Listening. Learning. Leading.®

Research Report
ETS RR–11-25

# Use of e-rater® in Scoring of the TOEFL iBT® Writing Test

Shelby J. Haberman

June 2011

# Use of e-rater® in Scoring of the TOEFL iBT® Writing Test

Shelby J. Haberman
ETS, Princeton, New Jersey

June 2011

As part of its nonprofit mission, ETS conducts and disseminates the results of research to advance quality and equity in education and assessment for the benefit of ETS's constituents and the field.

To obtain a PDF or a print copy of a report, please visit:

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

**Technical Review Editor:** Matthias von Davier

**Technical Reviewers:** Sandip Sinharay and Gautam Puhan



**Abstract**

Alternative approaches are discussed for use of e-rater® to score the TOEFL iBT® Writing test. These approaches involve alternate criteria. In the 1st approach, the predicted variable is the expected rater score of the examinee's 2 essays. In the 2nd approach, the predicted variable is the expected rater score of 2 essay responses by the examinee on a parallel form. This 2nd approach is related to prediction of the expected rater score of 2 essay responses on an actual form taken later by an examinee. The relationship of e-rater scores to scores on other sections of TOEFL® is also considered. These alternative approaches suggest somewhat different procedures for scoring.

Key words: classical test theory, reliability, validity

**Acknowledgments**

Data for the study was provided by Cathy Trapani and Vincent Weng. The selection of alternative weighting schemes and the evaluation criteria reflect suggestions by Yigal Attali, Brent Bridgeman, Tim Davey, Neil Dorans, Marna Golub-Smith, and David Williamson. The author thanks Sandip Sinharay, Gautam Puhan, and Matthias von Davier for helpful comments.

In the TOEFL iBT® Writing test, e-rater® is currently used in scoring of the two writing tasks. In this report, criteria are considered for evaluation of the current scoring procedure and for selection of other scoring procedures that make use of e-rater. In section 1, the criterion of scoring accuracy is considered. In section 2, behavior of e-rater and human scores on repeat examinations is explored. In section 3, analysis from sections 1 and 2 is applied to provide an analysis based on reliability measurement. In section 4, the relationship of e-rater and human scores to other portions of the TOEFL iBT test is examined. In section 5, some conclusions are provided. Analysis is somewhat restricted because only two prompts are available in the Writing assessment for a given administration. In addition, the data are of quite variable quality, and the results are quite dependent on the criteria used.

## 1 Scoring Accuracy

In the TOEFL iBT examination, writing is assessed by use of two tasks: an independent task in which an opinion must be supported in writing and an integrated task in which an examinee must respond to both reading material and spoken material. At the introduction of the TOEFL iBT examination, two human raters normally scored each task, and each rater assigned a holistic score of 1 to 5 to the response. Some exceptions to this situation arose in such cases as off-topic responses, blank responses, and scores assigned to the same response that differed by more than one point. During 2009, e-rater and a human rater were typically employed to score the independent task, although some exceptions arose due to essays not appropriate for e-rater or due to large discrepancies between e-rater and human scores. Beginning late in 2010, both in the case of the independent task and in the case of the integrated task, e-rater and a human rater have typically been employed to score the response.

To date, e-rater has been treated in the TOEFL® test as a substitute score for a human score, save that e-rater produces a continuous score and human raters produce only integer scores. In this report, e-rater will be regarded as a predictor of a human score rather than as a substitute. This approach can yield somewhat different results in terms of weighting of human and e-rater scores and in terms of definitions of unusual discrepancy. The basic data in this part of the analysis are obtained from a sample of 139,134 TOEFL examinees for whom two human scores from 1 to 5 and an e-rater score were available for the responses to both the independent and integrated tasks. The data are from administrations in the first 10 months of 2008. They are not

a random sample of examinees, for the number of responses per prompt is nearly constant but the number of examinees in an administration is quite variable. This procedure can be expected to overweight examinees in Western countries and to underweight examinees in Eastern countries. In addition, the data do not include several months in 2008 in which examinee performance is typically relatively high. A further bias results from the fact that the administrations with highest volumes are typically in times of the year that are relatively less represented in the sample. In short, any analysis based on these data should be approached with the caution appropriate for a biased sample.

