Listening. Learning. Leading.®

Research Report
ETS RR–15-15

Automated Trait Scores for *GRE*® Writing Tasks

Yigal Attali

Sandip Sinharay

June 2015

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

Automated Trait Scores for *GRE*® Writing Tasks

Yigal Attali & Sandip Sinharay

Educational Testing Service, Princeton, NJ

The *e-rater*® automated essay scoring system is used operationally in the scoring of the argument and issue tasks that form the Analytical Writing measure of the *GRE*® General Test. For each of these tasks, this study explored the value added of reporting 4 trait scores for each of these 2 tasks over the total e-rater score. The 4 trait scores are word choice, grammatical conventions, fluency and organization, and content. First, confirmatory factor analysis supported this underlying structure. Next, several alternative ways of determining feature weights for trait scores were compared: weights based on regression parameters of the trait features on human scores, reliability of trait features, and loadings of features from factor analytic results. In addition, augmented trait scores, based on information from other trait scores, were also analyzed. The added value of all trait score variants was evaluated by comparing the ability to predict a particular trait score on one task from either the same trait score on the other task or the e-rater score on the other task. Results supported the use of trait scores and are discussed in terms of their contribution to the construct validity of e-rater as an alternative essay scoring method.

Keywords Automated scoring; augmented trait scores; alternative feature weighting schemes

doi:10.1002/ets2.12062

For performance assessments in general and essay writing assessments in particular, the implementation of subscores usually implies the development of analytic (multitrait) scoring rubrics that can be useful for capturing examinees' specific weaknesses and strengths in writing (Weigle, 2002). Therefore, many educators believe that analytic scoring can be useful for generating diagnostic feedback to guide instruction and learning (Hamp-Lyons, 1991, 1995; Roid, 1994; Swartz et al., 1999). A well-known example of an analytic rubric for writing assessments is the 6+1 trait model (Education Northwest, 2011), which defines six traits: ideas, organization, voice, word choice, sentence fluency, conventions, and presentation (the +1 in 6+1).

However, analytic scoring has not been widely used for large-scale writing assessments for two main reasons. One reason has to do with the increased cost associated with multiple ratings of each essay instead of a single holistic score. Another is that analytic ratings have often proven less useful than expected because they are highly correlated among themselves and with holistic scores, thus rendering them redundant from a psychometric point of view (Bacha, 2001; Freedman, 1984; Huot, 1990; Lee, Gentile, & Kantor, 2008; Veal & Hudson, 1983).

Recent advances in automated essay scoring provide an opportunity to develop cost-effective trait scores that are also viable from a psychometric point of view. In particular, several aspects of the *e-rater*® scoring engine V.2 (Attali & Burstein, 2006) support the use of trait scores: The feature set used for scoring is small, and all of the features are indicators of generally acknowledged dimensions of good writing, essay scores are created by using a weighted average of the feature values, and a single scoring model is developed for a writing assessment across all assessment prompts.

In addition, factor analyses of both *TOEFL*® computer-based test (CBT) essays (Attali, 2007) and essays written by native English speakers from a wide developmental range (4th to 12th grade; Attali & Powers, 2008) revealed a similar underlying three-factor structure of the noncontent e-rater features. This three-factor structure has an attractive hierarchical linguistic interpretation with a word choice factor (measured by the vocabulary and word length features), a grammatical conventions within a sentence factor (measured by the grammar, usage, and mechanics features), and a fluency and organization factor (measured by the style, organization, and development features). Confirmatory factor analysis can help determine the subscores of a test (e.g., Grandy, 1992). That is, the number of factors is indicative of the number of subscores that can be justified and the pattern of item-factor relationships (which items load on which factors) indicates how the subscores should be scored.

Corresponding author: Y. Attali, E-mail: yattali@ets.org

Recently, Attali (2011) explored the feasibility of developing automated trait scores for the *TOEFL iBT*® independent task based on this three-factor structure. First, using a multiple-group confirmatory factor analysis, the three-factor structure was found to be quite stable across major language groups. Next, the trait scores based on these three factors were found to have added value in the context of repeater examinees by comparing the ability to predict a trait score on one test from the trait score on the other test or from the total e-rater score on the other test. For example, the correlation between the grammatical conventions scores on the first and second tests was .66, but the correlation between conventions on one test and e-rater scores on another test was .55 to .56. In other words, the conventions score in one test is the best single predictor of another conventions score in another test.

This approach to the evaluation of trait scores was inspired by Haberman (2008), who recently suggested a simple criterion to determine if subscores of a test have added value beyond the total score. The criterion is that the true subscore should be predicted better by a predictor based on the (observed) subscore than by a predictor based on the total score. Alternatively, the criterion is that the subscore on one test form should be predicted better by the corresponding subscore on a parallel form than by the total score on a parallel form (Sinharay, 2013). If these conditions are not satisfied, then instructional or remedial decisions based on the subscore will lead to more errors than those based on total scores. Analyses of test subscores often find that this condition is not satisfied (e.g., Haberman, 2008; Sinharay, 2010). One reason for this result is that subscores, often based on a small number of items, tend to have low reliability. Another reason is that the entire assessment is essentially unidimensional, with the effect that subscores, instead of measuring a unique subskill, are simply less reliable measures of the general skill measured by the total score.

The main goal of this paper was to evaluate the feasibility and added value of automated trait scores for the Analytical Writing measure of the *GRE*® General Test, which comprises two essay writing tasks, an issue task and an argument task.[1] In the issue task, the student is asked to discuss and express his or her perspective on a topic of general interest. In the argument task, a brief passage is presented in which the author makes a case for some course of action or interpretation of events by presenting claims backed by reasons and evidence. The student's task is to discuss the logical soundness of the author's case by critically examining the line of reasoning and the use of evidence.

