Listening. Learning. Leading.®

Research Report
ETS RR–11-39

Automated Subscores for TOEFL iBT® Independent Essays

Yigal Attali

October 2011

Automated Subscores for TOEFL iBT[®] Independent Essays

Yigal Attali

ETS, Princeton, New Jersey

October 2011

Technical Review Editor: Dan Eignor

Technical Reviewers: Jill Burstein and Sandip Sinharay

Abstract

The e-rater® automated essay scoring system is used operationally in the scoring of TOEFL iBT® independent essays. Previous research has found support for a 3-factor structure of the e-rater features. This 3-factor structure has an attractive hierarchical linguistic interpretation with a word choice factor, a grammatical convention within a sentence factor, and a fluency factor. The purpose of this study was to explore the feasibility of developing automated subscores for the TOEFL iBT independent task based on this 3-factor structure. First, using a multiple-group confirmatory factor analysis, the 3-factor structure was found to be quite stable across major language groups. Next, the subscores based on these 3 factors were found to have added value in the context of repeater examinees by comparing the ability to predict a subscore on one test from the subscore on the other test or from the total e-rater score on the other test. The results of this study could be used to develop and report subscores for the performance of examinees on TOEFL iBT independent prompts. Reporting subscores could have instructional and remedial value, both at the individual and institutional level.

Key words: automated essay scoring, e-rater, subscores, TOEFL®

For performance assessments in general, and essay writing assessments in particular, the implementation of subscores usually implies the development of analytic (multi-trait) scoring rubrics that can be useful for capturing examinees' specific weaknesses and strengths in writing (Weigle, 2002). This can be especially valuable for second language learners who are still developing their writing skills and who are thus likely to show uneven profiles across different aspects of writing. For this reason, 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).

Analytic rubrics for essays written by English as a second language (ESL) students define different sets of dimensions. The one developed by Jacobs and her colleagues (Jacobs, Zingraf, Wormuth, Hartfiel, & Hughey, 1981) includes five rating dimensions of writing quality: content, organization, vocabulary, language use, and mechanics. The Test in English for Educational Purposes (TEEP; Weir, 1990) consists of seven 4-point scales that cover four aspects of communicative effectiveness (relevance and adequacy of content, compositional organization, cohesion, and adequacy of vocabulary for purpose) and three accuracy dimensions (grammar, mechanical accuracy/punctuation, and mechanical accuracy/spelling). The Michigan Writing Assessment Scoring Guide (Hamp-Lyons, 1991) contains three 6-point scales: ideas and arguments, rhetorical features, and language control. The analytic rubrics developed by Gentile, Riazantseva, and Cline (2002) for a research study include six rating scales that cover five dimensions: development, organization, vocabulary, language use, and mechanics. The language use dimension is further divided into two subdimensions of sentence variety/construction and grammar/usage accuracy.

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 total (holistic) essay 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 subscores for TOEFL iBT® independent essays that are also viable from a psychometric point of view. The e-rater® V.2 scoring engine (Attali & Burstein, 2006) 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 all of the features are indicators of generally acknowledged dimensions of good 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 successfully applied on data from several essay prompts of the same assessment. This means that a single scoring model is developed for a writing assessment (a generic model), consistent with the human rubric that is usually the same for all assessment prompts.

The feature set used with e-rater V.2 includes measures of grammar, usage, mechanics, style, organization, development, vocabulary, and word length. All these measures are related to the form and structure of the essay. In addition, it is possible to use prompt-specific vocabulary usage measures. In order to produce an essay score, the feature values need to be combined. In e-rater V.2 this is accomplished by standardizing the feature values, followed by calculating a weighted average of the standardized feature values, and finally applying a linear transformation to achieve a desired scale (usually formed by matching some human scoring scale).