In the analysis in this report, the e-rater scores for each tasks are derived from a linear regression using the generic-model approach on the sampled data from 2008 (Attali & Burstein, 2005). For each task, a linear transformation is applied to e-rater scores to match the sample mean and sample variance of the average of two human ratings for the sample of responses used in the regression analysis. In current practice, e-rater scores are truncated so that no score is permitted to be less than 0.5001 or greater than 5.4999. This truncation affects a relatively small fraction of the sample, about 0.5% of examinees for the independent task and about 3.76 of examinees for the integrated task. For each task, about 80 of truncations arise because the e-rater score is less than 0.5001 without truncation. In this report, both e-rater scores with truncation and e-rater scores without truncation are considered.

### 1.1 Human Scoring

To begin analysis, it is appropriate to summarize some basic features of the human scoring for the sample under study. For the integrated task, the average of human scores is 3.09, and the average score for the independent task is 3.39. The tasks differ substantially in terms of variability. For the integrated task, the sample standard deviation of a human rater is 1.19, and the sample standard deviation of a human rater is 0.83 for the independent task. Prior to use of e-rater, the raw score $S_r$ for the Writing test was normally the sum $S$ of the average human rating for the integrated task and the average human rating for the independent task, although exceptions arose when two human scores on the same response differed by more than 1. The scaled Writing score was then obtained by rounding a linear transformation of the raw score to the nearest integer within the range 0 to 30. The sample correlation of $S$ and $S_r$ is 0.996, and $S$ and $S_r$ differ for only about 2.2% of the examinees in the sample. As a consequence, it appears reasonable to apply $S$

in subsequent analysis. This result is consistent with earlier work at ETS in the 1980s (Mazzeo, Schmitt, & Cook, 1986a, 1986b). Use of $S$ has the advantage that linear theory is readily applied.

Both $S_r$ and $S$ have sample mean 6.47 and standard deviation 1.71. The average of the two human scores for the integrated task has sample standard deviation 1.14, and the average of the two human scores for the independent task has sample standard deviation 0.76. The sample correlation of the average human score for the integrated task and the average human score on the independent task is 0.61. In the case of two items with unequal sample variances, the estimated Cronbach $\alpha$ (Lord & Novick, 1968, p. 204) of the score $S$ is

$$\alpha_S = \frac{4(1.14)(0.76)(0.61)}{1.71^2} = 0.72.$$

Thus reliability of the Writing test is an obvious cause for concern. Nonetheless, as is well known, $\alpha$ provides a lower bound to reliability, so that the actual reliability may be higher. Some further work on estimation of this reliability will be considered in section 2.

An appreciable fraction of the error of measurement of the raw score $S$ is clearly due to the human rating, although the reliability of $S$ is not very high even if variability of human raters would disappear. For a pair of examinee responses, let $U$ be the expected value of $S$ obtained by regarding the human raters as randomly drawn from the pool of raters used in scoring. The estimated Cronbach $\alpha$ of $S$ implies that the estimated variance of measurement of $S$ is no greater than $0.83 = 1.71^2(1 - 0.72)$. For either of the two tasks, the estimated variance of scoring error is one quarter of the mean squared difference of the two corresponding rater scores. The estimated variances are 0.12 for the integrated task and 0.11 for the independent task. The estimated variance of scoring error for $S$ is 0.24 (note rounding error), the sum of the estimated variances of scoring error for the two tasks, so that the estimated variance of measurement of $U$ is no greater than $0.83 - 0.24 = 0.59$, the estimated variance of $U$ is $1.71^2 - 0.24 = 2.70$, and the Cronbach $\alpha$ of $U$ is still only $\alpha_U = 1 - 0.59/2.70 = 0.78$.