The two tasks are scored by human raters using a holistic scoring rubric, and since October 2008, e-rater has been used operationally as part of the scoring process for both tasks. The GRE scoring rubric emphasizes ideas, development, and organization, word choice, sentence fluency, and conventions, as the following description of a typical high-scored response shows (Educational Testing Service [ETS], 2011):

- articulates a clear and insightful position on the issue in accordance with the assigned task
- develops the position fully with compelling reasons and/or persuasive examples
- sustains a well-focused, well-organized analysis, connecting ideas logically
- conveys ideas fluently and precisely, using effective vocabulary and sentence variety
- demonstrates facility with the conventions of standard written English (i.e., grammar, usage, and mechanics), but may have minor errors

There are several differences between the approach taken in Attali (2011) and the approach taken in this paper. First, the previous paper relied on test repeaters as a basis for defining reliability (and validity) coefficients. The generalizability of results of such an approach is limited because of the self-selected nature of the sample. In this paper, reliability coefficients are based on the relations between the two writing tasks, argument and issue, effectively treating these tasks as the two items of the writing assessment. Although there are noticeable differences in the demands of the two tasks, the GRE reports only one writing score, consistent with this interpretation. Accordingly, the added value of trait scores is assessed in this paper by comparing the cross-task correlations of a specific trait (e.g., the argument and issue conventions scores) to the correlations of the trait score in one task (e.g., the argument conventions score) with other scores on another task (e.g., the issue e-rater score).[2]

This paper also expands the coverage and definition of traits to include the content features of e-rater. In previous work (Attali, 2007, 2011; Attali & Powers, 2008), the three-factor structure was based only on the noncontent features. In this paper, we considered content features as well and performed factor analyses to determine whether a four-factor structure (word choice, conventions, fluency and organization, and content/ideas) is supported in the two writing tasks, and whether four trait scores based on these factors could have added psychometric value.

Finally, this paper also compares different ways to compute trait scores. First, different sources for determining the feature weights for trait scores were compared. In the traditional regression-based method, the weights (or relative

importance) of each feature in score calculation are based on a regression of the human essay scores on the essay features (those that contribute to a particular trait score). This choice is the most natural because it is also the method used to determine weights for the operational e-rater scores. A second set of weights was based on the idea that all features are equally important (and therefore should have equal weights), but that weighting should take into account differences in the reliability of features—a feature that is measured more reliably should contribute more significantly to the scores. In particular, by setting feature weights proportional to $\sqrt{r}/(1-r)$, where r is the reliability of the feature, maximum reliability of the trait score will be achieved when all traits measure the same underlying construct (Li, Rosenthal, & Rubin, 1996). A third set of weights was based on the standardized feature loadings from a confirmatory factor analysis. These loadings can be interpreted as the regression weights for predicting the underlying factor from the features that are designed to measure this factor. An important distinction between the regression-based weights and the two alternatives is that the former is based on an external criterion (prediction of human scores) whereas the latter are based on internal criteria—reliability or relation to underlying measure.

In addition to alternative feature weighting schemes, this paper also explores the use of augmented trait scores (Wainer, Sheehan, & Wang, 2000) as a way to improve the reliability of subscores. This method is based on a multivariate generalization of Kelley's classic regressed estimate of the true score (Kelley, 1927). The generalization involves multiple regression of a true subscore on all of the observed subscores on a test with the effect that the information in the observed subscore is augmented by all other observed subscores.

Method

Data

The analyses in this paper are based on 413,693 examinees who took the GRE between July 2009 and February 2010 and had a complete score record. For each test taker, several variables were available for analysis. Among them were all GRE test scores, test takers' answers to the biographical questionnaire, and other background information such as the country of the test center.

E-rater Features

The feature set used in this study is based on the features used in e-rater V.2 (see Table 1). Essay length was used in this study instead of the development feature of e-rater V.2 (Attali & Burstein, 2006) because the development feature is nearly a linear combination of the essay length and organization features.

Factor Analysis

Confirmatory factor analyses were conducted for all features. Four models were investigated, reflecting different degrees of separation between linguistic levels and content. The first model is a one-factor model without any separation. The second model is a two-factor model with the two word-choice features (vocabulary and word length) separated from the other features. The third model is a three-factor model with the fluency features separated from the grammar and content features. The fourth model has four factors, with grammar and content features separated. These models were chosen based on the intended measurement purpose of each feature, previous factor analytic results, and current exploratory factor analyses with different number of factors specified, suggesting the separation of features above.

Analyses were performed with LISREL 8.80 (Jöreskog & Sörbom, 2006), based on the covariance matrices for each writing task. The comparative fit index (CFI), nonnormed fit index (NNFI), and root mean square error of approximation (RMSEA) were used for overall model fit. Common rules of thumb were used in appraising the measures (Hoyle & Panter, 1995): .90 or more for CFI and NNFI, and .05 or less for RMSEA.

Argument Task

Table 2 presents the overall correlation matrix for the features used in this study. Correlations range from around 0 to .76.

In a preliminary principal component analysis, the first four eigenvalues were 3.71, 1.63, 1.28, and 0.82 and together accounted for 68% of total eigenvalues. The number of eigenvalues greater than 1 suggests a three-factor solution, but both a larger number of factors (four) and smaller number of factors was compared in confirmatory factor analyses.

Table 1 Features Used in This Study

Feature	Trait	Description
Vocabulary	Word choice	Based on frequencies of essay words in a large corpus of text
Word length	Word choice	Average word length
Grammar	Conventions	Based on rates of errors such as fragments, run-on sentences, garbled sentences, subject-verb agreement errors, ill-formed verbs, pronoun errors, missing possessives, and wrong or missing words
Usage	Conventions	Based on rates of errors such as wrong or missing articles, confused words, wrong form of words, faulty comparisons, and preposition errors
Mechanics	Conventions	Based on rates of spelling, capitalization, and punctuation errors
Col/prep	Conventions	Collocation and preposition use
Organization	Fluency	Based on detection of discourse elements (i.e., introduction, thesis, main points, supporting ideas, conclusion)
Essay length	Fluency	Based on number of words in essay
Style	Fluency	Based on rates of cases such as overly repetitious words, inappropriate use of words and phrases, sentences beginning with coordinated conjunctions, very long and short sentences, and passive voice sentences
Value cosine	Content	Based on similarity of essay vocabulary to prompt-specific vocabulary across score points. In this feature, the degree of similarity (values of cosine correlations) across points is analyzed.
Pattern cosine	Content	Based on similarity of essay vocabulary to prompt-specific vocabulary across score points. In this feature, the pattern of similarity across points is analyzed.