One of the main possible advantages of e-rater V.2 for TOEFL iBT is the possibility of implementing a single scoring model (generic model). Such a model is trained on a large set of essays written on multiple prompts and is implemented on new prompts without additional training. Generic models score essays written to different prompts using the same scoring standards, a psychometric advantage for a standardized test. A generic model is particularly useful for TOEFL iBT because of the large number of forms administered (around 50 a year) and the fast required turnover of scores for reporting purposes. In a recent large-scale evaluation of e-rater for the TOEFL iBT independent task (Attali, in press), the test-retest reliability of the generic e-rater scores (.80) was significantly higher than for a single human rater (.53), and the true-score correlation estimate between human and e-rater scores was very high (.95). In addition, the correlations of the e-rater scores with the other TOEFL subscores were similar to the human score correlations.

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 structure of the e-rater features.

This 3-factor structure has an attractive hierarchical linguistic interpretation with a word choice factor, a grammatical conventions within a sentence factor, and a fluency factor. 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, and the pattern of item-factor relationships (which items load on which factors) indicates how the subscores should be scored. Because of its meaningful interpretation, the 3-factor structure can serve as a basis for automatically produced subscores with a sound psychometric foundation.

The literature includes several approaches for the computation and evaluation of subscores (Dwyer, Boughton, Yao, Steffen, & Lewis, 2006; Wainer, Sheehan, & Wang, 2000; Yen, 1987). Haberman (2008) 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. If this condition is not satisfied, then instructional or remedial decisions based on the subscore will lead to more errors than those based on total scores.

The first goal of this paper is to determine whether the 3-factor structure is a feasible underlying structure for TOEFL iBT independent essays and, in particular, whether this structure holds across the major language groups that take the test. The second goal is to determine whether subscores based on the 3-factor structure have added value over total scores by testing a variant of Haberman’s criterion: does a subscore in the first test better predict the same subscore in the second test than the total score (in the first test)?

Data for the study came from a recent evaluation of e-rater with TOEFL iBT independent essays (Attali, in press). For this study, an entire year of essays was analyzed and e-rater essay scores were developed based on a generic model (that is, a single e-rater model across prompts).

Underlying Structure Across Language Groups

Method

Data. TOEFL iBT was gradually introduced worldwide from September 2005 to October 2006. The last administration of the TOEFL CBT was in September 2006. The analyses in this report include all computer-based (iBT) tests administered worldwide from the beginning of October 2006 until the middle of May 2007. The total number of test records was 205,566. In this period, 26 independent prompts were administered, ranging in the number of test takers from around 3,900 to around 15,000.

For each test taker, several variables were available for analysis, among them all TOEFL test scores; test takers' answers to the biographical questionnaire, including native language; and other background information such as the country of the test center.

Overall, 132 different native languages were reported by examinees. The 10 most popular self-reported native languages, accounting for 72% of examinees, were Korean (16.5%), Chinese (15.2%), Japanese (9.4%), Spanish (9.1%), Arabic (7.0%), German (3.6%), French (3.2%), Telugu (2.9%), Hindi (2.6%), and Turkish (2.4%).

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 organization and development features of e-rater V.2 (Attali & Burstein, 2006) because of the very high combined multiple-correlation of these two features with essay length. Table 1 also lists the linguistic level of each feature, with vocabulary and word length in the word choice first level; grammar, usage, and mechanics in the conventions second level; and essay length and style in the fluency third level.

Table 1

Features Used in the Present Study

Feature (linguistic level)	Description
Essay length (3)	Based on number of words in essay
Style (3)	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
Grammar (2)	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 (2)	Based on rates of errors such as wrong or missing articles, confused words, wrong form of words, faulty comparisons, and preposition errors
Mechanics (2)	Based on rates of spelling, capitalization, and punctuation errors
Vocabulary (1)	Based on frequencies of essay words in a large corpus of text
Word length (1)	Average word length