### 1.2 Prediction of Human Scoring by e-rater

By Kelley's formula (Kelley, 1947), the best linear predictor of $U$ by $S$ is estimated to be $6.47 + R^2(S - 6.47)$, where the estimated coefficient of determination $R^2$ is $1 - 0.24/2.70 = 0.91$. A basic question involving automated scoring is how well $U$ can be predicted by one human score $h_{g1}$ on the integrated prompt, one human score $h_{d1}$ on the independent prompt, the e-rater score

$e_g$ on the integrated prompt, and the e-rater score $e_d$ on the independent prompt. This question is closely related to a similar issue that has been studied for a single essay score (Haberman & Qian, 2007). The same methodology leads to the following prediction for $U$:

$$\hat{U} = 0.74 + 0.79h_{g1} + 0.16e_g + 0.36h_{d1} + 0.47e_d.$$

In this prediction formula, a best linear predictor of $U$ from $h_{g1}$, $h_{d1}$, $e_g$, and $e_d$ is constructed. The covariance matrix of the predictors is readily estimated from the available sample. The covariance of $U$ and $e_{g1}$ is the same as the covariance of $S$ and $e_{g1}$ because the e-rater score does not involve the rater error, and the covariance of $S$ and $e_{g1}$ is readily estimated from the sample data. Similarly, the covariance of $U$ and $e_{d1}$ can be estimated without difficulty. Let $h_{g1}$ be decomposed into the sum $H_g + r_{g1}$, where the scoring error $r_{g1}$ has mean 0 and is uncorrelated with $H_g$. Let $h_{d1}$ be decomposed into the sum $H_d + r_{d1}$, where the scoring error $r_{d1}$ has mean 0 and is uncorrelated with $H_g$, $H_d$, and $r_{g1}$. Then the covariance of $U$ and $h_{g1}$ is the sum of the variance of $H_g$ and the covariance of $h_{g1}$ and $h_{g1}$. The covariance of $U$ and $h_{d1}$ is the sum of the variance of $H_d$ and the covariance of $h_{g1}$ and $h_{d1}$. The covariance of $h_{g1}$ and $h_{d1}$ can be estimated easily from the sample data. The variance of $H_g$ is the variance of $h_{g1}$, which is estimated from the sample data, minus the variance of $r_{g1}$, which is twice the estimated variance of scoring error for the integrated task. A similar argument applies to $H_d$.

Given the existing estimate of the variance of $U$, one finds that the coefficient of determination $R^2$ is 0.83. By this criterion, prediction of $U$ by $h_{g1}$, $e_g$, $h_{d1}$, and $e_d$ rather than $S$ does entail an appreciable decrease in $R^2$.

The prediction of $U$ is almost unaffected if e-rater is not considered at all for the integrated prompt. In this case, the prediction for $U$ is

$$\hat{U}_1 = 0.59 + 0.83h_{g1} + 0.38h_{d1} + 0.61e_d.$$

The resulting $R^2$ is 0.82.

If e-rater scores and human scores are used interchangeably, so that $U$ is predicted by a linear function of

$$Q = (h_{g1} + e_g + h_{d1} + e_d)/2,$$

then the predictor is

$$\hat{U}_2 = 0.75 + 0.88Q$$

and $R^2$ decreases to 0.79. If human scores are used but no e-rater scores are employed, then the predictor is

$$\hat{U}_3 = 1.15 + 0.93h_{g1} + 0.72h_{d1}$$

and $R^2$ is also 0.79. A simple alternative predictor is a linear function of

$$Q_1 = h_{g1} + (h_{d1} + e_d)/2.$$

In this case, the predictor is

$$\hat{U}_4 = 0.75 + 0.88Q_1$$

and $R^2$ is 0.82.

During meetings in 2010 of the Technical Advisory Committee on Automatic Scoring, other alternative predictors with simple weights were suggested. These include the following:

$$Q_2 = \frac{2h_{g1} + e_g}{3} + \frac{h_{d1} + e_d}{2},$$

$$Q_3 = \frac{2h_{g1} + e_g + h_{d1} + e_d}{2.5},$$

and

$$Q_4 = \frac{2h_{g1} + e_g + h_{d1} + 2e_d}{3}.$$

One has the predictor

$$\hat{U}_5 = 0.63 + 0.90Q_2$$

with $R^2 = 0.81$, the predictor

$$\hat{U}_6 = 0.92 + 0.87Q_3$$

with $R^2 = 0.82$, and the predictor

$$\hat{U}_7 = 0.66 + 0.90Q_4$$

with $R^2 = 0.81$.

