Listening. Learning. Leading.®

Research Report
ETS RR–15-14

Automated Trait Scores for *TOEFL*® 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 *TOEFL*® 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 *TOEFL iBT*® independent and integrated tasks. In this study we explored the psychometric added value of reporting four trait scores for each of these two tasks, beyond the total e-rater score. The four trait scores are word choice, grammatical conventions, fluency and organization, and content. Trait scores were computed on the basis of several criteria for determining feature weights: regression parameters of the trait features on human scores, reliability of trait features, and coefficients of features from a principal component analysis. In addition, augmented trait scores, based on information from other trait scores, were also analyzed. The psychometric added value of trait scores beyond total e-rater scores was evaluated by comparing the ability to predict a particular trait score on one task from the same trait score on the other task versus 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; e-rater; augmented trait scores; alternative feature weighting schemes

doi:10.1002/ets2.12061

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 e-rater 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 testing (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 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, 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 (2011b) 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 in a comparison of the ability to predict a trait score on one test from the trait score on the other test to that from the total e-rater score on the other test. For example, the correlation between the grammatical conventions scores on the first and second test repeaters took was .66, but the correlation between grammatical conventions scores on one test and e-rater scores on another test was .55 to .56. In other words, the grammatical conventions score on one test is the best single predictor of another conventions score on a parallel 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. In analyses of subscores from operational tests, this condition is often 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 purpose of this paper was to evaluate the added value of automated trait scores for the TOEFL iBT writing section, which comprises two essay writing tasks.[1] The first is an *integrated* task that requires test takers to read, listen, and then write in response to what they have read and heard. The second is an *independent* task where test takers support an opinion on a topic.

The two tasks are scored by human raters using a holistic scoring rubric, and since 2009 (for the independent task) and 2010 (for the integrated task), e-rater has been used operationally as part of the scoring process.

There are several differences between the approach taken in Attali (2011b) and 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, independent and integrated, 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 TOEFL reports only one writing score, thereby supporting 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 independent and integrated conventions scores) to the correlations of the trait score in one task (e.g., the independent conventions score) with other scores on another task (e.g., the integrated e-rater score).[2] In other words, we would conclude that a trait score (e.g., the conventions score) has added value if it is predicted best, among all the scores on another task, by the score on the same trait (the conventions score).

In this paper, we also expand the coverage and definition of traits to include the content features of e-rater. In previous work (Attali, 2007, 2011b; Attali & Powers, 2008), the three-factor structure was based only on the noncontent features. In this paper, we considered content features as a fourth trait and evaluated the added value of all four traits: word choice, conventions, fluency and organization, and content/ideas.

Finally, in this paper we also compare 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 trait score calculation is based on a regression of the human essay scores on the essay features that contribute to the particular trait score. This is the most natural choice because this 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 coefficients from a principal component analysis (PCA). These coefficients can be interpreted as the regression weights for predicting the underlying component from the features that

are designed to measure this component. 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 (1927) classic regressed estimate of the true score. 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 all 37,390 examinees who took the TOEFL in November and December of 2010 around the world. In addition to test scores, the country of the test center was also available for analysis.

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.

Principal Component Analysis

In previous work, we employed factor analysis to investigate the underlying structure of e-rater features and possible invariances across developmental levels (Attali & Powers, 2008) and language groups (Attali, 2011b). In this paper, we are more interested in using the manifest features to compute trait scores. In this respect, an advantage of PCA is that it provides unique component scores (see, e.g., Widaman, 2007) that can be calculated as simple weighted sums of the manifest variables. Therefore, a four-component PCA was conducted for each task to estimate the coefficients (or loadings)

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.

Note: Col/prep = collocation and preposition use.

on the essay features. The coefficients obtained from PCA were later used for one of the weighting schemes of trait scores. It should be noted that maximum likelihood factor analyses resulted in similar feature loadings and factor correlations (results not shown here).

Independent Task

In Table 2, we present the overall correlation matrix for the features used in this study. Correlations range from −.15 to .88.

The first four eigenvalues for the PCA were larger than 1 (3.28, 2.11, 1.16, and 1.03) and together accounted for 69% of total variance. Table 3 presents the coefficients of the four-component PCA after an oblique promax rotation. The coefficients are, for the most part, as expected from a four-component solution, with the exception of a relatively high collocation and preposition use coefficient on the first component. Table 4 shows that correlations among the obliquely rotated principal components are relatively low, especially for word choice.

Table 2 Independent Task Feature Correlation Matrix

Feature	WL	G	U	M	CP	OR	EL	S	VC	PC
Vocabulary	.88	.09	−.10	.06	.42	.05	−.01	.34	.28	.21
Word length (WL)		.03	−.15	−.02	.36	.06	−.08	.27	.23	.15
Grammar (G)			.40	.47	.24	.18	.34	.19	.33	.28
Usage (U)				.35	.21	.11	.27	.10	.23	.16
Mechanics (M)					.29	.20	.34	.09	.37	.24
Col/prep (CP)						.09	.07	.20	.24	.09
Organization (OR)							.39	.13	.18	.10
Essay length (EL)								.39	.38	.19
Style (S)									.42	.21
Value cosine (VC)										.54
Pattern cosine (PC)										

Note: $N = 37{,}390$. Col/prep = collocation and preposition use.

