Listening. Learning. Leading.®

Research Report
ETS RR–11-36

A Differential Word Use Measure for Content Analysis in Automated Essay Scoring

Yigal Attali

August 2011

A Differential Word Use Measure for Content Analysis in Automated Essay Scoring

Yigal Attali

ETS, Princeton, New Jersey

August 2011

Technical Review Editor: Joel Tetreault

Technical Reviewers: Paul Deane and Derrick Higgins

Abstract

This paper proposes an alternative content measure for essay scoring, based on the *difference* in the relative frequency of a word in high-scored versus low-scored essays. The *differential word use* (DWU) measure is the average of these differences across all words in the essay. A positive value indicates the essay is using vocabulary more typical of high-scoring essays in this task, and vice versa. In addition to the traditional prompt level, this measure is also computed at the generic task level, where the content of an essay is evaluated in the context of the general vocabulary of the writing task (e.g., GRE® issue), *across* different prompts. Evaluation results across four GRE and TOEFL® tasks are presented. Factor analyses show that both the prompt and task DWU measures load on the same factor as the prompt-specific content analysis measures of e-rater®. Regression results on the human essay scores show that the generic task DWU measure is a strong predictor of human scores, second only to essay length among noncontent e-rater features. This measure provides a way to conduct a task-level content analysis that is related to the prompt-specific content analysis of e-rater but does not require prompt-specific training. Regression results for the prompt-level DWU show that it can be used as a replacement to prompt-specific content-vector analysis (CVA) features.

Key words: automated scoring, content analysis, writing assessment, e-rater

Overview

Content Features in e-rater®

In e-rater® V.2 (Attali & Burstein, 2006), prompt-specific content analysis is based on comparing the lexical content of an essay to the lexical content of samples of essays with known score levels that were written to the same prompt. This is accomplished through content vector analysis (Salton, Wong, & Yang, 1975), where the vocabulary of essays, or sets of essays, is converted to a vector whose elements are based on the frequency of each word. More specifically, the vocabulary of sets of training essays from each score point (usually the score scale includes six points) is converted to vectors whose elements are weights for each word in the set of training essays (some function words are removed prior to vector construction). For each of the score categories, the weight for word *i* in score category *s* is:

$$W_{is} = \frac{F_{is}}{MaxF_s} \log(\frac{N}{N_i})$$

where F_{is} is the frequency of word *i* in score category *s*, $MaxF_s$ is the maximum frequency of any word at score point *s*, *N* is the total number of essays in the training set, and N_i is the total number of essays having word *i* in all score points in the training set.

Then, for each individual essay, a similar vector is produced from the vocabulary of the essay, where the weight for word *i* in the essay is:

$$W_i = \frac{F_i}{MaxF} \log(\frac{N}{N_i})$$

where F_i is the frequency of word *i* in the essay, and *MaxF* is the maximum frequency of any word in the essay.

Finally, for each essay, six cosine correlations are computed between the vector of word weights for that essay and the word weight vectors for each score point. These six cosine values indicate the degree of similarity between the words used in an essay and the words used in essays from each score point.

In Attali and Burstein (2006), two content analysis features were defined from these six cosine correlations and used in e-rater V.2. The first is the *score point value* (1-6) for which the maximum cosine correlation over the six score point correlations was obtained (referred to as *max. cos.*). This feature indicates the score point level to which the essay text is most similar with regard to vocabulary usage. The second is the *cosine correlation value* between the essay vocabulary and the sample essays at the highest score point, in many cases 6 (referred to as *cos. w/6*). This feature indicates how similar the essay vocabulary is to the vocabulary of the best essays. Together these two features provide a measure of the level of prompt-specific vocabulary used in the essay.

Modifications to Content-Vector Analysis Features

Recently, Attali (2011) modified the definition of the two features. The rationale of the max. cos. feature is that the score point for which the highest cosine correlation is obtained should be a predictor of human essay score. However, if the score point with the highest cosine correlation is positively correlated with human essay scores, so should the score point with the *second* highest cosine correlation. Additionally, the score point with the *lowest* cosine correlation should be *negatively* correlated with human essay scores. In other words, the complete ranking of the cosine correlations should provide more complete information on the pattern of *k* cosine correlations (*k* is the number of score points). The *pattern cosine* feature quantifies these rankings by summing the product of score points (S_i) and rankings (R_i):

$$Pat.\,Cos. = \sum_{i}^{k} S_i R_i$$

This score can be transformed to conform to the original scale of 1 to *k*.