The current use of e-rater for scoring both TOEFL prompts is not exactly a linear function of the human scores $h_{g1}$ and $h_{d1}$ and the e-rater scores $e_g$ and $e_d$. As already noted, the e-rater scores are truncated. In addition, sufficiently large discrepancies between e-rater and corresponding human scores lead to additional use of human raters. The resulting approximation to $S$ will be written as $V$. The lack of linearity prevents use of the analytical methods in this section. Nonetheless, the functions $S$, $\hat{U}$, $\hat{U}_1$, $\hat{U}_2$, $\hat{U}_3$, $\hat{U}_4$, $\hat{U}_5$, $\hat{U}_6$, $\hat{U}_7$, and $V$ used in prediction of $U$ are all

quite highly correlated, as is evident from Table 1. The relationship of $\hat{U}_2$ to $S$ is relatively weaker than is the case for the other estimates.

**Table 1**
***Correlations of Predictors of Score*** $U$

| Predictor | $S$ | $\hat{U}$ | $\hat{U}_1$ | $\hat{U}_2$ | $\hat{U}_3$ | $\hat{U}_4$ | $\hat{U}_5$ | $\hat{U}_6$ | $\hat{U}_7$ | $V$ |
|---|---|---|---|---|---|---|---|---|---|---|
| $S$ | 1.00 | 0.96 | 0.96 | 0.93 | 0.96 | 0.96 | 0.95 | 0.95 | 0.94 | 0.95 |
| $\hat{U}$ | 0.96 | 1.00 | 1.00 | 0.98 | 0.98 | 1.00 | 0.99 | 1.00 | 0.99 | 0.98 |
| $\hat{U}_1$ | 0.96 | 1.00 | 1.00 | 0.98 | 0.98 | 0.99 | 0.98 | 0.99 | 0.98 | 0.97 |
| $\hat{U}_2$ | 0.93 | 0.98 | 0.98 | 1.00 | 0.92 | 0.95 | 0.99 | 0.99 | 0.99 | 0.98 |
| $\hat{U}_3$ | 0.96 | 0.98 | 0.98 | 0.92 | 1.00 | 0.99 | 0.95 | 0.96 | 0.94 | 0.95 |
| $\hat{U}_4$ | 0.96 | 1.00 | 0.99 | 0.95 | 0.99 | 1.00 | 0.98 | 0.98 | 0.98 | 0.97 |
| $\hat{U}_5$ | 0.95 | 0.99 | 0.98 | 0.99 | 0.95 | 0.98 | 1.00 | 1.00 | 1.00 | 0.99 |
| $\hat{U}_6$ | 0.95 | 1.00 | 0.99 | 0.99 | 0.96 | 0.98 | 1.00 | 1.00 | 1.00 | 0.98 |
| $\hat{U}_7$ | 0.94 | 0.99 | 0.98 | 0.99 | 0.94 | 0.98 | 1.00 | 1.00 | 1.00 | 0.98 |
| $V$ | 0.95 | 0.98 | 0.97 | 0.98 | 0.95 | 0.97 | 0.99 | 0.98 | 0.98 | 1.00 |

By the criterion of prediction of human scoring, it follows that e-rater has modest utility and nearly all value of e-rater is provided by e-rater for the independent prompt.

## 2 Analysis of Repeaters

Data from the TOEFL examinees included 7,747 examinees who repeated the TOEFL examination and had two human scores from 1 to 5 and e-rater scores for both Writing prompts for both administrations studied. These data are obviously rather biased given that most examinees do not repeat the TOEFL examination. As evident from Table 2, the distribution of test country in the sample of repeaters is very different than the distribution for the main sample.