Table 2 Argument Task Feature Correlation Matrix

Feature	WL	G	U	M	CP	OR	EL	S	VC	PC
Vocabulary	.61	.19	.09	.14	.20	.11	.08	.10	.33	.37
Word length (WL)		.11	.04	.01	.08	.09	−.01	.03	.23	.29
Grammar (G)			.42	.39	.27	.21	.38	.25	.41	.35
Usage (U)				.33	.28	.14	.34	.22	.38	.33
Mechanics (M)					.31	.11	.27	.16	.39	.28
Col/prep (CP)						.06	.13	.11	.29	.25
Organization (OR)							.59	.28	.29	.19
Essay length (EL)								.49	.47	.29
Style (S)									.29	.23
Value cosine (VC)										.76
Pattern cosine (PC)										

Note. $N = 413{,}693$; Col/prep = collocation and preposition use.

Table 3 Argument Task Overall Fit of Models

Model	*df*	χ^2	CFI	NNFI	RMSEA
Four factors	38	74,422	.963	.947	.069
Three factors	41	204,490	.912	.882	.110
Two factors	43	419,644	.809	.756	.154
One factor	44	569,095	.730	.663	.177

Note. CFI = comparative fit index; NNFI = nonnormed fit index; RMSEA = root mean square error of approximation.

Table 3 presents the overall fit indices for the four confirmatory factor analysis models. The overall fit for the one- and two-factor models were unsatisfactory for all three fit indices. The overall fit for the three-factor model was better, with satisfactory CFI value. The overall fit for the four-factor model was even better, with satisfactory CFI and NNFI values, but still high RMSEA (.07). In summary, only the four-factor model showed reasonable fit.

Table 4 shows the feature loadings and error variances for the four-factor model, and Table 5 shows the interfactor correlations. Table 4 shows that some features, such as the col/prep and style features, have higher error variances. Table 5 shows that the word choice factor has the lowest correlations with other factors (.07 – .39) and content the highest (.39 – .66).

Table 4 Argument Task Four-Factor Model—Loadings and Error Variances

Factor	Feature	Loadings[a]	Error variances
Word choice	Vocabulary	.90	.19
Word choice	Word length	.67	.55
Conventions	Grammar	.67	.54
Conventions	Usage	.61	.63
Conventions	Mechanics	.58	.67
Conventions	Col/prep	.44	.81
Fluency	Organization	.59	.65
Fluency	Essay length	.99	.00
Fluency	Style	.49	.76
Content	Value cosine	.96	.08
Content	Pattern cosine	.80	.37

Note. Col/prep = collocation and preposition use.

[a]Loadings on corresponding factor. Loadings on other factors are set to zero.

Table 5 Argument Task Four-Factor Model—Factor Correlations

	Grammar	Fluency	Content
Word choice	.26	.07	.39
Grammar		.51	.66
Fluency			.47

Table 6 Issue Task Feature Correlation Matrix

Feature	WL	G	U	M	CP	OR	EL	S	VC	PC
Vocabulary	.71	.26	.08	.13	.27	.06	.02	.16	.23	.48
Word length (WL)		.18	.00	.01	.14	.05	−.06	.09	.18	.37
Grammar (G)			.44	.46	.32	.13	.30	.22	.41	.40
Usage (U)				.46	.33	.06	.28	.18	.46	.41
Mechanics (M)					.37	.07	.25	.11	.50	.37
Col/Prep (CP)						.02	.09	.10	.30	.32
Organization (OR)							.49	.14	.15	.04
Essay length (EL)								.35	.43	.20
Style (S)									.24	.24
Value cosine (VC)										.69
Pattern cosine (PC)										

Note. $N = 413{,}693$; Col/prep = collocation and preposition use.

Issue Task

Table 6 presents the overall correlation matrix for the features used in this study. Correlations range from around −.06 to .71.

In a preliminary principal component analysis, the first four eigenvalues were 3.69, 1.75, 1.36, and 0.87 and together accounted for 70% of total eigenvalues. The number of eigenvalues greater than 1 suggests a three-factor solution, but both a larger number of factors (four) and smaller number of factors were compared in confirmatory factor analyses.

Table 7 presents the overall fit indices for the four models. The overall fit for the one-, two-, and three-factor models were unsatisfactory for all three indices. The overall fit for the four-factor model was better, with satisfactory CFI value, but still low NNFI (.87) and high RMSEA (.11). In summary, only the four-factor model showed reasonable fit.

Table 8 presents the feature loadings and error variances for the four-factor model, with similar results to the argument task, except for higher error variance for pattern cosine. Table 9 presents the interfactor correlations, with similar results to the argument task—with even lower correlations for the word choice factor (−.02 – .46) and higher for the content factor (.33 – .76).

Table 7 Issue Task Overall Fit of Models

Model	*df*	χ^2	CFI	NNFI	RMSEA
Four factors	38	180,669	.909	.868	.107
Three factors	41	258,067	.873	.829	.123
Two factors	43	387,147	.810	.757	.148
One factor	44	649,298	.687	.609	.189

Note. CFI = comparative fit index; NNFI = nonnormed fit index; RMSEA = root mean square error of approximation.

Table 8 Issue Task Four-Factor Model — Loadings and Error Variances

Feature	Loadings	Error variances
Vocabulary	.90	.08
Word length	.79	.45
Grammar	.65	.57
Usage	.68	.54
Mechanics	.70	.51
Col/prep	.49	.76
Organization	.45	.80
Essay length	.99	.01
Style	.31	.90
Value cosine	.84	.30
Pattern cosine	.83	.32

Note. Col/prep = collocation and preposition use.