The rationale to the cos. w/6 feature is that the *value* of the cosine correlation with the highest score point should be a predictor of human essay score. The modification to this feature was similar to the one proposed to max. cos. Instead of using only the information for the highest score point, a weighted sum of all cosine correlation values is used as a predictor of human essay scores, with a +1 weight for the three highest score points and a -1 weight for the lowest three score points:

$$\text{Val. Cos.} = 1\ \text{Cos6} + 1\ \text{Cos5} + 1\ \text{Cos4} - 1\ \text{Cos3} - 1\ \text{Cos2} - 1\ \text{Cos1}$$

Performance of Content-Vector Analysis Features

The original content-vector analysis (CVA) features had a modest contribution to e-rater performance. In Attali and Burstein (2006), regression analyses of the e-rater features on human scores across several assessment programs (including 6th to 12th grade writing assessments, Graduate Management Admission Test [GMAT], and TOEFL® prompts) were performed. The average relative weights (relative contribution of standardized regression weights) of the two CVA features was 15%. In Attali, Bridgeman, and Trapani (2010), the combined relative weights of the two CVA features for the GRE® issue, GRE argument, and TOEFL independent tasks were 11%, 20%, and 6%, respectively. The inclusion of the two CVA features improved the correlation of e-rater scores with human scores by .01 (for GRE issue and TOEFL independent) and .03 (for GRE argument). Finally, in the evaluation work of the modified CVA features (Attali, 2011), performance of original and modified CVA features was evaluated on datasets for the GRE issue and argument tasks and TOEFL independent and integrated tasks. The original CVA features increased the correlation of e-rater scores with human scores by .02, .04, .01, and .12, respectively. The modified features further increased correlations by .003, .010, .002, and .028, respectively.

These results show that the prompt-specific content analysis is less important for the GRE issue and TOEFL independent tasks, which emphasize writing quality and present prompts that are less constrained by specific content. Content analysis is more important for the GRE argument task, and even more so for the TOEFL integrated task, which presents a more detailed prompt and requires the student to write a critique that is more intimately related to the prompt's content.

Differential Word Use Measure

This paper introduces an alternative approach to CVA. It is based on a comparison of the relative frequency of a word in essays of high quality versus essays of low quality. This approach is based on the assumption that the use of a word that appears more frequently in high-quality essays than in low-quality essays should be an indicator of better vocabulary.

More technically, the measure is computed by first developing word indices for their differential frequency in high- and low-quality essays. This is accomplished by counting, for each word (indexed i), occurrences in a large set of high- and low-scored (f_{il} and f_{ih}) essays, and computing the differences of log-transformed relative frequencies of the words:

$$d_i = \log(f_{ih} / f_{\bullet h}) - \log(f_{il} / f_{\bullet l})$$

A d_i value of zero indicates that a word is equally likely to appear in a low- or high-scored essay. For an individual essay, the *differential word use* (DWU) measure is computed by averaging the d_i values over all the words in the essay.

Several differences can be noted between the CVA approach and DWU. First, CVA uses the so-called *tf–idf* (term frequency–inverse document frequency) weighting for individual words. That is, the importance of a word (term) is related to its frequency in a category, but inversely related to its overall document frequency in all categories. Conversely, d_i values do not explicitly take into account term frequency. Second, in CVA, indices for individual words are combined by computing a cosine correlation between all words in an essay and in a category. On the other hand, the DWU measure is a simple average of all d_i values in an essay. Finally, with CVA it is not clear how to use the six resulting cosine correlations and, as noted above, several methods have been used. With the proposed approach, the average of all d_i values is the final measure.

Note that d_i values are closely related to the logarithm of the odds ratio (OR). In the OR, the two denominators of d_i are replaced by the frequencies that the word does *not* appear in the (high or low) category. In the present context, frequencies of appearance of words in essays are very low relative to the frequencies of absence. Therefore, d_i values are virtually identical to logarithms of OR (and results obtained with either one of them are identical). There is a slight advantage of ease of interpretation for d_i values over OR: Odds are less intuitive than simple rates. On the other hand, OR have important advantages from a statistical theory perspective. For example, Fleiss (1981) discusses different ways to combine multiple OR. A simple alternative is based on a weighted average of logs OR, where the weights are the inverse of the standard errors of the log OR. These weights are highly related to the overall frequency of occurrence. Therefore, this alternative differs from the above proposal for DWU by taking into account word frequency. However, a comparison of weighted versus unweighted DWU shows a clear performance advantage for the unweighted version.