**Table 2**
***Distribution of Examinees by Test Country***

| | Percent of sample | |
|---|---|---|
| Country | Main | Repeater |
| China | 12.0 | 8.0 |
| India | 7.8 | 2.5 |
| Japan | 6.9 | 13.8 |
| South Korea | 13.4 | 26.6 |
| United States | 18.1 | 24.1 |
| Other | 41.8 | 25.1 |

In addition, the distribution of examinee scores is somewhat different in the repeater sample.

In the main sample, $S$ has a mean of 6.47, a standard deviation of 1.71, and a Cronbach $\alpha$ of 0.72. In the repeater sample, for the first administration for an examinee, the mean of $S$ is 6.02, the standard deviation is 1.56, and $\alpha$ is 0.66. For the second administration, $S$ has mean 6.35, standard deviation 1.52, and $\alpha$ of 0.65. In view of the bias of the sample, considerable caution must be used in application of the data.

To begin, consider prediction of $S$ for the second TOEFL test of the examinee from $S$, $\hat{U}$, $\hat{U}_1$, $\hat{U}_2$, $\hat{U}_3$, $\hat{U}_4$, and $V$ from the first TOEFL test. Results are summarized in Table 3.

**Table 3**
***Correlations of TOEFL Scores on a Repeat Administration to TOEFL Scores on an Initial Administration***

| Predictor at first administration | $S$ at second administration $R$ | $R^2$ |
|---|---|---|
| $S$ | 0.73 | 0.53 |
| $\hat{U}$ | 0.73 | 0.54 |
| $\hat{U}_1$ | 0.73 | 0.53 |
| $\hat{U}_2$ | 0.73 | 0.54 |
| $\hat{U}_3$ | 0.69 | 0.48 |
| $\hat{U}_4$ | 0.72 | 0.52 |
| $\hat{U}_5$ | 0.74 | 0.55 |
| $\hat{U}_6$ | 0.73 | 0.54 |
| $\hat{U}_7$ | 0.74 | 0.55 |
| $V$ | 0.74 | 0.54 |

These results suggest that, with the exception of $\hat{U}_3$, the predictors are all quite comparable in terms of prediction of $S$ at the second administration. The extent to which sample bias affects the results remains an important question.

## 3 Reliability Analysis

In typical applications of augmentation, sample means, sample variances, sample covariances, and estimated Cronbach $\alpha$ values of components of a composite score are used to examine appropriate linear weighting of the observed components in order to estimate a true score of one or more test components (Haberman, 2008; Wainer et al., 2001). Best linear predictors are used. In the case under study, complications arise because each test component includes only one item, so that a Cronbach $\alpha$ cannot be estimated for each test component. In this section, an attempt at augmentation analysis is made by use of some data on repeaters in order to estimate reliability

of each item score. Because the repeater analysis involves sample bias, estimation of reliability of tasks is obtained with minimal use of the information based on repeater data. The analysis in this section is somewhat different than the analysis in Section 1, for errors due to examinee variation on item responses are considered along with variation due to scoring error.

The following decompositions divide scores into true scores, errors exclusive of rater errors, and rater errors:

$$h_{gj} = H_{Tg} + H_{Eg} + r_{gj},\ 1 \leq j \leq 2,$$

$$h_{dj} = H_{Td} + H_{Ed} + r_{dj},\ 1 \leq j \leq 2,$$

$$e_g = e_{Tg} + e_{Eg},$$

$$e_d = e_{Td} + e_{Ed}.$$

The true scores $H_{Tg}$, $H_{Td}$, $e_{Tg}$, and $e_{Td}$ are uncorrelated with the errors $H_{Eg}$, $r_{gj}$, $H_{Ed}$, $r_{dj}$, $e_{Eg}$, and $e_{Ed}$. Errors from different prompts are uncorrelated, so that $H_{Eg}$, $r_{gj}$, and $e_{Eg}$ are not correlated with $H_{Ed}$, $r_{dj}$, and $e_{Ed}$. In addition, the rater errors $r_{g1}$ and $r_{g2}$ are uncorrelated with each other and with $H_{Eg}$ and $e_{Eg}$, and the rater errors $r_{d1}$ and $r_{d2}$ are uncorrelated with each other and with $H_{Ed}$ and $e_{Ed}$. The expected value of each error component is 0, the variances of $r_{g1}$ and $r_{g2}$ are both $\sigma^2_{rg}$ and the variances of $r_{d1}$ and $r_{d2}$ are both $\sigma^2_{rd}$. The variance of $H_{Tg}$ is $\sigma^2_{HTg}$, the variance of $H_{Ed}$ is $\sigma^2_{HEd}$, and similar notation is used for other variances. The covariance of $H_{Tg}$ and $H_{Td}$ is $\gamma_{THHgd}$, the covariance of $e_{Tg}$ and $e_{Td}$ is $\gamma_{TEEgd}$, and similar conventions are applied to other covariances. The true sum

$$W = H_{Tg} + H_{Td}$$

is to be estimated by use of the human scores $h_{g1}$ and $h_{d1}$ and the e-rater scores $e_g$ and $e_d$. Note that $U$ in Section 1 is $W + H_{Eg} + H_{Eg}$. In the analysis of the main sample, it is a straightforward matter to estimate the covariance matrix of the vector with elements $h_{g1}$, $h_{d1}$, $e_g$, and $e_d$ and to estimate the variances $\sigma^2_{rg}$ and $\sigma^2_{rd}$. On the one hand, the methods of this section also demand estimation of the covariance matrix of the vector with elements $H_{Tg}$, $H_{Td}$, $e_{Tg}$, and $e_{Td}$. In the latter case, estimation of $\gamma_{THHgd}$, $\gamma_{THEgd}$, $\gamma_{TEHgd}$, and $\gamma_{TEEgd}$ is readily accomplished with the complete sample. For example, $\gamma_{THHgd}$ is also the covariance of $h_{g1}$ and $h_{d1}$. On the other hand, other elements of the covariance matrix of the true scores cannot be obtained from conventional analysis without very strong assumptions.

Nonetheless, the repeater data can be employed to obtain plausible estimates of the covariance matrix of the true scores. Consider the case of $\gamma_{THEgg}$. The repeater data provide an estimate $\tilde{\gamma}_{THEgg}$ equal to the average of the sample covariances of $h_{gj}$ from the first administration and $e_g$ from the second administration and $e_g$ from the first administration and $h_{gj}$ from the second administration for $j$ equal 1 or 2. Because bias is a concern, it is probably prudent also to consider an estimate $\tilde{\gamma}_{HEgg}$ based on the average of the sample covariances of $h_{gj}$ and $e_g$ for $j$ equal 1 or 2 for the same administration. Let $\hat{\gamma}_{HEgg}$ be the average of the sample covariances from the main sample of $e_g$ and $e_d$. Then the estimate of $\gamma_{THEgg}$ is

$$\hat{\gamma}_{THEgg} = \tilde{\gamma}_{THEgg}\hat{\gamma}_{HEgg}/\tilde{\gamma}_{HEgg}.$$

This estimate is appropriate if the ratio $\gamma_{THEgg}/\gamma_{HEgg}$ between covariances of true scores $H_{Tg}$ and $e_{Tg}$ and covariances of observed scores $h_{g1}$ and $e_g$ is the same as the corresponding ratio of covariances conditional on being from the repeater population. It is not assumed that the covariance of $h_{g1}$ and $e_g$ is the same for the complete and repeater populations. Indeed it is clear from the data that these covariances are different. To be sure, no way exists to be certain that the assumption used to derive $\hat{\gamma}_{THRgg}$ is actually valid, but the assumption is at least more limited than the assumption of equal covariances for complete and repeater populations.

Similar arguments can be used to estimate all needed variances and covariances of true scores. For this analysis, the optimal prediction of $W$ is

$$\hat{W} = 1.42 + 0.44h_{g1} + 0.27e_g + 0.27h_{d1} + 0.58e_d.$$

The resulting $R^2$ is 0.79. For prediction of $U$ rather than $W$, a linear function of $\hat{W}$ can be obtained with an $R^2$ of 0.80, so that $\hat{W}$ is a bit less effective as a predictor of scores.