Language groups. Several major individual languages were analyzed separately: Korean, Chinese, Japanese, Arabic, and Turkish. In addition, in order to avoid the analysis of too many separate languages, several language groups were formed based on language families. The group of western European languages, based on Latin and Germanic languages, included Spanish (9% of essays), German (4%), French (3%), Italian (2%), Portuguese (2%), Romanian (.6%), Dutch (.3%), and Swedish (.3%). Other languages that belong to this group represented less than .1% of the data and were not included in this group. A second language group included Indian languages that belong to the Indo-Arian branch of the Indo-European family: Hindi (3%), Gujarati (2%), Urdu (1%), Bengali (.6%), Marathi (.5%), and Punjabi (.4%). A third language group was the Dravidian languages spoken in south India: Telugu (3%), Tamil (.8%), Malayalam (.5%), and Kannada (.3%). A fourth language group included Slavic languages: Russian (2%), Bulgarian (.6%), Polish (.5%), and Ukrainian (.3%). Table 2 summarizes the relative frequency of all languages and language groups analyzed below.

Table 2

Language Groups

Language group	Frequency (%)	Languages included
Latin-Germanic	21.1	Spanish, German, French, Italian, Portuguese, Romanian, Dutch, Swedish
Korean	16.5	Korean
Chinese	15.2	Chinese
Other languages	15.1	104 different languages
Japanese	9.4	Japanese
Arabic	7.0	Arabic
Indian (Indo-Arian family)	6.6	Hindi, Gujarati, Urdu, Bengali, Marathi, Punjabi
Indian (Dravidian family)	4.1	Telugu, Tamil, Malayalam, Kannada
Slavic	2.6	Russian, Bulgarian, Polish, Ukrainian
Turkish	2.4	Turkish

Alternative models. Four models were investigated, reflecting different degrees of separation in the three linguistic levels. The first model is a 1-factor model without any separation. The next two models are 2-factor models with partial separation. In one model, the two level-3 features, essay length and style, are separated from the other features. In the second model, the two level-1 features, vocabulary and word length, are separated from the other features. The fourth model is a 3-factor model with full separation of the three linguistic levels, with the grammar, usage, and mechanics features forming the second factor.

Results

Table 3 presents descriptive statistics about the seven features. In general, the features exhibit low levels of skewness and kurtosis, except for essay length with higher levels of positive skewness and negative kurtosis.

Table 4 presents the overall correlation matrix for the seven features used in this study. Correlations range from around 0 to .70.

The multiple-group confirmatory factor analyses were conducted for the 10 language groups. First, the alternative models presented above were tested to determine the number of factors that was supported in the data: one, two (with either level 1 or level 3 separated from the rest of the features), or three factors. Based on the model that was best supported, several nested models were employed to test the invariance of the factors across language groups. The nested models tested invariance in factor loadings, error variances, and factor correlations. Analyses were performed with LISREL 8.80 (Joreskog & Sorbom, 2006), based on the covariance matrices for each language group.

Several goodness-of-fit indices were used to evaluate model results (Hoyle & Panter, 1995). The comparative fit index (CFI), nonnormed fit index (NNFI), and root mean square error of approximation (RMSEA) were used for overall model fit. The standardized root mean square residual (SRMR) and goodness of fit index (GFI) were used for individual subsamples. The χ^2, χ^2/df, and the expected cross-validation index (ECVI) were used to compare overall and subsample models. Common rules of thumb were used in appraising the measures (Hoyle & Panter, 1995): .90 or more for CFI, NNFI, and GFI; .05 or less for RMSEA; .10 or less for SRMR; .05 alpha level for the χ^2, and 3 or less for χ^2/df.

Table 3

Feature Descriptive Statistics

Feature	M	SD	Skewness	Kurtosis
Essay length	5.71	0.29	7.28	-1.73
Style	-0.24	0.13	-0.19	0.19
Grammar	-0.11	0.04	0.40	-0.69
Usage	-0.11	0.04	-0.14	-0.47
Mechanics	-0.19	0.07	0.67	-0.73
Vocabulary	-62.48	2.87	-0.47	0.09
Word length	4.47	0.29	-0.02	0.08

Note. *N* = 203,804.