Rationale for a Task-Level Content Analysis

Content analysis in the context of automated essay scoring applications has always been conceptualized as prompt-specific. Similarly to e-rater's CVA analysis, the Intelligent Essay

Assessor (IEA; Landauer, Laham, & Foltz, 2003) by Pearson Knowledge Technologies uses latent semantic analysis (Landauer, Foltz, & Laham, 1998), a dimensionality-reduction method to represent the content of the essay as a vector in multidimensional space. The content score is based on the proximity of the vector to vectors of sets of prescored essays in the same dimensionality space.

This paper proposes a generic task-level content analysis, where the content of an essay is evaluated in the context of the general vocabulary of the writing task, *across* different prompts. It is reasonable to assume that specific tasks elicit vocabulary that has commonalities across specific prompts. Furthermore, parts of this task lexicon may be more typical of higher or lower quality writing. Characterizing an essay's vocabulary in this respect may identify aspects of the essay's content beyond the prompt-specific analysis and improve automated scoring.

A simple approach for testing the usefulness of a task-level content analysis is to apply the CVA analysis on a cross-prompt training set. That is, create score point word vectors based on essays from different prompts. Unfortunately, the cosine correlations computed at this level are not related to human scores of essays.

The alternative, explored in this paper, is to compute the DWU measure both at the prompt level and at the task level. In other words, in addition to a traditional prompt-specific content analysis based on DWU (PDWU), a task-level DWU measure (TDWU) will also be evaluated.

Vocabulary Measure Interpretation

Although the TDWU measure was conceived as a generic content measure, it can also be interpreted as a vocabulary measure. The current e-rater vocabulary measure is based on (log transformed) word frequencies from a large collection of published printed media of more than 500 million words. The vocabulary measure computes the median frequency (from the list) of the words used in an individual essay. Essays associated with less frequent words (i.e., their median word frequency is lower) are deemed to possess better vocabulary and are predicted to be of higher quality.

The proposed TDWU measure has an advantage over the current vocabulary measure because it is based on actual essays written by students and creates word indices based on actual evaluations of the essays these words were embedded in. On the other hand, it is not clear why a word that appears less frequently in published books should be associated with better vocabulary in a typical persuasive essay. For example, the word *brainy* is much less frequent than *smart*, but it does not seem as an indicator of better vocabulary.

Evaluation Setup

The DWU measures were evaluated on the GRE argument and issue tasks and on the TOEFL independent and integrated tasks. For three tasks (GRE issue and argument and TOEFL independent), training of the DWU measures was based on 25 prompts and 500 essays per prompt. For TOEFL integrated, the training set included 38 prompts and 500 essays per prompt. These training sets were also used to train the e-rater CVA features. All evaluations were based on a separate set of essays written to the same prompts.

In the evaluations of the DWU, a high-quality essay was defined as an essay with a score in the 4-6 range for GRE and 4-5 range for TOEFL, and low-quality essays were defined in the 1-3 range (alternative definitions resulted in less favorable results). In addition, evaluations used lemmatized words, that is, inflected forms of words were grouped together. The current evaluation included common noncontent stop words, such as *an*, *and*, *any*, *are*. Analyses showed that removing these words resulted in very similar values with slightly lower performance.

Several types of analyses are reported below. First, factor analyses were performed to investigate the structure of e-rater features when the DWU measures are added to the set. Factor analyses of both TOEFL computer-based test essays (Attali, 2007) and essays written by native English speakers from a wide developmental range (4th to 12th grade; Attali & Powers, 2008, 2009) revealed a similar underlying structure of the e-rater features. This three-factor structure has an attractive hierarchical linguistic interpretation with a word choice factor, a grammatical-conventions-within-a-sentence factor, and a fluency factor. These three factors can be complemented with a fourth, content factor. The purpose of the factor analyses in this paper was to find out on which of these factors the DWU measures would load. In particular, although the TDWU was conceived as a content feature, it is also interpretable as a vocabulary measure.