Table 9 Issue Task Four-Factor Model — Factor Correlations

	Grammar	Fluency	Content
Word choice	.25	−.02	.46
Grammar		.34	.76
Fluency			.33

The Three Sets of Trait Scores

The performance of several sets of trait scores was compared. The first set of scores was based on a regression analysis of the human score on the relevant features. Table 10 (columns 2 and 3) shows the relative weights (standardized regression weight divided by sum of weights of the relevant features for each trait score) for each feature. The second set of weights was based on the cross-task reliabilities (correlation between the argument and issue feature score) of the features that are shown in column 4. Weights that were based on these reliabilities were derived to achieve maximum reliability (Li et al., 1996). The weight on a feature is proportional to $\sqrt{r}/(1-r)$, where r is the reliability of the feature. The relative weights based on these reliabilities are presented in columns 5 and 6. Note that with this method the relative weights are the same for argument and issue. A third set of scores based on the factor loadings from Tables 4 and 8 are presented in the last two columns.

Inspection of Table 10 shows that the two sets of regression-based and factor analysis (FA)-based weights (for argument and issue) are similar, with the possible exception of the content features. The most important difference across types of weights is that regression-based weights are less homogeneous than any of the two other types of scores. In addition, the regression-based weights on word length and organization for the issue task are negative. This finding constitutes a serious problem because all features are expected to have a positive influence on essay scores. In an operational setting, this difference might be resolved by eliminating the feature from the score.

Regression-Based Trait Scores

Table 11 presents, for trait scores, e-rater scores, and human scores, the within-task score correlations for argument (above the diagonal) and issue (below the diagonal), as well as the cross-task correlations or reliabilities (the diagonals). For

Table 10 Alternative Relative Weights

	Regression-based		Reliability-based			FA-based	
Feature	Arg.	Issue	Reliability	Arg.	Issue	Arg.	Issue
Vocabulary	92%	114%	.46	50%	50%	57%	56%
Word length	8%	−14%	.46	50%	50%	43%	44%
Grammar	37%	32%	.47	25%	25%	29%	26%
Usage	32%	36%	.48	26%	26%	26%	26%
Mechanics	20%	25%	.58	35%	35%	25%	28%
Col/prep	11%	7%	.29	15%	15%	19%	19%
Organization	4%	−2%	.54	31%	31%	29%	26%
Essay length	88%	89%	.71	56%	56%	48%	57%
Style	8%	13%	.25	13%	13%	24%	18%
Value cosine	81%	68%	.45	48%	48%	55%	66%
Pattern cosine	19%	32%	.48	52%	52%	45%	34%

Note. FA = factor analysis; arg. = argument; col/prep = collocation and preposition use.

Table 11 Regression-Based Weights — Within-Task Score Correlations and Cross-Task Reliabilities

Score	H	E	W	G	F	C
man (H)	.56	.76	.22	.53	.66	.63
e-rater (E)	.77	.75	.28	.67	.88	.74
Word choice (W)	.27	.32	.46	.20	.08	.35
Grammar (G)	.58	.73	.21	.70	.43	.52
Fluency (F)	.64	.85	.05	.35	.70	.46
Content (C)	.59	.61	.33	.59	.39	.50

Note. Figures above diagonal are for argument, below diagonal for issue, and on diagonal for cross-task reliabilities.

example, the first value of 0.56 in the diagonal denotes the correlation between the human score on the issue task and the human score on the argument task; the value of 0.76 toward the right of it denotes the correlation between the human score on the argument task and the e-rater score on the argument task; the number 0.77 below the first diagonal value denotes the correlation between the human score on the issue task and the e-rater score on the issue task. The reliability of e-rater scores is significantly higher than human scores (.75 versus .56). The reliability of the grammar and fluency trait scores (.70) is not much lower than e-rater (.75), whereas the reliability of word choice and content is even lower than human scores (.46 and .50). Fluency shows the highest correlation with e-rater scores (.88 and .85), reflecting the high weights of essay length in e-rater scores, whereas word choice has the lowest correlations with e-rater scores (.28 and .32). The highest correlations between trait scores are between grammar and content (.52 and .59) and content and fluency (.46 and .39). Overall, the corresponding correlations in argument and issue are similar — an exception is the higher correlation between e-rater and content in argument, reflecting the higher weight of content features in this task.

Table 12 presents the cross-task correlations that form the basis for evaluating the value of trait scores. The numbers in bold show correlations between the same trait scores across the two tasks or, in other words, the reliabilities of the trait scores that were also shown in Table 11. The question for each trait score is whether these reliabilities are higher than other correlations between the trait score and other scores. The table shows that for word choice, grammar, and fluency, this is indeed the case for both tasks. For example, the highest correlation of argument word choice is with issue word choice (.46), and the next highest correlation is with the verbal score (.31). For argument fluency, the difference between the correlation with issue fluency and issue e-rater is small (.70 versus .69) but in the other five cases the differences are quite large. On the other hand, for content there are two other scores (denoted with an underline) that show higher correlations than its reliability (.50). For argument, both the issue e-rater (.50) and verbal (.53) scores show higher correlations, and for issue, both the argument grammar (.55) and argument e-rater (.53) scores show higher correlations.

Table 13 shows subgroup standardized mean scores. That is, for any variable and subgroup, the table shows the difference between the mean value of the variable in the subgroup and in the full sample, divided by the standard deviation of the variable in the full sample. For ethnic and gender comparisons, only domestic examinees are included in the population. As might be expected, international examinees show relatively small differences for word choice and fluency (compared

Table 12 Regression-Based Weights — Cross-Task Score Correlations

	Argument				Issue			
Score from other task	W	G	F	C	W	G	F	C
Word choice (W)	**.46**	.27	.20	.35	**.46**	.18	.03	.21
Grammar (G)	.18	**.70**	.34	.49	.27	**.70**	.34	.55
Fluency (F)	.03	.34	**.70**	.30	.20	.34	**.70**	.37
Content (C)	.21	.55	.37	**.50**	.35	.49	.30	**.50**
e-rater	.21	.61	.69	.50	.34	.55	.61	.53
Human	.18	.55	.53	.47	.30	.49	.45	.48
Verbal	.31	.49	.41	.53	.40	.48	.33	.48
Quant.	.16	.06	.28	.19	.15	.01	.16	−.03

Note. Boldface figures are cross-task reliabilities and expected to be highest in column. Underlined figures contradict this expectation.