Table 4

Overall Feature Correlation Matrix

Feature	S	G	U	M	V	WL
Essay length	.40	.28	.22	.30	.09	-.02
Style (S)		.22	.16	.12	.29	.21
Grammar (G)			.34	.45	.26	.19
Usage (U)				.31	.02	-.03
Mechanics (M)					.22	.09
Vocabulary (V)						.70
Word length (WL)						

Note. *N* = 203,804.

Table 5 presents the overall fit indices for the four models. The overall fit for the 1-factor model and the two 2-factor models was unsatisfactory for all three indices (CFI, NNFI, and RMSEA). The overall fit for the 3-factor model was better, although it still showed low NNFI (.84) and high RMSEA (.11). In addition, Table 6 shows that the 3-factor solution had satisfactory separate-group fit for all indices (the separate-group fit for the previous models was unsatisfactory in all cases). In summary, only the 3-factor model showed reasonable fit.

A detailed view of the 3-factor model is presented in Table 7. The table shows that in all language groups the vocabulary feature has loadings larger than 1 (Heywood effects) and corresponding negative error variances. The table shows similar loadings, factor correlations, and error variances across language groups.

Table 5

Overall Tests of Invariance in Number of Factors

Model	df	χ^2	CFI	NNFI	RMSEA
Three factors	11	26,726	.92	.84	.11
Level 1 separated	13	42,506	.86	.77	.13
Level 3 separated	13	126,997	.69	.49	.22
One factor	14	161,208	.50	.25	.24

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

Table 6

Language Group Tests for Three-Factor Model

Language	df	χ^2	SRMR	GFI
Romance-Germanic	11	8,206	.08	.95
Korean	11	3,386	.05	.97
Chinese	11	4,483	.06	.96
Other	11	5,433	.07	.95
Japanese	11	2,813	.07	.96
Arabic	11	2,186	.06	.96
Indian (Indo-Arian family)	11	1,566	.06	.97
Indian (Dravidian family)	11	1,045	.06	.97
Slavic	11	1,142	.09	.94
Turkish	11	775	.07	.96

Note. GFI = goodness of fit index, SRMR = standardized root mean square residual.

It should be noted that both the Heywood cases and the fit of the model are improved when the correlations between several pairs of error variances are allowed to differ from 0—in particular, when the error correlations between the word level features and essay length are allowed to vary (with correlation estimates of around -.4) and the error correlations between the word level features and usage are allowed to vary (with correlation estimates of around -.15). For this model (df = 6, χ^2 = 3,266), the fit indices are all satisfactory (CFI = .99, NNFI = .97, RMSEA = .05) and the Heywood effect disappears.

Table 7

Three-Factor Model

	RG	KOR	CHI	OTH	JAP	ARA	IND	SIN	SLA	TUR
Feature loadings										
Essay length	.58	.81	.66	.69	.95	.77	.56	.55	.69	.66
Style	.51	.53	.59	.51	.46	.51	.70	.66	.46	.61
Grammar	.72	.71	.75	.64	.71	.66	.77	.67	.63	.70
Usage	.53	.46	.37	.49	.42	.51	.60	.50	.51	.47
Mechanics	.65	.66	.57	.63	.58	.73	.71	.62	.66	.64
Vocabulary	1.03	1.51	1.01	1.23	1.20	1.02	1.20	2.09	1.46	1.87
Word length	.73	.43	.67	.59	.53	.71	.57	.30	.50	.38
Factor correlations										
3 ↔ 2	.58	.63	.53	.60	.57	.59	.56	.56	.72	.59
2 ↔ 1	.32	.15	.36	.26	.14	.45	.32	.16	.24	.15
3 ↔ 1	.29	.11	.32	.19	.06	.23	.30	.15	.18	.13
Error variances										
Essay length	.67	.35	.56	.52	.11	.41	.69	.70	.53	.57
Style	.74	.72	.65	.74	.79	.74	.51	.56	.79	.62
Grammar	.48	.50	.44	.59	.49	.57	.40	.55	.60	.51
Usage	.72	.79	.86	.76	.83	.74	.64	.74	.74	.78
Mechanics	.57	.56	.68	.60	.67	.47	.50	.48	.57	.59
Vocabulary	-.07	-1.27	-.02	-.51	-.44	-.04	-.44	-3.35	-1.13	-2.48
Word length	.47	.82	.55	.66	.72	.49	.68	.91	.75	.86

Note. ARA = Arabic, CHI = Chinese, IND = Indian, JAP = Japanese, KOR = Korean, OTH = other, RG = Romance-Germanic, SIN = South Indian, SLA = Slavic, TUR = Turkish.