The second type of analysis was a comparison of four different e-rater regression models for the prediction of the human score from different sets of features. Model 1 includes all noncontent features. Model 2 introduces the TDWU feature as the only (generic) content feature. Model 3 replaces TDWU with the two existing prompt-specific content (CVA) features. Model 4 includes all three TDWU and CVA features. Finally, Models 5 and 6 replace CVA features from Models 3 and 4. Both the correlations and relative weights of the features are used to compare models. Relative weights displayed are from a generic model, that is, a single regression model across all 25 prompts. Note that although the prompt-specific content features are usually used in prompt-specific models (because their computation is based on prompt-specific information), it is possible to use them in a generic model. The correlations for both the generic and prompt-specific models are displayed for comparison.

The third analysis presents examples of extreme task-level *d* values for the issue and argument tasks, and the fourth analysis reports correlations between *d* values across levels and tasks, and between *d* values and word frequency.

Results

GRE Issue

All results are based on a validation sample of 500 essays for each of 25 GRE issue prompts. In this dataset, the TDWU measure had a skewness of -.67, a kurtosis of 0.16, and a *D* of .08 for the Kolmogorov-Smirnov goodness-of-fit test to the normal distribution. Table 1 shows the results of the four-factor analysis for GRE issue. The standardized coefficients of the oblique-rotated solution show the expected loadings for the existing features. Both DWU content measures load on the content factor, together with the two existing prompt-specific content features.

Table 2 shows the results of the six regression models for the prediction of the human score from different sets of features. Results show that the relative weight for TDWU is 25% and 19% in Models 2 and 4, respectively. These weights are second only to essay length. The correlation in Model 2 is substantially higher than in Model 1, and higher than in Model 3. In terms of TDWU's effect on other relative weights, the table shows that in Model 2 it lowers the weights of all features (compared to Model 1), except for essay length (whose weight is even higher). In Model 4, DWU mostly lowers the weights of the prompt-specific content features, from 20% (in Model 3) to 9%. Models 5 and 6 show that PDWU performs almost identically to the CVA features in Models 3 and 4, respectively.

GRE Argument

All results are based on a validation sample of 500 essays for each of 25 GRE argument prompts. In this dataset, the TDWU measure had a skewness of -.41, a kurtosis of .60, and a *D* of .04 for the Kolmogorov-Smirnov goodness-of-fit test to the normal distribution. Table 3 shows the results of the four-factor analysis. As with the issue task, the solution shows the expected loadings for the existing features, and the DWU content measures load on the content factor, together with the two existing prompt-specific content features.

Table 1

GRE Issue Factor Loadings

	Factor			
Feature	Word	Conventions	Fluency	Content
Word length	**0.75**	-0.05	-0.04	0.05
Vocabulary	**0.92**	0.07	0.06	0.00
Grammar	0.11	**0.58**	0.13	0.03
Mechanics	-0.09	**0.75**	-0.04	0.10
Usage	-0.14	**0.46**	0.08	0.29
Col/prep	0.15	**0.54**	-0.10	0.01
Essay length	0.00	-0.01	**0.96**	-0.03
Style	0.02	-0.01	**0.37**	0.22
Task DWU	0.13	0.16	-0.01	**0.75**
Prompt DWU	-0.02	0.01	-0.03	**0.97**
Value cosine	0.10	0.01	0.00	**0.86**
Pattern cosine	-0.06	0.13	0.17	**0.71**

Note. Col/prep = collocation/preposition, DWU = differential word use. Boldface indicates the expected high loadings for each factor.

Table 2

GRE Issue Models: Correlations and Relative Weights

			Word		Conventions				Fluency		Content			
Model	R_G	R_{PS}	WL	Voc	G	M	U	C/P	EL	S	T_D	P_D	Pat	Val
1	.805	.812	.08	.08	.07	.13	.15	.04	.42	.03	.	.	.	.
2	.827	.833	.05	.02	.05	.07	.08	.04	.45	.00	.25	.	.	.
3	.822	.830	.06	.04	.05	.09	.10	.03	.42	.01	.	.	.05	.15
4	.829	.836	.05	.02	.05	.06	.08	.03	.44	.00	.19	.	.03	.06
5	.824	.830	.07	.04	.05	.09	.08	.03	.43	.00	.	.20	.	.
6	.829	.834	.05	.02	.05	.07	.07	.03	.44	-.01	.18	.09	.	.