Some comparisons with alternative estimates are worth consideration. Consider estimation by just the human scores $h_{g1}$ and $h_{d1}$ and the e-rater score $e_d$ for the independent prompt. In this case, the optimal prediction of $W$ is

$$\hat{W}_1 = 1.16 + 0.51h_{g1} + 0.30h_{d1} + 0.80e_d.$$

The $R^2$ is 0.77. If just two human scores are used, then the optimal prediction of $W$ is

$$\hat{W}_2 = 1.91 + 0.65h_{g1} + 0.76h_{d1}.$$

The $R^2$ is 0.69.

The optimal linear function of the equally weighted score $Q$ is

$$\hat{W}_3 = 1.54 + 0.76Q,$$

and the resulting $R^2$ is 0.79. The optimal linear function of $S$ is

$$\hat{W}_4 = 1.65 + 0.74S,$$

and $R^2$ is 0.74.

If a linear function of $\hat{U}$ is employed, then the optimal predictor is

$$\hat{W}_5 = 1.70 + 0.83\hat{U},$$

and $R^2$ is 0.77.

If a linear function of $Q_1$ is employed, then the optimal predictor is

$$\hat{W}_6 = 1.78 + 0.72Q_1,$$

and $R^2$ is 0.74.

If a linear function of $Q_2$ is employed, then the optimal predictor is

$$\hat{W}_7 = 1.52 + 0.77Q_2,$$

and $R^2$ is 0.79.

If a linear function of $Q_3$ is employed, then the optimal predictor is

$$\hat{W}_8 = 1.80 + 0.73Q_3,$$

and $R^2$ is 0.78.

If a linear function of $Q_4$ is employed, then the optimal predictor is

$$\hat{W}_9 = 1.52 + 0.77Q_4,$$

and $R^2$ is 0.79.

The reliability analysis thus suggests different weights than suggested by the analysis of scoring accuracy in Section 1. One issue of note here is that e-rater scores are much more reliable than are human scores. Recall that $S$ has a Cronbach $\alpha$ of 0.72. In contrast, from the complete

data, one finds that an assessment score that is a weighted linear combination of the two e-rater scores can have a Cronbach $\alpha$ as high as 0.86.

It should be noted that in any actual application of weighted scores, linking to the score $S$ is required in order to preserve an appropriate reporting scale.

## 4 Relationships to Other Section Scores

It is helpful to examine the relationship of raw scores for Writing with the scaled scores for Reading, Listening, and Speaking. For this purpose, the main sample can be employed, and linear regressions on the three scores are appropriate. A summary of results is provided in Table 4. This table does not discriminate very much between different estimates, but it does suggest some weakness in $\hat{U}_3$, which does not use e-rater and uses only two human scores.

**Table 4**
***Regressions of Scores on Writing on Other Scaled Section Scores***

| Dependent variable | $R^2$ |
|---|---|
| $S$ | 0.65 |
| $\hat{U}$ | 0.64 |
| $\hat{U}_1$ | 0.63 |
| $\hat{U}_2$ | 0.63 |
| $\hat{U}_3$ | 0.60 |
| $\hat{U}_4$ | 0.62 |
| $\hat{U}_5$ | 0.64 |
| $\hat{U}_6$ | 0.64 |
| $\hat{U}_7$ | 0.64 |
| $\hat{W}$ | 0.63 |

## 5 Conclusions

The analysis does not provide entirely consistent conclusions, but it appears that the current implementation of e-rater scoring has no obvious advantage over implementations that do not require additional human scorers. Table 5 summarizes different criteria for performance of alternative scoring systems. Note that this table adds a few analyses not previously described, and note that $\hat{U}_2$ and $\hat{W}_3$ are equivalent, for both are linear functions of $Q$. Similar issues affect $\hat{U}_4$, $\hat{U}_5$, $\hat{U}_6$, $\hat{U}_7$, $\hat{W}_6$, $\hat{W}_7$, $\hat{W}_8$, and $\hat{W}_9$. Scoring accuracy involves estimation of the score $U$, while

reliability analysis involves estimation of $W$.