Table 8 presents fit indices for initial tests of invariance in the 3-factor model for the 10 language groups considered in a multigroup analysis. The three types of invariance, in factor loadings, factor correlations, and error variance, should be compared with the basic 3-factor solution results in the first row. These comparisons show that for all three types of invariance, the overall χ^2 and χ^2/df differences were statistically (all p values smaller than .001) and practically (all χ^2/df differences larger than 31) significant. An additional model fit indicator for comparing different models is the ECVI, which takes into account model complexity (as reflected by *df*) and sample size (smaller penalty for larger samples) in addition to model fit (as reflected by χ^2). The ECVIs for the three types of invariance (.19, .16, and .18) were similar to the ECVI of the basic model (.15), indicating that the basic 3-factor solution (without any of the invariance constraints) is expected to cross-validate only slightly better in a new sample than the solutions that presume invariance. In addition, the CFI, NNFI, and RMSEA values were similar or even better than the basic solution, and nearly all subgroup GFI and SRMR values (not shown in Table 8) were satisfactory.

When all three types of invariance are tested simultaneously (fifth row), both the CFI value is lower and the number of unsatisfactory SRMR values is higher. In this model a single set of factor correlations and feature loadings exist across all language groups. The three invariant correlations between factors were .60 between fluency and conventions, .27 between conventions and word choice, and .21 between fluency and word choice. The invariant feature loadings were .70, .54, .70, .48, .65, 1.15, and .60 for essay length, style, grammar, usage, mechanics, vocabulary, and word length, respectively.

Finally, when all three types of invariance are tested simultaneously with several correlated error variances (between essay length and the word level features, around -.4, and between usage and the word level features, around -.15), model fit is again improved. Both the CFI and NNFI values are satisfactory, the RMSEA is lower (.08), and most GFI and SRMR values are satisfactory. In this model, the three invariant correlations are .56 between fluency and conventions, .36 between conventions and word choice, and .60 between fluency and word choice. The feature loadings are .79, .48, .68, .51, .64, 1.02, and .68 for essay length, style, grammar, usage, mechanics, vocabulary, and word length, respectively.

Table 8

Tests of Invariance for the Three-Factor Model

Invariance	df	χ^2	CFI	NNFI	RMSEA	# GFI > .9	# SRMR < .1
None	110	30,846	.91	.83	.12	10	10
Factor loadings	173	38,328	.89	.87	.10	10	9
Factor correlations	137	31,969	.91	.86	.11	10	10
Error variances	173	36,471	.89	.87	.10	10	10
All	263	55,548	.84	.87	.10	9	4
All with correlated error variances	259	33,596	.91	.93	.08	10	6

Note. CFI = comparative fit index, GFI = goodness of fit index, NNFI = non-normed fit index, RMSEA = root mean square error of approximation, SRMR = standardized root mean square residual.

In summary, the overall fit for the 3-factor model was better than the 2-factor and 1-factor models, although it still showed low NNFI (.84) and high RMSEA (.11). The 3-factor solution also had satisfactory separate-group fit for all indices, as compared to unsatisfactory fit for the previous models. For the 3-factor model, statistical tests failed to show invariance in factor loadings, factor correlations, and error variance across the different language groups. However, model fit for invariant models was not much different than the basic noninvariant model. Finally, allowing several correlated error variances further improved model fit (both overall fit and cross-group fit).

