human ratings) and n is the number of essays in the sample. It should be noted that the number of raters that rate every essay influence the value of σ_H, with smaller values for higher number of raters.

In the case of PP scoring, each human score is related to a standardized weighted score. Thus, the conditional distributions of human scores given their initial standardized weighted scores have smaller variability than the SD of a random sample of human scores. Their SD is equal to the standard error of estimating human scores from *e-rater* scores. The standard error of estimate when predicting a human score H from a given value of *e-rater* score E is denoted $\sigma_{H.E}$ and computed as shown in Equation 5:

$$\sigma_{H.E} = \sigma_H \sqrt{1 - \rho_{HE}^2} \tag{5}$$

Where σ_H is the SD of the human scores and ρ_{HE} is the correlation between human and *e-rater* scores.

Finally, ρ_{HE}, the correlation between human scores and *e-rater* scores, can be shown to be dependent on the correlation between a human score based on a single human rating and the *e-rater* scores (ρ_{SE}), the reliability of human scores based on a single rating (ρ_{SS}), and the number of raters (k). This follows from the correction for attenuation formula for validity coefficients and from the Spearman-Brown formula for the reliability of a composite (see Lord & Novick,

1968, p. 114, for a discussion of the effect of test length on the correlation between two variables).

Specifically, the correlation between the human and *e-rater* scores is related to their true-score correlations and their reliabilities, as shown in Equation 6:

$$\rho_{HE} = \rho_{T_H T_E} \sqrt{\rho_{HH}} \sqrt{\rho_{EE}} \tag{6}$$

Since the true-score correlation is not influenced by the number of raters that form the human scores, the relation between ρ_{SE} and ρ_{HE} is related only to the increased reliability of human scores based on more raters, through the Spearman-Brown formula shown in Equation 7:

$$\rho_{HH} = \frac{k\rho_{SS}}{1+(k-1)\rho_{SS}} \tag{7}$$

Therefore, using the Spearman-Brown formula, we can express the relation between ρ_{SE} and ρ_{HE} as Equation 8:

$$\rho_{HE} = \rho_{SE} \sqrt{\frac{k}{1+(k-1)\rho_{SS}}} \tag{8}$$

The standard error of the mean of the human scores that are assigned to the scaling sample is given by Equation 9:

$$\sigma_M = \frac{\sigma_{H.E}}{\sqrt{n}} = \frac{\sigma_H \sqrt{1-\rho_{HE}^2}}{\sqrt{n}} \tag{9}$$

Where the previous formula can be plugged into ρ_{HE}.

The two parameters that affect the size of σ_M are the sample size of essays n and the number of raters that score each essay k. This is apart from σ_H, ρ_{SE}, and ρ_{SS}, which can be regarded as constants in a specific application.

Figure 1 shows the actual values of σ_M for typical n and k values, when σ_H for a single rater (k = 1) was set to 1.0 points; ρ_{SE} was set to .80, a typical correlation between a single human rating and machine scores; and ρ_{SS} was set to .64.