**Table 5**
***Coefficients of Determination of Scores Based on Alternate Criteria***

| Variable | Scoring accuracy | Repeaters | Reliability analysis | Other sections |
|---|---|---|---|---|
| $S$ | 0.91 | 0.53 | 0.74 | 0.65 |
| $V$ | | 0.54 | | 0.64 |
| $\hat{U}$ | 0.83 | 0.54 | 0.77 | 0.64 |
| $\hat{U}_1$ | 0.82 | 0.53 | 0.75 | 0.63 |
| $\hat{U}_2$ | 0.79 | 0.54 | 0.79 | 0.63 |
| $\hat{U}_3$ | 0.79 | 0.48 | 0.68 | 0.60 |
| $\hat{U}_4$ | 0.82 | 0.52 | 0.74 | 0.62 |
| $\hat{U}_5$ | 0.81 | 0.55 | 0.79 | 0.64 |
| $\hat{U}_6$ | 0.82 | 0.54 | 0.78 | 0.64 |
| $\hat{U}_7$ | 0.81 | 0.55 | 0.79 | 0.64 |
| $\hat{W}$ | 0.80 | 0.54 | 0.79 | 0.63 |

The evenly weighted option $\hat{U}_2$ appears to be viable, although it exhibits some weakness in terms of correlation with the actual human score. Options $\hat{U}$, $\hat{U}_5$, $\hat{U}_6$, $\hat{U}_7$, and $\hat{W}$ all appear reasonable, and $\mathrm{Z}\hat{U}_3$ is relatively weak on all criteria other than scoring accuracy. Option $V$, when it can be evaluated, is comparable to options $\hat{U}$, $\hat{U}_5$, $\hat{U}_6$, $\hat{U}_7$, and $\hat{W}$; however, it is far more expensive to employ due to the greatly increased use of human scoring. The choice of options depends on the priorities assigned to the reliability analysis and the scoring analysis. The scoring analysis makes fewer assumptions. The reliability analysis appears to provide reasonable results, but its use of the repeater data for some calculations is problematic.

It is recognized that the TOEFL program may desire added human scoring in some cases in which e-rater and human scores appear discrepant, but such scoring should be minimized as much as feasible. Of course, double scoring of essays is needed to estimate rater reliability and to study other issues concerning rater behavior.

## References


Attali, Y., & Burstein, J. (2005). *Automated essay scoring with e-rater® v. 2.0* (ETS Research Report No. RR-04-45). Princeton, NJ: ETS.

Haberman, S. J. (2008). When can subscores have value? *Journal of Educational and Behavioral Statistics*, *33*, 204–229.

Haberman, S. J., & Qian, J. (2007). Linear prediction of a true score from a direct estimate and several derived estimates. *Journal of Educational and Behavioral Statistics*, *32*, 6–23.

Kelley, T. L. (1947). *Fundamentals of statistics.* Cambridge, MA: Harvard University Press.

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

Mazzeo, J., Schmitt, A., & Cook, L. (1986a, April). *The compatibility of adjudicated and non-adjudicated essay scores.* Paper presented at the annual meeting of the National Council on Measurement in Education, San Francisco, CA.

Mazzeo, J., Schmitt, A., & Cook, L. (1986b). *The compatibility of adjudicated and non-adjudicated essay scores on the ATP English Composition Test with Essay.* Unpublished manuscript, Educational Testing Service, College Board Statistical Analysis, Princeton, NJ.

Wainer, H., Vevea, J. L., Camacho, F., Reeve, B. B., Swygert, K. A., & Thissen, D. (2001). Augmented scores—"Borrowing strength" to compute scores based on small numbers of items. In D. Thissen & H. Wainer (Eds.), *Test scoring* (pp. 343–387). Mahwah, NJ: Erlbaum.