Added Value of Subscores

Method

Data. All 151,128 records of examinees that were tested in the first 4 months of the global administration of TOEFL iBT (October 2006 to January 2007) were examined to find repeaters. In this period, 25 prompts were administered. A total of 13,899 examinees were found to repeat the test in that period. In the few rare cases that these repeaters took more than two tests in the period, the first two tests were included in the study sample.

E-rater scores. The same feature set that was used in the previous section was used in these analyses. The generic e-rater score was based on *optimal* weights derived from a regression analysis of the e-rater features on the first human essay score (excluding all cases with a 0 human score). For this evaluation a prompt-fold score was produced whereby, for each prompt, a model based on all other prompt data was developed and then applied only on the prompt data that was excluded from model development. In this way, scores for each essay were not based on essays written to the same essay prompt. However, it should be noted that the results of a simple model that is trained on the entire sample are almost identical to the results of the prompt-fold model. E-rater subscores were derived by standardizing all features and computing a weighted sum of the relevant features for each subscore, using the invariant feature loadings from the model with all three types of invariance tested simultaneously. For the word level subscore, 1.15 and .60 were used as weights for the vocabulary and word length features, respectively. For the grammatical conventions subscore, .70, .48, and .65 were used as weights for the grammar, usage, and mechanics features, respectively. For the fluency subscore, .70 and .54 were used as weights for the essay length and style features, respectively. Alternative weighting schemes, including equal weighting, were also explored.

Results

Table 9 shows that, as expected, repeaters achieved lower scores than the total group and improved their scores on the second test. However, these effects were not large. For example, the average first independent essay score of repeaters is .37 *SD*s below the total group average, and the average second score is .19 *SD*s below the overall average. The gains repeaters achieved in the different types of scores are roughly comparable, around .2 of their *SD*.

Table 10 presents the test-retest reliabilities and intercorrelations for different essay scores and other TOEFL subscores for repeaters. Included are human essay scores, both single and double; the e-rater generic score; the three e-rater subscores (word, grammar, and fluency); and the three TOEFL scores—reading, listening, and speaking—and the sum of these three scores (labeled RLS). Rows present correlations of a particular second test score with all first test scores. Columns present correlations of a particular first test score with all second test scores. Figures on the main diagonal represent test-retest reliabilities of a particular score.

The usefulness of a subscore is determined by whether predicting the subscore value at one test occasion from the same subscore at the other test occasion is more accurate than predicting it from another score at the other test occasion. In general, if the highest correlation value of a particular score on a row or a column is found on the diagonal, then prediction of that score is most accurate from the same score at the other test occasion. For example, prediction of a single human essay score at one test occasion from scores at the other occasion is more accurate from the e-rater score (.61), the double human score (.58), and the RLS score (.55) than from a single human score (.52).

Table 9

Descriptive Statistics for All Examinees and for Repeaters

	All		Repeaters 1st test		Repeaters 2nd test		Repeaters
Measure	*M*	*SD*	*M*	*SD*	*M*	*SD*	*d*
Independent essay	3.38	0.89	3.05	0.78	3.21	0.75	0.21
Integrated essay	3.01	1.22	2.57	1.07	2.85	1.09	0.26
Reading	19.49	8.13	16.74	7.74	18.46	7.63	0.22
Listening	20.35	7.29	17.96	6.96	18.93	6.90	0.14
Speaking	19.17	5.17	16.91	4.40	17.79	4.31	0.20

Note. All *N* = 151,128, Repeaters 1st test *N* = 13,899, Repeaters 2nd test *N* = 13,899.