Table 13 Regression-Based Weights — Subgroup Differences (Standardized Scores)

				Argument						Issue					
Group	%	Q	V	H	E	W	G	F	C	H	E	W	G	F	C
International	21	.64	−.41	−.48	−.42	−.02	−.74	−.11	−.55	−.69	−.53	−.15	−.86	−.15	−.91
Asian	6	.43	.03	.05	.10	.18	.00	.09	.06	−.06	.02	.12	.00	−.02	−.11
Black	7	−.89	−.66	−.56	−.70	−.12	−.47	−.65	−.51	−.46	−.51	−.18	−.38	−.42	−.34
Hispanic	6	−.41	−.33	−.24	−.25	−.06	−.19	−.21	−.20	−.18	−.20	−.09	−.19	−.14	−.14
Female	62	−.20	−.09	−.01	.01	−.02	.05	.00	.02	−.01	.01	−.08	.06	.00	.03

Note. Results for ethnicity and gender are limited to domestic examinees. Q = quantitative; V = verbal; H = human; E = e-rater; W = word choice; G = grammar; F = fluency; C = content.

to the reference group of domestic examinees), but large differences for grammar and content. For content, there is also a large difference between argument (−.55) and issue (−.91). Asian examinees have slightly lower scores for issue than for argument (both human and e-rater), and these differences are mainly due to lower argument content and fluency scores. Black examinees have higher issue than argument scores, but their word choice scores are lower for issue. Hispanic examinees show similar argument and issue scores and relatively homogeneous trait score patterns. Female examinees also show similar argument and issue scores and relatively homogeneous trait score patterns.

Reliability-Based Trait Scores

Tables 14 – 16 present the same results as in Tables 11 – 13 for reliability-based trait scores. Reliability-based trait scores have less heterogeneous weights, higher cross-task reliabilities (except for fluency), lower within-task correlations with other trait scores (the median difference is .04), lower cross-task correlations (the median difference is .01), and show similar subgroup differences (the median difference is .00) compared to regression-based trait scores. The combination of higher reliabilities and lower correlations with other scores slightly increases the added value of reliability-based trait scores, but the results are very similar.

Factor Analysis-Based Trait Scores

Tables 17 – 19 present similar results for FA-based trait scores. The results are very similar to the reliability-based scores, with slightly higher added value compared to regression-based scores.

Augmented Reliability-Based Trait Scores

Augmented scores (Wainer et al., 2000) were computed for the reliability-based trait scores to see if the value of the trait scores could be improved by "borrowing strength" from other trait scores, especially for the fluency and content scores. To compute augmented scores, the standardized observed trait scores were used, together with the cross-task reliability

Table 14 Reliability-Based Weights — Within-Task Score Correlations and Cross-Task Reliabilities

Score	H	E	W	G	F	C
Human (H)	.56	.76	.20	.51	.63	.61
e-rater (E)	.77	.75	.26	.65	.84	.71
Word choice (W)	.24	.28	.52	.16	.08	.36
Grammar (G)	.58	.72	.17	.72	.38	.50
Fluency (F)	.60	.80	.03	.29	.67	.40
Content (C)	.58	.59	.38	.59	.31	.51

Note. Figures above diagonal are for argument, below diagonal for issue, and on diagonal for cross-task reliabilities.

Table 15 Reliability-Based Weights — Cross-Task Score Correlations

	Argument				Issue			
Score from other task	W	G	F	C	W	G	F	C
Word choice (W)	**.52**	.22	.18	.35	**.52**	.16	.03	.23
Grammar (G)	.16	**.72**	.30	.48	.22	**.72**	.29	.54
Fluency (F)	.03	.29	**.67**	.25	.18	.30	**.67**	.30
Content (C)	.23	.54	.30	**.51**	.35	.48	.25	**.51**
e-rater	.19	.60	.63	.48	.31	.55	.59	.52
Human	.16	.53	.49	.46	.27	.48	.44	.48
Verbal	.31	.48	.40	.53	.37	.48	.32	.50
Quantitative	.18	.06	.32	.17	.17	.02	.20	−.03

Note. Figures above diagonal are for argument, below diagonal for issue, and on diagonal for cross-task reliabilities.

Table 16 Reliability-Based Weights — Subgroup Differences (Standardized Scores)

				Argument						Issue					
Group	%	Q	V	H	E	W	G	F	C	H	E	W	G	F	C
International	21	.64	−.41	−.48	−.42	.00	−.74	−.01	−.58	−.69	−.53	−.09	−.84	−.03	−.93
Asian	6	.43	.03	.05	.10	.19	.02	.11	.04	−.06	.02	.15	.02	−.00	−.10
Black	7	−.89	−.66	−.56	−.70	−.15	−.43	−.68	−.49	−.46	−.51	−.21	−.36	−.48	−.33
Hispanic	6	−.41	−.33	−.24	−.25	−.07	−.18	−.22	−.18	−.18	−.20	−.11	−.19	−.16	−.13
Female	62	−.20	−.09	−.01	.01	−.02	.05	−.00	.00	−.01	.01	−.06	.07	−.01	.00

Note. Results for ethnicity and gender are limited to domestic examinees. Q = quantitative; V = verbal; H = human; E = e-rater; W = word choice; G = grammar; F = fluency; C = content.

estimates and the trait score correlation matrix (Wainer et al., 2000). These parameters result (using formulae derived in Wainer et al., 2000) in a set of weights placed on the different observed trait scores. Augmented trait scores are computed as the sum of the products of these weights and observed trait scores.

Table 20 shows the weights placed on the different observed trait scores in the computation of the augmented trait scores for the two tasks. For example, the numbers in the first column of the table denote that in the computation of the augmented word-choice score for the argument task, a weight of .44 was placed on the observed word-choice score and smaller weights were placed on the other three observed trait scores. In other words, for the argument task,[3]

$$\text{Augmented word choice score} = .44 \times \text{Observed word choice score} - .01 \times \text{Observed grammar score} - .03 \times \text{Observed fluency score} + .23 \times \text{Observed content score}.$$

The weight of the corresponding observed score in any column is marked in bold. Table 20 shows that in the computation of any augmented trait score, the corresponding observed trait score receives the largest weight, except for the case of the augmented content scores for both tasks. Though the observed content score provides direct information about the true content score, the former has low reliability. Therefore, some of the other trait scores, which provide only indirect information about the true content score, receive a larger weight due to their high reliability.