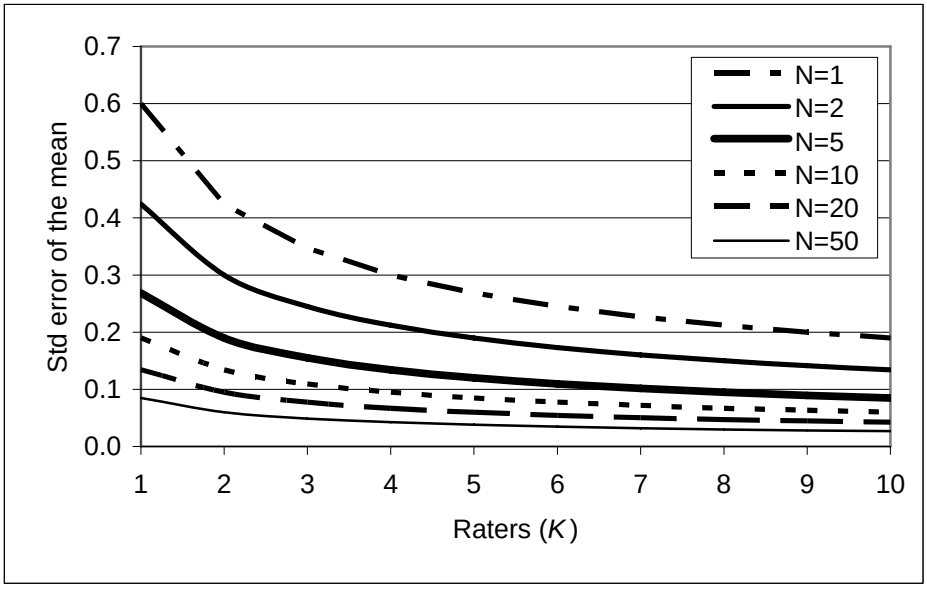

Figure 1. **Standard error of means for various number of essays (N) and number of raters (K).**

Figure 1 shows that the gain in σ_M by using more than 20 essays or more than 5 raters is very small. For 20 essays and 5 raters, the calculated σ_M is .06. For 50 essays and 5 raters, the calculated σ_M is .04.

It is instructive to compare a typical σ_M value under PP scoring, where it is determined by $\sigma_{H.E}$, to the σ_M that would be obtained if a random sample of human scores was used to scale the *e-rater* scores, based on σ_H (see Equation 4). The difference between a PP-based σ_M and an EP-based σ_M is dependent on the value of ρ_{HE} (higher values lower the PP-based σ_M), which in turn depends on *k* (higher number of raters raises the value of ρ_{HE}). Beginning with the original value of ρ_{HE} (or ρ_{SE}) for one rater (.80), the value of ρ_{HE} is .88 for two raters, .92 for three raters, .94 for four raters, and .97 for 10 raters.

Based on these values of ρ_{HE}, we can compute how much larger σ_H would be than σ_{HE} for different numbers of raters. From that, we can deduce how much larger the EP sample size would have to be, compared to the PP sample size, to have the same σ_M. Higher number of raters entail a larger advantage for EP scoring in terms of sample size. For example, for two raters, σ_H will be more than two times (2.1) larger than σ_{HE}. In other words, under EP scoring, we would need a random sample 4.5 times (2.1^2) larger to get the same σ_M under EP scoring. For five raters, σ_H will be more than three times (3.2) larger than σ_{HE}. Thus, under EP scoring we would

need a random sample more than 10 times (3.2^2) larger to get the same σ_M under EP scoring. These are very significant gains in sample sizes required for developing a new scoring application.

Evaluations of PP Scoring

In this section, several empirical evaluations of PP scoring are presented. In all these evaluations, real essay data were used to develop scores based on previous parameters and, for scaling, on very small sets of training samples. The agreement between these PP scores and human scores was compared to the agreement performance of other scores, either EP scores or human scores.

The K–12 Experiment

In the first evaluation, PP scoring was applied to samples of essays written by students using the *Criterion* application at different grades (see Table 4). The dataset included about 7,600 essays written on 36 topics from Grades 6–12, with an average of about 200 essays per topic and 5 topics per grade. The essays were scored by two trained human raters according to grade-level rubrics.

Table 4

Descriptive Statistics on Essays and Average Human Score

Grade	Prompts	Mean # of essays per prompt	M	SD
6	5	203	3.01	1.16
7	4	212	3.21	1.20
8	5	218	3.50	1.29
9	4	203	3.65	1.24
10	7	217	3.39	1.23
11	6	212	3.90	1.08
12	5	203	3.61	1.22

PP scoring was applied in the following manner. The parameters that would be used for PP scoring were obtained from a single EP model that was built for all ninth-grade essays in the sample. The following optimal weights were obtained for this EP model: grammar, 11%; usage, 15%; mechanics, 11%; style, 8%; organization, 28%; development, 13%; vocabulary, 9%; and word length, 6%. These relative weights, together with the feature distributions for ninth-grade essays, were used throughout the experiment.

For each of the remaining 32 topics (from Grades 6–8 and 10–12), a random sample of 30 essays was chosen as the prompt-specific scaling sample for PP scoring. For each of the essays in the scaling sample, a standardized weighted score was computed (based on the parameters from the ninth-grade model) in addition to the human scores available for the essays. As described above, the discrepancy between the human scores and the standardized weighted scores was used to produce the scaling parameters for new essays. Both the predetermined parameters and the scaling parameters then were applied to the remaining essays of the prompt.