Note. C/P = collocation/preposition, EL = essay length, G = grammar, M = mechanics, Pat = pattern cosine, P_D ,= prompt-based differential word use (PDWU), R_G = correlation for generic model, R_{PS} = correlation for prompt-specific model, S = style, T_D = task-based differential word use (TDWU), U = usage, Val = value cosine, Voc = vocabulary, WL = word length.

Table 3

GRE Argument Factor Loadings

	Factor			
Feature	Word	Conventions	Fluency	Content
Word length	**0.70**	-0.10	0.00	0.08
Vocabulary	**0.86**	0.08	0.04	-0.02
Grammar	0.07	**0.58**	0.17	-0.01
Mechanics	-0.10	**0.69**	-0.03	0.08
Usage	-0.08	**0.50**	0.10	0.12
Col/prep	0.07	**0.52**	-0.08	0.03
Essay length	-0.03	-0.02	**0.99**	-0.02
Style	0.09	0.05	**0.46**	0.07
Task DWU	0.15	0.15	-0.04	**0.65**
Prompt DWU	0.04	0.08	-0.05	**0.86**
Value cosine	-0.01	-0.04	-0.01	**0.94**
Pattern cosine	-0.04	0.04	0.16	**0.83**

Note. Col/prep = collocation/preposition; DWU = differential word use. Boldface indicates the expected high loadings for each factor.

Table 4 shows the results of the same six regression models for the prediction of the human score from different sets of features. As with the issue task, the DWU feature displays a high weight in Model 2 (25%), second only to essay length. However, Models 3 and 4 show a stronger effect of the prompt-specific content features, reflected in a higher correlation in Model 3 than Model 2, and in slightly higher weights (12% and 15%) than DWU (11%) in Model 4. In addition, Models 5 and 6 show lower performance than Models 3 and 4, reflecting a slightly lower performance of PDWU than CVA features.

TOEFL Independent

All results are based on a validation sample of 500 essays for each of 25 prompts. In this dataset, the TDWU measure had a skewness of .04, a kurtosis of .47, and a D of .01 for the Kolmogorov-Smirnov goodness-of-fit test to the normal distribution. Table 5 shows the results of the four-factor analysis. Unexpectedly, the word length feature does not load on the word level factor. Instead, it loads negatively on the conventions factor, and both word level features have low but sizable loadings on the content factor. However, all other features load on the expected factors, and the DWU content measures load on the content factor.

Table 4

GRE Argument Models: Correlations and Relative Weights

			Word		Conventions				Fluency		Content			
Model	R_G	R_{PS}	WL	Voc	G	M	U	C/P	EL	S	T_D	P_D	Pat	Val
1	.748	.762	.09	.06	.07	.10	.10	.07	.48	.03	.	.	.	.
2	.778	.795	.05	.00	.04	.06	.06	.05	.48	.01	.25	.	.	.
3	.791	.802	.04	.00	.04	.03	.05	.04	.43	.01	.	.	.19	.15
4	.795	.807	.04	-.01	.03	.03	.05	.04	.45	.00	.11	.	.15	.12
5	.782	.795	.05	.00	.04	.05	.06	.04	.48	.01	.	.27	.	.
6	.790	.801	.04	-.02	.03	.04	.04	.04	.48	.01	.15	.19	.	.

Note. C/P = collocation/preposition, EL = essay length, G = grammar, M = mechanics, Pat = pattern cosine, P_D ,= prompt-based differential word use (PDWU), R_G = correlation for generic model, R_{PS} = correlation for prompt-specific model, S = style, T_D = task-based differential word use (TDWU), U = usage, Val = value cosine, Voc = vocabulary, WL = word length.

Table 5

TOEFL Independent Factor Loadings

	Factor			
Feature	Word	Conventions	Fluency	Content
Word length	**-0.07**	-0.48	-0.03	0.29
Vocabulary	**0.73**	0.02	0.00	0.26
Grammar	0.05	**0.69**	0.08	0.06
Mechanics	0.04	**0.81**	-0.09	0.06
Usage	-0.25	**0.42**	0.14	0.15
Col/prep	0.05	**0.37**	-0.08	0.18
Essay length	-0.14	0.25	**0.56**	-0.02
Style	0.13	-0.09	**0.73**	0.04
Task DWU	0.27	0.08	0.16	**0.62**
Prompt DWU	0.01	-0.01	0.13	**0.84**
Value cosine	-0.06	0.01	-0.02	**0.92**
Pattern cosine	0.02	-0.07	-0.12	**0.93**

Note. Col/prep = collocation/preposition, DWU = differential word use. Boldface indicates the expected high loadings for each factor.