Table 17 Factor Analysis (FA) – Based Weights — Within-Task Score Correlations and Cross-Task Reliabilities

Score	H	E	W	G	F	C
Human (H)	.56	.76	.20	.52	.63	.62
e-rater (E)	.77	.75	.27	.65	.84	.72
Word choice (W)	.24	.28	.51	.18	.09	.36
Grammar (G)	.58	.71	.19	.71	.38	.51
Fluency (F)	.61	.81	.04	.31	.66	.41
Content (C)	.59	.61	.33	.59	.36	.50

Note. Figures above diagonal are for argument, below diagonal for issue, and on diagonal for cross-task reliabilities.

Table 18 Factor Analysis (FA)-Based Weights — Cross-Task Score Correlations

	Argument				Issue			
Score from other task	W	G	F	C	W	G	F	C
Word choice (W)	**.51**	.24	.19	.36	**.51**	.18	.04	.20
Grammar (G)	.18	**.71**	.30	.49	.24	**.71**	.30	.54
Fluency (F)	.04	.30	**.66**	.27	.19	.30	**.66**	.33
Content (C)	.20	.54	.33	**.50**	.36	.49	.27	**.50**
e-rater	.19	.60	.62	.49	.32	.55	.60	.53
Human	.16	.54	.48	.47	.28	.49	.44	.48
Verbal	.31	.49	.40	.53	.38	.49	.32	.48
Quantitative	.18	.06	.31	.18	.17	.02	.20	−.03

Note. Boldface figures are cross-task reliabilities and expected to be highest in column. Underlined figures contradict this expectation.

Table 19 Factor Analysis (FA)-Based Weights — Subgroup Differences (Standardized Scores)

				Argument						Issue					
Group	%	Q	V	H	E	W	G	F	C	H	E	W	G	F	C
International	21	.64	−.41	−.48	−.42	.00	−.74	−.02	−.57	−.69	−.53	−.10	−.84	−.06	−.91
Asian	6	.43	.03	.05	.10	.19	.01	.09	.05	−.06	.02	.15	.01	−.01	−.11
Black	7	−.89	−.66	−.56	−.70	−.15	−.44	−.66	−.50	−.46	−.51	−.21	−.36	−.47	−.34
Hispanic	6	−.41	−.33	−.24	−.25	−.07	−.18	−.22	−.19	−.18	−.20	−.11	−.19	−.15	−.14
Female	62	−.20	−.09	−.01	.01	−.02	.05	−.01	.00	−.01	.01	−.07	.06	−.01	.03

Note. Results for ethnicity and gender are limited to domestic examinees. Q = quantitative; V = verbal; H = human; E = e-rater; W = word choice; G = grammar; F = fluency; C = content.

Table 21 shows the correlations between observed and augmented trait scores. The very high correlations for grammar and fluency mostly reflect the higher reliabilities of these scores (and to some extent, modest correlations with other scores), whereas the relatively low correlation for content reflects both the lower reliability and higher correlations with other trait scores, which allows one to borrow more information for the content score.

Tables 22 – 24 present the same analyses with previous trait scores. The cross-task reliabilities increase by .01 to .04 for the first three scores but increase dramatically (from .51 [Table 14] to .76 [Table 22]) for content. Naturally, all within-task correlations are higher. High within-task correlations of around .9 can be observed between some of the scores. An examination of Table 23 and comparison to Table 15 shows that the added value of the content score is increased with the use of augmented scores, but the situation is less clear for other scores. For these other three scores, the correlation between a trait score and e-rater score increases more than the reliability of the trait score, slightly reducing their added value.

Examination of Table 24 shows a peculiar situation with respect to subgroup differences. In several cases, subgroup standardized scores are markedly different than the corresponding scores in Table 16, especially for word use and content. For example, word use scores for international examinees are close to average for regular scores (.00 and −.09) but are substantially lower for augmented scores (−.23 and −.44). In other cases, only the scores of one task are different. For

Table 20 Weights for Computation of Augmented Trait Scores

	Argument augmented score				Issue augmented score			
Observed score	W	G	F	C	W	G	F	C
Word choice (W)	**.44**	−.01	−.04	.22	**.43**	−.04	−.05	.25
Grammar (G)	−.01	**.61**	.08	.17	−.02	**.57**	.05	.25
Fluency (F)	−.03	.10	**.59**	.12	−.03	.06	**.62**	.09
Content (C)	.23	.29	.18	**.21**	.25	.43	.14	**.12**

Note. The figures in bold indicate the weight of the corresponding observed score in any column.

Table 21 Correlations of Observed and Augmented Trait Scores

Score	Argument	Issue
Word choice	.93	.93
Grammar	.96	.96
Fluency	.98	.99
Content	.80	.76

Table 22 Augmented Reliability-Based Weights—Within-Task Score Correlations and Cross-Task Reliabilities

Score	H	E	W	G	F	C
Human (H)	.56	.76	.35	.64	.70	.71
e-rater (E)	.77	.75	.42	.80	.91	.89
Word choice (W)	.35	.35	.55	.40	.22	.69
Grammar (G)	.67	.79	.43	.76	.67	.89
Fluency (F)	.66	.85	.09	.53	.68	.78
Content (C)	.70	.85	.67	.92	.59	.76

Note. Figures above diagonal are for argument, below diagonal for issue, and on diagonal for cross-task reliabilities.

example, issue augmented content scores for international examinees are .09 higher than the corresponding regular scores, but issue augmented scores are .19 higher.

We also computed the proportional reduction in mean squared errors (PRMSEs) of the augmented trait scores. Table 25 shows the reliabilities of the observed trait scores and the PRMSEs of the augmented trait scores. Haberman (2008) stated that a necessary condition for an augmented subscore to have added value is that its PRMSE is substantially higher than the reliability of the corresponding observed subscore. Table 25 shows that all PRMSEs are higher than their corresponding reliabilities with especially large differences for content and very small differences for fluency.