For comparison with the PP scoring, EP *e-rater* scoring was implemented on the remaining essays from each topic (excluding the 30 essays in the PP scaling sample). A six-fold method was used for building and cross-validating EP scoring. In this method the *e-rater* model is built on 5/6 of all essays, and then the model is applied to the 1/6 of essays that were left out. The procedure is repeated six times.

Table 5 presents a summary of the results in comparing PP and EP performances on the cross-validation samples (for EP scoring, every essay is used once in a cross-validation sample). Table 5 shows that the PP approach performance based on 30 essays is very similar to the EP performance that was based on around 150 essays (5/6 of the remaining essays).

Table 5

Summary of Model Performance, Relation Between e-rater and Human Scores, for 32 Topics

Scoring	Kappa	Correlation	Exact agreement
Estimated parameters	.39	.78	.53
predetermined-parameters	.38	.78	.52

The State Assessment Experiment

The purpose of the Indiana experiment was to evaluate the PP scoring approach in a context where content experts score benchmark essays specifically for *e-rater* PP scoring. In the previous evaluation, the human scores were given and were produced as part of a previous research effort. The writing assessment that was used in this evaluation was Indiana's Core 40 End-of-Course Assessment in English 11 writing test. This test is scored operationally by *e-rater*. The raters were 12 Indiana teachers chosen to conduct the scaling sessions.

The data used for this experiment included four sources:

1. Source of standardization and weighting parameters: All 11th-grade essays in the *Criterion* application dataset described above were used to develop an EP *e-rater* model, from which parameters were retrieved.

2. Topics: *e-rater* scoring was developed for two topics. Topic A was the operational topic in the spring 2004 administration of the Indiana test, and Topic B was a candidate topic for the 2005 administration.

3. PP scaling sample: For scaling purposes, the Indiana teachers rated sets of 25 essays. Four sets were used, two for Topic A (A1 and A2) and two for Topic B (B1 and B2).

4. Validation samples: Two sets of 300 essays were used for validation of PP scoring, one for each of the Topics A and B.

The scoring sessions took place on 2 consecutive days. On the 1st day, after an introduction to the Indiana rubrics, the teachers scored each essay in the four scaling sets (25-essay sets) and discussed their scoring. For each set, the teachers started by individually scoring each essay in the set and then continued with discussions of problematic essays, after which they could correct their scores (although all scores were recorded). The teachers were allowed to assign half-point scores if they wished.

On the 2nd day, every teacher scored a random sample from the validation sets. The plan was that each essay would be scored twice by different raters. However, in practice not all validation essays were scored.

Table 6 presents descriptive statistics for the scaling sample scoring. In addition to the average of 12 raters before and after revision, Table 6 shows results of 9 select raters before revision. The 3 raters excluded showed biases in their scores compared to the other 9 raters. The

differences between the different measures of human ratings were small, but there were differences between the first and second set for each topic. The scores for the second set were higher than for the first set. The order of scoring on the 1st day was A1, B1, A2, and B2; it seems the raters were not calibrated fully from the start of the sessions.

The last row in Table 6 shows information about the *anchor* score. The anchor score is the *e-rater* score from the 11th-grade *Criterion* model, whose parameters were used for PP scoring in this experiment. A remarkable result in Table 6 is the very large difference between the human scores and the *e-rater* anchor scores, about .9 even for the A2 and B2 sets. These differences indicated that the scoring standards of the human raters were much higher than the *Criterion* scoring standards. The columns labeled *r* in Table 6 present the correlations between average human scores and *e-rater* anchor scores. These were around .97 and .93 for A2 and B2, respectively.

Table 6

Descriptive Statistics for Benchmark Scoring

	A1			A2			B1			B2		
Raters	M	SD	*r*	M	SD	*r*	M	SD	*r*	M	SD	*r*
12 raters	2.30	1.25	.91	2.68	1.27	.97	2.59	1.20	.92	2.71	1.42	.94
9 raters	2.30	1.28	.91	2.75	1.31	.97	2.59	1.27	.91	2.70	1.43	.93
12 rev. raters	2.22	1.16	.91	2.66	1.24	.97	2.56	1.18	.89	2.69	1.43	.92
Anchor score	3.60	1.46		3.61	1.48		3.60	1.46		3.58	1.47	

Note. Anchor score is the *e-rater* score from the 11th-grade *Criterion* model; *r* is correlation between average human score across raters and *e-rater* anchor score. All scores on a 1–6 scale.