Table 6 shows the results of the four regression models. Similarly to the issue task, the TDWU feature displays a high weight in Models 2 and 4 (22% and 16%), and these models perform better than Models 1 and 3, respectively. Also, similarly to the issue task, Models 5 and 6 (with PDWU) perform identically to Models 3 and 4 (with CVA features).

Table 6

TOEFL Independent Models: Correlations and Relative Weights

			Word		Conventions				Fluency		Content			
Model	R_G	R_{PS}	WL	Voc	G	M	U	C/P	EL	S	T_D	P_D	Pat	Val
1	.758	.770	.09	.09	.12	.14	.12	.07	.33	.04	.		.	.
2	.773	.782	.06	-.01	.10	.11	.11	.07	.34	.00	.22		.	.
3	.769	.780	.07	.03	.10	.13	.10	.07	.32	.03	.		.08	.06
4	.775	.785	.06	-.02	.10	.11	.10	.07	.33	.00	.16		.06	.02
5	.771	.781	.08	.03	.10	.12	.10	.07	.33	.01	.	.16	.	.
6	.775	.784	.06	-.01	.10	.11	.10	.07	.33	.00	.15	.09	.	.

Note. C/P = collocation/preposition, EL = essay length, G = grammar, M = mechanics, Pat = pattern cosine, P_D ,= prompt-based differential word use (PDWU), R_G = correlation for generic model, R_{PS} = correlation for prompt-specific model, S = style, T_D = task-based differential word use (TDWU), U = usage, Val = value cosine, Voc = vocabulary, WL = word length.

TOEFL Integrated

All results are based on a validation sample of around 4,000 essays for each of 38 prompts. In this dataset, the TDWU measure had a skewness of -.61, a kurtosis of 2.14, and a D of .04 for the Kolmogorov-Smirnov goodness-of-fit test to the normal distribution. In this dataset, the collocation/preposition feature was not available. Table 7 shows the results of the four-factor analysis. Unexpectedly, the essay length feature has a relatively high loading (.53) on the conventions factor and a low loading (.31) on the fluency factor. However, all other features load on the expected factors, and the DWU content measures load on the content factor, although TDWU also has minor loadings on the word (.18) and conventions (.35) factors.

Table 7

TOEFL Integrated Factor Loadings

	Factor			
Feature	Word	Conventions	Fluency	Content
Word length	**0.82**	-0.10	0.01	-0.06
Vocabulary	**0.60**	0.07	0.05	0.12
Grammar	0.04	**0.62**	0.11	0.01
Mechanics	0.01	**0.74**	-0.20	0.06
Usage	-0.08	**0.37**	0.14	0.02
Col/prep	-	-	-	-
Essay length	-0.11	0.53	**0.31**	0.02
Style	0.07	0.05	**0.67**	0.02
Task DWU	0.18	0.35	-0.03	**0.53**
Prompt DWU	-0.03	0.02	0.02	**0.92**
Value cosine	-0.02	-0.01	0.02	**0.92**
Pattern cosine	-0.02	-0.04	-0.01	**1.00**

Note. Col/prep = collocation/preposition, DWU = differential word use. Boldface indicates the expected high loadings for each factor.

Table 8 shows the results of the regression models. The prompt-specific content features have the largest weights and the strongest contribution to performance across all four tasks (Models 3 and 4). Similarly, the TDWU feature has the highest weight in Model 2 (46%), higher than essay length (31%). It also has a relatively larger effect on performance than in other tasks (correlation of .69 vs. .62 in Model 1). Similarly to the argument task, Models 5 and 6 (with PDWU) have lower performance than Models 3 and 4 (with CVA features).

Relation Between d Values

For task-level *d* values, the correlation between word values in the issue and argument tasks is quite low, .39 (.36 for words appearing at least 100 times in training). The correlations between *d* values and word frequency (log transformed), for words appearing at least 100 times in training, are quite low for issue (-.19) and argument (-.09).