Discussion

Test takers are very interested in receiving additional information on their performance beyond the total test score. Subscores of meaningful aspects of test performance are seen as valuable aids in interpreting test performance. However, subscores are often highly correlated with other subscores, rendering them less useful from a psychometric perspective. In addition, in the context of essay writing assessments, reporting of subscores based on human analytic scoring rubrics can be very costly.

This paper extends an approach for reporting essay trait scores that is based on the e-rater automated essay scoring system. Previous analyses showed support for a three-factor structure of the noncontent features of e-rater. This paper extends these results to the two GRE writing tasks and shows that the content features measure a fourth separate factor. Altogether, the four factors measured by e-rater cover the major constructs that are identified in many writing rubrics, including those for the GRE.

There are certainly differences in the degree that each factor can be interpreted as a sound measure of the construct it is intended to represent. Although the grammar, usage, mechanics, and collocation/preposition use features can be thought of as directly measuring conformity to conventions, other factors are measured in a less straightforward way. For example, content measures in e-rater (and other automated essay scoring systems) are based on the similarity of the

Table 23 Augmented Reliability-Based Weights—Cross-Task Score Correlations

	Argument				Issue			
Score from other task	W	G	F	C	W	G	F	C
Word choice (W)	**.55**	.40	.25	.52	**.55**	.33	.10	.46
Grammar (G)	.33	**.76**	.48	.68	.40	**.76**	.47	.76
Fluency (F)	.10	.47	**.68**	.52	.25	.48	**.68**	.57
Content (C)	.46	.76	.57	**.76**	.52	.68	.52	**.76**
e-rater	.28	.68	.68	.69	.38	.62	.64	.70
Human	.27	.60	.55	.60	.35	.55	.49	.59
Verbal	.18	.13	.31	.25	.10	.02	.18	.12
Quantitative	.42	.56	.47	.62	.45	.54	.37	.61

Note. Boldface figures are cross-task reliabilities and expected to be highest in column. The underlined figure contradicts this expectation.

Table 24 Augmented Reliability-Based Weights—Subgroup Differences (Standardized Scores)

				Argument						Issue					
Group	%	Q	V	H	E	W	G	F	C	H	E	W	G	F	C
International	21	.64	−.41	−.48	−.42	−.23	−.72	−.18	−.50	−.69	−.53	−.44	−.93	−.20	−.73
Asian	6	.43	.03	.05	.10	.15	.05	.10	.13	−.06	.02	.08	−.02	−.02	.06
Black	7	−.89	−.66	−.56	−.70	−.24	−.58	−.71	−.64	−.46	−.51	−.22	−.44	−.50	−.50
Hispanic	6	−.41	−.33	−.24	−.25	−.11	−.22	−.24	−.24	−.18	−.20	−.11	−.21	−.17	−.23
Female	62	−.20	−.09	−.01	.01	−.02	.04	.00	.01	−.01	.01	−.05	.05	−.00	.01

Note. Results for ethnicity and gender are limited to domestic examinees. Q = quantitative; V = verbal; H = human; E = e-rater; W = word choice; G = grammar; F = fluency; C = content.

Table 25 Proportional Reductions in Mean Squared Errors (PRMSEs) of Augmented Trait Scores

	Argument		Issue	
Score	Observed reliability	Augmented PRMSE	Observed reliability	Augmented PRMSE
Word choice	.52	.59	.52	.60
Grammar	.72	.77	.72	.78
Fluency	.67	.71	.67	.69
Content	.51	.80	.51	.89

(prompt-specific) vocabulary of an essay to the vocabulary used in high-scored versus low-scored essays. This measure of the ideas of an essay is crude, but for tasks that emphasize the importance of ideas (such as GRE argument and TOEFL integrated), these measures are strong predictors of human holistic scores. Similarly, the fluency and organization features in e-rater are crude measures of organization, development, and fluency. The fluency and organization factor was measured in this paper by the style feature and an essay length measure. The style feature measures sentence fluency and variety through sentence length and structure, and essay length measures writing fluency. An additional feature that was not included in the present analyses but loads on the same factor is the organization feature of e-rater. This feature measures organization by identifying discourse elements (such as introduction, thesis, main points, and conclusion) in the essay text.

Despite these limitations in representing the different aspects of the writing construct, the consistent factor analytic results show that e-rater is measuring important aspects of the construct across a wide range of tasks and populations. These results can be seen as providing convergent evidence for the construct validity of e-rater as an alternative method for scoring essays. Convergent evidence of this kind protects against the general threat of construct underrepresentation in interpretive meaning of scores (Messick, 1989).

The factor analytic results and their construct validity interpretation are supported by the results pertaining to the value of different trait scores based on these factors. In other words, the added value of the trait scores is evidence that

the factors are relatively independent and that they are measured relatively reliably. Factor intercorrelations are especially low between the word choice factor and other factors, and are higher among the three other factors (50s to 70s). Trait score reliabilities are lower for word choice and content (around .5) and higher for conventions and fluency (around .7). These results correctly predict that the content trait score will have lower added value than other trait scores because the content trait score is less reliable and more highly correlated with other trait scores. However, a third factor that affects added value of trait scores is the importance of the trait score in e-rater scores. Because the features of the fluency factor have higher relative weights in e-rater, the correlation of e-rater scores with the fluency trait score is higher, and the added value of the fluency trait score is thus reduced.

This paper explored another issue with implications for the validity of scores. The criterion for determining the importance of features in automated essay scoring applications has traditionally been focused on optimal prediction of human essay scores. Although the rationale for this criterion (optimal prediction) has been criticized on the grounds that it does not contribute to the validity and defensibility of automated scores (Ben-Simon & Bennett, 2007), performance issues have dominated the choice of criterion. This paper showed that a broader conception of performance, one that looks beyond a single essay, can change perception in this matter. The trait scores that were based on internal criteria (reliability or factor analytic results) had slightly higher value than those based on prediction of human scores, apart from possessing more homogeneous sets of feature weights.