Because of the differences in average scores between the first (A1 and B1) and second scaling sets (A2 and B2), only A2 and B2 results were used for PP scaling. In addition, the average of the 9 raters was used as the basis for scaling instead of the full 12 raters (although there were very small differences in the means and SDs of scores). The scaling was performed separately for each topic, although, as Table 6 shows, the scaling for the two topics was very similar.

Table 7 presents the distribution of human and *e-rater* PP scores for the validation essays. The human raters assigned some half-point scores, which were rounded up. Table 7 shows that the average PP *e-rater* scores were higher than the human scores by about 0.2 points and had SDs about 0.2 smaller than those of the human scores.

Table 8 shows the agreement results between human and *e-rater* scaled scores for the evaluation essays. The agreement statistics between the two human raters were very low, and the *e-rater* agreement with the human scores was higher than the interhuman agreement.

Table 7

Descriptive Statistics for Validation Scoring, With Human Scores Rounded Up

Scoring	N	Mean	SD
Topic A			
H1	288	3.26	1.18
H2	289	3.26	1.17
e-rater	291	3.11	1.03
Topic B			
H1	264	3.13	1.22
H2	263	3.07	1.21
e-rater	266	2.92	1.03

Note. H1 and H2 are first and second human scores. *e-rater* score is the scaled score based on PP scoring.

Table 8

Agreement Results for Validation Scoring (Human Scores Rounded)

	Kappa	Correlation	Exact agreement
H1-Scaled	.27	.64	.45
H2-Scaled	.26	.65	.45
H1-H2	.19	.59	.38

A Computerized Interface for On-the-Fly Modeling

The principles that underlie the PP scoring approach could be implemented through a computerized interface that allows users to customize *e-rater* scoring through example essays of the user's choice. Such an interface was developed as a Web-based application that allows users to load benchmark essays and adjust the scoring parameters to produce a customized *e-rater* scoring model. Figure 2 shows a screen-capture from this application. After loading a few benchmark essays (Step 1), the user determines relative weights to each of the dimensions measured by *e-rater* (Step 2; in this application, the word length feature was not represented). Then the scoring standards (Step 3) and score variability (the difference in scores between essays with different qualities, Step 4) are adjusted. These adjustments are reflected continuously in the essay scores to the left of the essay text. Finally, the user can select a reference program (*Criterion*'s ninth-grade program is shown in Figure 2) to see immediately the effect of the changing standards on the entire distribution of scores for this program. The score distribution is also updated continuously with any adjustments in scoring standards.