Table 8

TOEFL Integrated Models: Correlations and Relative Weights

			Word		Conventions				Fluency		Content			
Model	R_G	R_{PS}	WL	Voc	G	M	U	C/P	EL	S	T_D	P_D	Pat	Val
1	.620	.643	.03	.13	.15	.17	.09	-	.36	.08	.	.	.	.
2	.689	.710	-.04	.05	.09	.02	.06	-	.31	.05	.46	.	.	.
3	.766	.774	.02	-.01	.06	.04	.04	-	.25	.03	.	.	.35	.22
4	.767	.776	.01	-.02	.05	.03	.04	-	.25	.03	.07	.	.33	.21
5	.749	.760	.02	.02	.06	.05	.05	-	.26	.03	.	.52	.	.
6	.750	.761	.01	.01	.06	.03	.05	-	.26	.03	.08	.48	.	.

Note. C/P = collocation/preposition, EL = essay length, G = grammar, M = mechanics, Pat = pattern cosine, P_D ,= prompt-based differential word use (PDWU), R_G = correlation for generic model, R_{PS} = correlation for prompt-specific model, S = style, T_D = task-based differential word use (TDWU), U = usage, Val = value cosine, Voc = vocabulary, WL = word length.

For prompt-level *d* values, the 300 pair-wise correlations across the 25 prompts are quite low, both for issue (*M* = .21, *SD* = .04) and argument (*M* = .25, *SD* = .03). The 25 correlations between prompt-level *d* values and task-level values have a mean of .43 (*SD* = .03) for issue and a mean of .42 (*SD* = .03) for argument, similar to the task-level correlation across tasks.

To get a sense of the kind of words that are characteristic of high- and low-scoring essays, examples of extreme (task-level) *d* values for the GRE issue task are shown in Appendix A and for the GRE argument task are provided in Appendix B.

Summary

With respect to the task-level DWU measure, TDWU was shown to be a strong predictor of human essay scores. For the two tasks that are less dependent on content, the issue and independent tasks, TDWU increased the correlation by .022 and .015 beyond noncontent features (Models 1 and 2). For these tasks, the inclusion of TDWU resulted in better performance than the inclusion of the CVA features (Model 2 compared to Model 3). The relative weight of TDWU was second only to essay length for these tasks (25% and 22% in Model 2).

For the two tasks that are more dependent on content, the argument and integrated tasks, the DWU had an even larger effect beyond noncontent features, with an increase in a correlation of .030 and .049 (Models 1 and 2). However, for these tasks, the inclusion of TDWU resulted in lower performance than the inclusion of the prompt-specific CVA features (Model 2 compared to Model 3). Nevertheless, the inclusion of both TDWU and CVA features increased performance even further (Model 4 compared to Model 3). For these tasks, the relative weight of DWU was again very high, particularly for integrated (46% in Model 2).

The prompt-specific DWU measure was very similar in performance to the CVA features for the two tasks that are less dependent on content. For the two tasks that are more dependent on content, its performance was inferior to CVA, especially for integrated.

The factor analyses confirmed that both DWU measures load on the content factor, together with the prompt-specific CVA features. However, for the integrated task, the loading of TDWU on the conventions factor was not insignificant (.35).

Although TDWU loaded on the content factor, its conceptual similarity to the vocabulary measure was confirmed in the regression analyses. The inclusion of the TDWU measure (in Model 2) decreased the relative weight of the vocabulary measure to near zero for issue and argument and below zero for independent. For integrated, the weight of the word length feature decreased to -.04.

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. (2011). *Modified CVA features.* Unpublished report. Princeton, NJ: ETS.

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

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

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

Attali, Y., & Powers, D. (2009). Validity of scores for a developmental writing scale based on automated scoring. *Educational and Psychological Measurement*, *69*, 978–993.

Fleiss, J. L. (1981). *Statistical methods for rates and proportions* (2nd ed.). New York, NY: John Wiley.

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

Landauer, T. K., Laham, D., & Foltz, P. W. (2003). Automated scoring and annotation of essays with the Intelligent Essay Assessor. In M. D. Shermis & J. Burstein (Eds.), *Automated essay scoring: A cross-disciplinary perspective* (pp. 87–112). Mahwah, NJ: Lawrence Erlbaum Associates, Inc.