Table 10

Intercorrelations and Test-Retest Reliabilities for Repeaters

	First test									
Second test	SH	DH	ER	W	G	F	R	L	S	RLS
Single human (SH)	.52	.58	.61	.26	.46	.46	.45	.47	.47	.55
Double human (DH)	.58	.64	.67	.28	.51	.52	.49	.52	.53	.60
e-rater (ER)	.61	.67	.79	.31	.56	.63	.48	.47	.49	.57
Word (W)	.24	.27	.30	.46	.23	.15	.36	.24	.14	.32
Grammar (G)	.45	.50	.55	.22	.66	.28	.39	.38	.31	.44
Fluency (F)	.47	.51	.62	.16	.28	.64	.33	.36	.46	.44
Reading (R)	.46	.50	.50	.37	.41	.35	.74	.62	.42	.74
Listening (L)	.48	.52	.49	.26	.38	.38	.59	.74	.60	.77
Speaking (S)	.51	.55	.52	.18	.34	.48	.41	.61	.82	.68
RLS	.56	.61	.58	.34	.45	.46	.72	.77	.67	.86

Note. RLS = the sum of the TOEFL reading, listening, and speaking scores.

For the three e-rater subscores, the table shows that they are most accurately predicted from the subscores themselves, with test-retest reliabilities of .46, .66, and .64, although the fluency subscore was predicted almost as well from the e-rater score. For the fluency subscore, alternative weighting schemes resulted in more significant differences in test-retest reliability. A higher weight for essay length resulted in higher reliability: .60 for equal weights versus .69 for a 70-30 scheme.

The results of the e-rater subscores can be compared to the usefulness of the TOEFL subscores. The reading subscore reliability is .74, and it can be predicted about equally well from the RLS score. The listening subscore reliability is also .74, but it can be better predicted from the RLS score (.77). The speaking subscore reliability is higher, .82, and it is best predicted from the speaking subscore.

Conclusions

Test takers are very keen on 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, development of subscores based on human analytic scoring rubrics is very costly. In addition, analytic scores for essay writing assessments are often highly correlated with other analytic scores, rendering them less useful from a psychometric perspective.

Recent advances in automated essay scoring provide an alternative way to develop subscores for TOEFL iBT independent essays. The reliability of e-rater generic scores for TOEFL iBT independent essays is high (.80 versus .53 for single human scores), and automated scores are highly correlated with human scores (.95 after correction for unreliability). In addition, factor analyses of e-rater features for essays written by both ESL (TOEFL CBT examinees) and native speakers of English from 4th to 12th grade revealed a stable 3-factor structure for its features, with an attractive hierarchical linguistic interpretation with a word choice factor, a grammatical conventions within a sentence factor, and a fluency factor. These factors can be used as the basis for three subscores for independent essays, providing added instructional and remedial value to TOEFL test users, both at the individual and institutional level.

The present study replicated previous factor analysis results and found support for the 3-factor structure across a wide range of native languages. In addition, the test-retest correlations of subscores based on the three factors were higher than the correlations between subscores and any other test score, including the total e-rater score, human essay scores, and other TOEFL iBT subscores. These results indicate that reporting e-rater subscores has a slight advantage over reporting the total e-rater score from a measurement perspective. It would be interesting to validate these results with other writing tasks, such as the TOEFL integrated task and the GRE argument and issue tasks.

References

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

Attali, Y. (in press). *E-rater evaluation for TOEFL iBT independent essays* (ETS Research Report). Princeton, NJ: ETS.

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

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

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

Dwyer, A., Boughton, K. A., Yao, L., Steffen, M., & Lewis, D. (2006). *A comparison of subscale score augmentation methods using empirical data.* Paper presented at the Annual Meeting of the National Council on Measurement in Education, San Francisco, CA.

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.

Gentile, C., Riazantseva, A., & Cline, F. (2002). *A comparison of handwritten and word processed TOEFL essays: Final report.* Unpublished manuscript, ETS, Princeton, NJ.

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

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.

Jacobs, H. L., Zingraf, S. A., Wormuth, D. R., Hartfiel, V. F., & Hughey, J. B. (1981). *Testing ESL composition: A practical approach*. Rowley, MA: Newbury House.Joreskog & Sorbom (2006).

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

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

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.

Weir, C. J. (1990). *Communicative language testing*. Englewood Cliffs, NJ: Prentice Hall Regents.

Yen, W. M. (1987, June). *A Bayesian/IRT index of objective performance*. Paper presented at the annual meeting of the Psychometric Society, Montreal, Quebec, Canada.