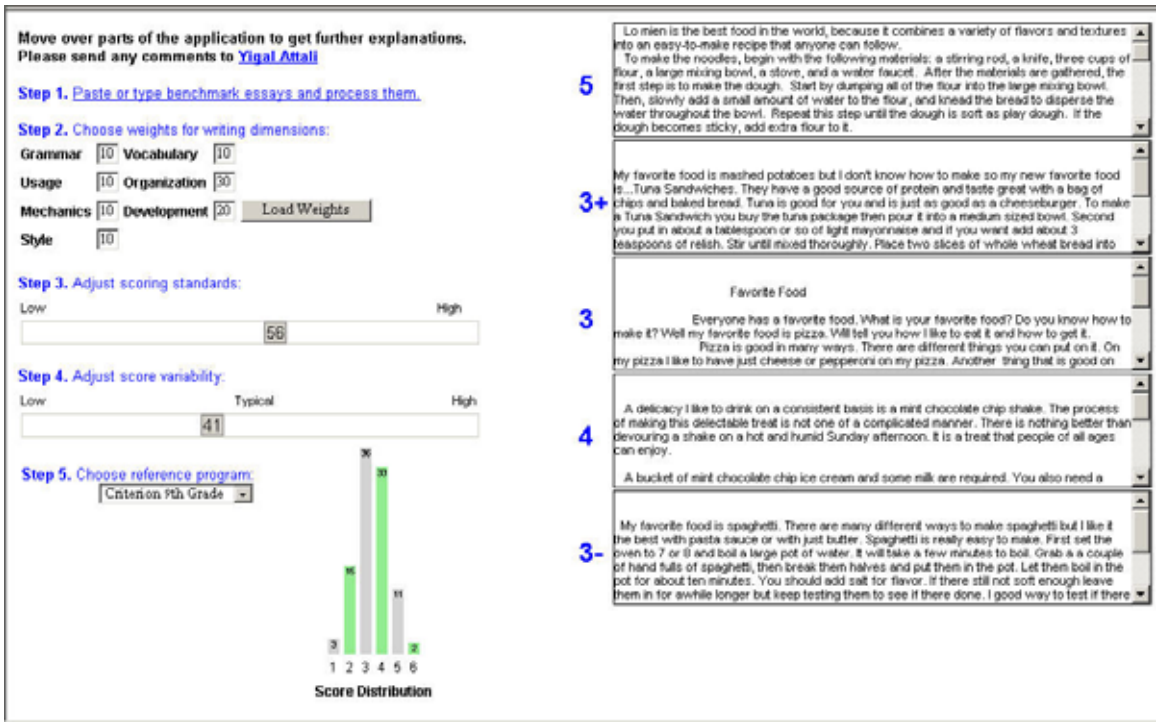

***Figure 2*. On-the-fly modeling application, ninth-grade *Criterion* program.**

The application computes scores in the following way. Feature distributions and intercorrelations are based on the large dataset that was described in the beginning of the Statistical Issues section of this paper. All parameters are computed from the average statistic values across the 64 prompts. By combining these parameters with the relative weights chosen by the user in Step 2, the standardized weighted scores can be computed. The adjustments in Steps 3 and 4 change the scaling parameters of the final scores. The score distributions of specific programs in Step 5 are approximated from the feature distributions of each program.

The GRE Experiment

The purpose of this experiment was to evaluate PP scoring with content experts who use the computerized interface with a very small number of benchmark essays. Five GRE test developers used this application to develop a scoring model for a single topic, "Present Your Perspective on an Issue." Each rater used the application five times with different sets of benchmark essays. Each set included five essays. The models developed for each set by the raters were validated on a validation set of about 500 essays. All benchmark and validation essays were scored previously by two raters.

The procedure each rater followed was to load in turn the essays from each set and adjust the scoring standards and score variability of the essays. The raters did not adjust the component weights, which were set to the values shown in Figure 2.

The application was slightly altered in order to prevent the raters from copying their settings from one benchmark set to the other. Every time a set of essays was loaded into the application, the scaling of the two sliders in Steps 3 and 4 of Figure 2 were changed randomly, so that the participants would have to find the best settings for every set independently. Therefore, if the same set of essays were loaded two different times, and the same setting for the sliders were chosen in these two occasions, the scores shown for the essays would be different.

In addition to scaling through the application, the raters provided independent scores of each essay. These scores were not necessarily identical to the application scores, because the participants were not able to accommodate any combination of scores in using the application. For example, if a participant thought that essay *x* should get a higher score than essay *y* but the application score of *x* was lower than *y*, the participant could not reverse the rank order of the two essay scores through the two slides. Such a reversal could be achieved only with changes in the relative weights of components, which was not possible in this experiment. The participants

reported that such cases where they were not able fully to accommodate their scoring preferences were common.

Table 9 presents the mean and SD of the application scores of each rater for each set. Because the essays in each set were not necessarily of the same quality, the average scores of different sets, as well as their variability, should not be the same. Similarity of scores should be expected between raters (columns). Overall, the most significant differences could be found in the lower mean score of Rater 1 and in the higher SD of Rater 4. Rater 1 gave consistently lower scores than the other raters, thus this rater's results were not included in the computation of the scaling parameters.