Salton, G., Wong, A., & Yang, C. S. (1975). A vector space model for automatic indexing. *Communications of the ACM, 18*, 613–620.

Appendix A

Examples of Extreme *d* Values for the GRE Issue Task

For the GRE issue task, the 50 highest *d* values with at least 100 occurrences in the training data are (highest value first):

inherently, perpetuate, conformist, underlying, incredibly, commission, capitalism, censorship, translate, constituent, vietnam, volunteer, founding, patron, elsewhere, twentieth, suburban, angeles, seemingly, los, recipient, curriculum, inability, essentially, governmental, shakespeare, wage, societal, versus, california, farm, potentially, shareholder, gallery, jr., continually, federal, ad, progression, increasingly, band, heavily, array, framework, programming, mexico, participant, specifically, dinner, neighbor.

The 50 highest *d* values with at least 1,000 occurrences in the training data are:

americans, often, argue, perhaps, american, throughout, certainly, states, allow, simply, continue, conform, ideal, likely, potential, united, employee, argument, seek, corporation, while, offer, greater, upon, understanding, community, america, opportunity, within, longer, challenge, rely, cell, impact, desire, truly, structure, available, drive, overall, figure, force, story, similar, focus, off, home, funding, current, yet.

The 50 lowest *d* values with at least 100 occurrences in the training data are (lowest value first):

cant, u, stud, olden, dont, socity, firstly, thirdly, mobile, lots, till, korea, economical, eventhough, belive, behaviour, admittedly, speaker, sum, inorder, reseach, benifits, india, gandhi, assert, reasearch, develope, indian, nowadays, concede, totally, poeple, conclude, partially, astrologer, analyse, teller, invent, partly, bother, prevail, telling, lot, usa, according, thing, deeply, mahatma, days, etc.

The 50 lowest *d* values with at least 1,000 occurrences in the training data are (lowest value first):

> lot, thing, days, etc, kind, author, agree, nonmainstream, get, useful, things, astrology, being, leader, famous, painting, bad, he, main, solve, some, big, good, so, him, above, depend, finally, tell, country, development, think, should, like, every, heroine, let, know, living, easily, wrong, we, creation, various, any, role, very, problem, start, which.

Appendix B

Examples of Extreme *d* Values for the GRE Argument Task

For the GRE argument task, the 50 highest *d* values with at least 100 occurrences in the training data are (highest value first):

causation, socio-economic, translate, leap, socioeconomic, adequately, methodology, likewise, monetary, skew, equate, disprove, imp, accurately, conception, demographic, heavily, subjective, unfounded, unrelated, faulty, length, unsupported, additionally, frame, define, wealthy, variable, presence, irrelevant, similarly, comparable, strenuous, outline, random, flower, necessarily, interpretation, preference, assessment, unanswered, campaign, inaccurate, correlation, definition, administer, examine, logic, characteristic, participant.

The 50 highest *d* values with at least 1,000 occurrences in the training data are (highest value first):

necessarily, participant, perhaps, shy, simply, melatonin, similar, determine, significant, sample, furthermore, address, assumption, account, assume, blue, size, impact, while, comparison, protein, shyness, often, additional, karp, parson, actually, infant, whether, upon, district, either, question, population, claim, influence, group, likely, fail, could, nothing, themselves, Dickens, contribute, each, sanlee, flaw, current.

The 50 lowest *d* values with at least 100 occurrences in the training data are (lowest value first):

cant, auther, dont, programme, &, cure, cholestrol, dr.field, doesnot, dr, disadvantage, disagree, climatic, agree, page, our, refrain, concentrate, sit, fast, front, your, mega, milk, non, days, doesnt, century, forget, win, selling, strongly, loose, gas, fat, progress, enviroment, good, think, according, thing, you, thier, poorest-selling, totally, s, alot, ca, earth, get.

The 50 lowest *d* values with at least 1,000 occurrences in the training data are (lowest value first):

our, good, think, thing, you, get, I, my, so, arguer, always, we, pill, old, every, feel, like, conclude, lot, news, can, here, medium, us, reduce, save, now, moreover, view, product, college, place, condition, hall, people, read, cover, say, book, things, keep, some, membership, country, above, develop, food, print, give, restaurant.