Alternative criteria for trait score definition had a relatively minor effect on their value. The main course of action for increasing the value of the trait scores would be to increase their measurement reliability by developing new and improved features. This paper explored the use of augmented trait scores as a statistical method for improving reliability. The augmented trait scores showed higher reliabilities than their nonaugmented counterparts, and the difference was large for content. However, possible limitations of augmented trait scores are that (a) they are difficult to explain to test score users, (b) the correlations between them and other augmented trait scores are high, and (c) the subgroup differences of augmented trait scores may show different patterns than their nonaugmented counterparts.

Acknowledgments

Dr. Sinharay participated in the conduct of this study while on staff at ETS. He is currently at Pacific Metrics Corporation.

Notes

1 A companion paper (Attali & Sinharay, 2015) explored the same issues with the two TOEFL prompts.
2 Note that in these analyses we are estimating reliability of a score consisting of only two items. These values of reliability can be unstable, especially because the variances of the scores on the items may differ occasionally. However, given the design of the test, we have no better way to estimate reliabilities.
3 Note that the observed trait scores were standardized before this computation. Therefore, there is no intercept in this equation.

References

Attali, Y. (2007). *Construct validity of e-rater in scoring TOEFL essays* (Research Report No. RR-07-21). Princeton, NJ: Educational Testing Service.

Attali, Y. (2011). *Automated subscores for TOEFL iBT independent essays* (Research Report No. RR-11-39). Princeton, NJ: Educational Testing Service.

Attali, Y., & Burstein, J. (2006). Automated essay scoring with e-rater V.2. *Journal of Technology, Learning, and Assessment, 4*(3). Retrieved from http://ejournals.bc.edu/ojs/index.php/jtla

Attali, Y., & Powers, D. (2008). *A developmental writing scale* (Research Report No. RR-08-19). Princeton, NJ: Educational Testing Service.

Attali, Y., & Sinharay, S. (2015). *Automated trait scores for TOEFL writing tasks* (Research Report No. RR-15-14). Princeton, NJ: Educational Testing Service.

Bacha, N. (2001). Writing evaluation: What can analytic versus holistic essay scoring tell us? *System*, *29*, 371 – 383.

Ben-Simon, A., & Bennett, R. E. (2007). Toward more substantively meaningful automated essay scoring. *Journal of Technology, Learning, and Assessment*, *6*(1). Retrieved from http://ejournals.bc.edu/ojs/index.php/jtla

Education Northwest. (2011). *6+1 trait writing*. Retrieved from http://educationnorthwest.org/traits

Educational Testing Service. (2011). *GRE scoring guide for the issue task*. Retrieved from http://www.ets.org/gre/revised_general/prepare/analytical_writing/issue/scoring_guide

Freedman, S. W. (1984). The registers of student and professional expository writing. Influences on teacher responses. In R. Beach, & S. Bridwell (Eds.), *New directions in composition research* (pp. 334–347). New York, NY: Guilford Press.

Grandy, J. (1992). *Construct validity study of the NTE Core Battery using confirmatory factor analysis* (Research Report No. RR-92-03). Princeton, NJ: Educational Testing Service.

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

Hamp-Lyons, L. (1991). Scoring procedures for ESL contexts. In L. Hamp-Lyons (Ed.), *Assessing second language writing in academic contexts* (pp. 241–276). Norwood, NJ: Ablex.

Hamp-Lyons, L. (1995). Rating nonnative writing: The trouble with holistic scoring. *TESOL Quarterly*, *29*, 759–762.

Hoyle, R. H., & Panter, A. T. (1995). Writing about structural equation models. In R. H. Hoyle (Ed.), *Structural equation modeling: Concepts, issues, and applications* (pp. 158–176). Thousand Oaks, CA: Sage Publications.

Huot, B. (1990). The literature of direct writing assessment: Major concerns and prevailing trends. *Review of Educational Research*, *60*, 237–263.

Jöreskog, K. G., & Sörbom, D. (2006). *LISREL 8.80 for Windows [Computer software]*. Lincolnwood, IL: Scientific Software International, Inc..

Kelley, T. L. (1927). *Interpretation of educational measurements*. Yonkers, NY: World.

Lee, Y., Gentile, C., & Kantor, R. (2008). *Analytic scoring of TOEFL CBT essays: Scores from humans and e-rater* (Research Report No. RR-08-01). Princeton, NJ: Educational Testing Service.

Li, H., Rosenthal, R., & Rubin, D. B. (1996). Reliability of measurement in psychology: From Spearman-Brown to maximal reliability. *Psychological Methods*, *1*, 98–107.

Messick, S. (1989). Validity. In R. L. Linn (Ed.), *Educational measurement* (3rd. ed., pp. 13–103). New York, NY: Macmillan.

Roid, G. H. (1994). Patterns of writing skills derived from cluster analysis of direct writing assessments. *Applied Measurement in Education*, *7*(2), 159–170.

Sinharay, S. (2010). When can subscores be expected to have added value? Results from operational and simulated data. *Journal of Educational Measurement*, *47*, 150–174.

Sinharay, S. (2013). A note on added value of subscores. *Educational Measurement: Issues and Practice*, *32*(4), 38–42.

Swartz, C. W., Hooper, S. R., Montgomery, J. W., Wakely, M. B., De Kruif, R. L., Reed, M., et al. (1999). Using generalizability theory to estimate the reliability of writing scores derived from holistic and analytic scoring methods. *Educational and Psychological Measurement*, *59*(3), 492–506.

Veal, L. R., & Hudson, S. A. (1983). Direct and indirect measures for large scale evaluation of writing. *Research in the Teaching of English*, *17*, 285–296.

Wainer, H., Sheehan, K., & Wang, X. (2000). Some paths toward making Praxis scores more useful. *Journal of Educational Measurement*, *37*, 113–140.

Weigle, S. C. (2002). *Assessing writing*. New York, NY: Cambridge University Press.

Suggested citation:

Attali, Y., & Sinharay, S. (2015). *Automated trait scores for GRE® writing tasks* (Research Report No. RR-15-15). Princeton, NJ: Educational Testing Service. http://dx.doi.org/10.1002/ets2.12062

Action Editor: Daniel Eignor

Reviewers: Shelby Haberman and Donald Powers