Table 9

Descriptive Statistics for Application Scores

	Mean						SD					
	Rater						Rater					
Set	1	2	3	4	5	All	1	2	3	4	5	All
1	2.4	3.6	3.3	3.4	3.4	3.2	1.3	1.0	1.1	2.0	1.5	1.4
2	2.8	3.5	3.4	3.4	3.3	3.3	1.3	1.1	1.3	1.3	1.2	1.2
3	3.2	3.3	3.8	3.7	3.5	3.5	1.0	1.0	1.0	1.2	1.2	1.1
4	3.4	4.4	3.8	4.0	3.5	3.8	1.0	0.9	1.0	1.5	0.9	1.1
5	3.2	3.7	3.9	4.0	3.0	3.6	0.7	0.6	0.7	0.9	0.6	0.7
All	3.0	3.7	3.6	3.7	3.4	3.5	1.0	0.9	1.0	1.4	1.1	1.1

Table 10 presents the same information about the independent scores of the raters. The independent scores were somewhat lower and more variable than the application scores. Note also that the independent scores of Rater 1 were closer to the scores of the other raters than the scaled scores are. The results of Tables 9 and 10 also can be compared with the original human scores for the benchmark essays. Table 11 presents the mean and SD of the average of the two human scores for each set. Table 11 shows that the original human scores were higher than the new panel scores.

Table 10

Descriptive Statistics for Independent Scores

	Mean						SD					
	Rater						Rater					
Set	1	2	3	4	5	All	1	2	3	4	5	All
1	2.6	3.1	3.3	2.9	3.3	3.0	1.5	1.1	0.9	1.7	1.7	1.4
2	3.1	3.5	3.3	3.2	3.3	3.3	2.0	1.6	1.4	1.6	1.5	1.6
3	3.1	3.4	3.5	3.5	3.5	3.4	0.8	1.2	1.1	1.8	1.2	1.2
4	3.6	3.7	3.6	3.5	3.5	3.6	1.5	1.1	1.1	1.6	1.3	1.3
5	3.5	3.8	3.8	3.2	3.4	3.5	1.6	1.1	1.0	1.8	1.0	1.3
All	3.2	3.5	3.5	3.3	3.4	3.4	1.5	1.2	1.1	1.7	1.3	1.4

Table 11

Descriptive Statistics for Original Human Scores

	Mean	SD
Set 1	3.7	1.4
Set 2	3.6	1.4
Set 3	3.8	1.4
Set 4	3.9	1.5
Set 5	3.9	1.4
All	3.8	1.4

The scores (both application and independent) that the raters produced for the benchmark essays were used as the scaling sample to generate *e-rater* scores for the validation set of 496 essays that were available for this topic. The scaling parameters were determined for each set separately based on the scores of Raters 2–5.

Table 12 summarizes the agreement results of various scores with the operational H1 score on the validation set. The first score to be compared with H1 is H2, the second operational human score for these essays. Next is an *e-rater* EP score based on optimal weights that was developed from the validation sample. The third score is an *e-rater* EP score, which was developed from the validation sample but with the same (nonoptimal) weights that were used in

the application (see Figure 2) by the raters. Following these scores are the application and independent *e-rater* scores from the five scaling sets.

Table 12

Agreement With Operational H1 (M = 3.5, SD = 1.0) on Validation Sample

Set	M	SD	Kappa	Exact agreement	Correlation
H2	3.6	1.0	0.68	0.76	0.89
EP optimal	3.5	1.0	0.54	0.66	0.83
EP semi-optimal	3.5	1.1	0.43	0.58	0.78
PP application					
Set 1	3.6	1.0	0.41	0.57	0.77
Set 2	3.3	1.1	0.39	0.55	0.79
Set 3	3.8	1.1	0.33	0.50	0.79
Set 4	3.2	1.0	0.38	0.55	0.79
Set 5	3.3	1.0	0.40	0.56	0.79
PP independent					
Set 1	3.3	0.9	0.43	0.59	0.79
Set 2	3.2	1.2	0.33	0.49	0.80
Set 3	3.8	1.4	0.27	0.43	0.81
Set 4	2.8	1.1	0.13	0.33	0.79
Set 5	3.0	1.5	0.20	0.36	0.80

Table 12 shows that the human agreement (H1/H2) was significantly higher than any of the human-to-machine agreements. Even the EP optimal scores showed lower agreement with H2 than H1 did, and the optimal scores performed better than the semi-optimal scores.

Semi-optimal EP score performance can be used as benchmark for PP score performance because they share the same relative weights. The average kappa for the application scores was .38, and the average kappa for the independent scores was .27. It seems that the main reason for lower performance of PP scores was discrepancies in the mean and SDs of scores, compared with the human scores. This was most evident with independent scores. The scaling of

application scores was more consistent and similar to that of the human scores. Considering the very small sample the application scores were based on (4 raters and five essays), their level of agreement with human scores is remarkable.

Summary

The three evaluations that were presented in this paper were significantly different from each other, but all three provided evidence that the on-the-fly approach is feasible. The *Criterion* simulation was based on samples of 30 essays and used actual operational scores, two per essay. The PP performance results were almost identical to EP performance based on around 200 essays. The Indiana evaluation was based on new scores produced by 9 raters for training samples of 25 essays. The human–machine agreement of the PP scores on the validation data was comparable to the human–human agreement. Finally, the GRE evaluation was based on new scores for five essays by 4 raters and was validated on previously available operational scores. Although in this evaluation the agreement of the PP scores with human scores fell below human–human agreement, it was only slightly lower than the agreement of an optimal model with the same feature weights as the PP scores.

This rapid approach to *e-rater* modeling may be used by prospective users either to customize *e-rater* to a new assessment or to adapt the scoring standards of an existing assessment. An example of the former is a state assessment considering the use of *e-rater*. An example of the latter is teachers interested in adjusting scoring standards for their students who use an application like *Criterion*. In either case, the essays used for the customization can be provided by the application itself or loaded by the user. As a first step in the implementation of such a system, Redman, Leahy, and Jackanthal (2006) performed a usability study of the application with *Criterion* teachers. They reported that the teachers were very enthusiastic about using the computerized application for customizing the *e-rater* standards used to score their students' essays. It is also clear that a detailed user manual would have to be created for teachers to use this application.

This paper does not provide a definite answer to the question of how many essays and raters are needed to achieve reasonable confidence in the accuracy of standards. The answer to this question also depends on the stakes involved in scoring decisions. However, Figure 1 suggests that the effect of increasing the number of essays is stronger than an increase in the number of raters; this is similar to the finding that an increase of one to the number of essays in a

writing assessment has a larger effect on reliability than an increase of one to the number of raters per essay (Breland, Camp, Jones, Morris, & Rock, 1987). The three experiments do not allow a systematic evaluation of this hypothesis. In the K–12 experiment, 30 essays and 2 ratings per essay were used. In the state assessment experiment, 25 essays and 9 ratings per essay were used. In the GRE experiment, 5 essays and 4 ratings per essay were used. An interesting replication of the GRE experiment that would test the minimal settings for customization could use 10 essays instead of 5.

Two scoring and scaling approaches were used in the evaluations. The state assessment raters scored each essay independently of others and did not directly set *e-rater* standards. The GRE raters, on the other hand, directly set standards in a computerized interface, and their scores were derived collectively from these standards. It seems that the "standards-first" approach is more suited to small numbers of essays, but it also may be more frustrating to users because they are not free to set individual essay scores.

The computerized interface allows a third approach to scaling that relies on the ability to examine the resulting score distributions of reference programs as scoring standards are being changed. This ability could serve as an important tool for potential users. In certain applications, the scoring of example essays could serve only a secondary purpose of providing examples of the standards, whereas the main adjustments of standards are performed vis-à-vis the reference programs deemed relevant to the user.

References

Attali, Y., & Burstein, J. (2006). Automated essay scoring with *e-rater*® V.2. *Journal of Technology, Learning, and Assessment, 4*(3). Retrieved October 12, 2007, from http://www.jtla.org

Ben-Simon, A., & Bennett, R.E. (2006, April). *Toward theoretically meaningful automated essay scoring*. Paper presented at the annual meeting of the National Council of Measurement in Education, San Francisco.

Breland, H. M., Camp, R., Jones, R. J., Morris, M. M., & Rock, D. A. (1987). *Assessing writing skill* (Research Monograph No. 11). New York: College Entrance Examination Board.

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

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

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

Lord, F. M., & Novick, M. R. (1968). *Statistical theories of mental test scores*. Reading, MA: Addison-Wesley.

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

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

Redman, M., Leahy, S., & Jackanthal, A. (2006). *A usability study of a customized e-rater score modeling prototype*. Unpublished manuscript, ETS.