Table 3 Coefficients From Principal Component Analysis (PCA) After Oblique Rotation: Independent Task

Feature	Component 1	Component 2	Component 3	Component 4
Vocabulary	.91	−.07	−.01	.12
Word length	.91	−.14	−.01	.07
Grammar	.02	.68	.05	.18
Usage	−.15	.73	−.01	.04
Mechanics	.02	.74	.05	.08
Col/prep	.66	.56	−.04	−.23
Organization	.05	.02	.85	−.24
Essay length	−.18	.13	.74	.18
Style	.25	−.16	.44	.39
Value cosine	.11	.17	.10	.73
Pattern cosine	−.04	.09	−.20	.88

Note: Col/prep = collocation and preposition use.

Table 4 Correlations Among the Rotated Principal Components: Independent Task

Feature	Grammar	Fluency	Content
Word choice	.02	.09	.21
Grammar		.28	.23
Fluency			.31

Table 5 Integrated Task Feature Correlation Matrix

Feature	WL	G	U	M	CP	OR	EL	S	VC	PC
Vocabulary	.33	.27	.10	.25	.19	.12	.19	.16	.26	.23
Word length (WL)		−.04	−.01	−.22	.10	.12	−.09	.07	.09	.12
Grammar (G)			.32	.47	.30	.17	.41	.24	.26	.22
Usage (U)				.22	.29	.10	.28	.24	.13	.13
Mechanics (M)					.30	.19	.38	.13	.27	.21
Col/prep (CP)						.11	.17	.17	.20	.19
Organization (OR)							.42	.15	.12	.12
Essay length (EL)								.41	.23	.22
Style (S)									.18	.15
Value cosine (VC)										.79
Pattern cosine (PC)										

Note: $N = 37{,}390$. Col/prep = collocation and preposition use.

Table 6 Coefficients From Principal Component Analysis (PCA) After Oblique Rotation: Integrated Task

Feature	Component 1	Component 2	Component 3	Component 4
Vocabulary	.57	.33	.06	.16
Word length	.90	−.08	.08	.01
Grammar	−.06	.65	.15	.06
Usage	.05	.69	.03	−.17
Mechanics	−.31	.58	.10	.17
Col/prep	.24	.77	−.19	−.03
Organization	.12	−.20	.86	−.02
Essay length	−.13	.18	.75	.04
Style	.18	.19	.53	−.07
Value cosine	.03	−.02	−.02	.94
Pattern cosine	.06	−.06	−.03	.95

Note: Col/prep = collocation and preposition use.

Table 7 Correlations Among the Rotated Principal Components: Integrated Task

Feature	Grammar	Fluency	Content
Word choice	−.04	−.04	.07
Grammar		.37	.35
Fluency			.26

Integrated Task

In Table 5, we present the overall correlation matrix for the features used in this study. Correlations range from around −.22 to .79.

The first four eigenvalues for the PCA were larger than 1 (3.20, 1.51, 1.23, and 1.07) and together accounted for 64% of total variance. Table 6 presents the coefficients of the four-component PCA after an oblique promax rotation. The coefficients are, for the most part, as expected from a four-component solution. Table 7 shows that correlations among the rotated principal components are relatively low, especially for word choice.

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 8 (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 (note that for each trait score and weighting scheme the sum of weights is 100%). The second set of weights was based on the cross-task reliabilities (correlation between the independent and integrated feature score) of the features that are shown in

Table 8 Alternative Relative Weights

	Regression-based		Reliability-based			PCA-based	
Feature	Ind.	Int.	Reliability	Ind.	Int.	Ind.	Int.
Vocabulary	159%	114%	.12	15%	15%	50%	39%
Word length	−59%	−14%	.65	85%	85%	50%	61%
Grammar	33%	37%	.49	26%	26%	25%	24%
Usage	20%	21%	.35	17%	17%	27%	26%
Mechanics	29%	26%	.64	42%	42%	27%	21%
Col/prep	18%	16%	.29	14%	14%	21%	28%
Organization	11%	−2%	.44	27%	27%	42%	40%
Essay length	68%	76%	.66	54%	54%	36%	35%
Style	21%	26%	.32	19%	19%	22%	25%
Value cosine	77%	55%	.26	55%	55%	45%	50%
Pattern cosine	23%	45%	.21	45%	45%	55%	50%

Note: Ind. = independent; Int. = integrated; PCA = principal component analysis.

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 5–6. Note that, with this method, the relative weights are the same for independent and integrated because the reliabilities of the two tasks are the same as an outcome of the way they were computed. A third set of scores based on the PCA coefficients from Tables 3 and 6 are presented in the two last columns.

Inspection of Table 8 shows that the two sets of regression-based and PCA-based weights (for independent and integrated) are similar for most of the features, with the exception of the word level 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. However, PCA-based weights are even more homogeneous than reliability-based weights, especially for word-choice features. In addition, the regression-based weights on word length and organization (only for the integrated task) are negative. This constitutes a serious problem because all features are expected to have a positive influence on essay scores. In an operational setting, this might be resolved by eliminating the feature with a negative weight from the score.

Regression-Based Trait Scores

In Table 9, we present the cross-task correlations or reliabilities (the diagonals), the within-task score correlations for the independent task (above the diagonal), and the integrated task (below the diagonal for regression-based trait scores, e-rater scores, and human scores). For example, the first value of 0.51 in the diagonal denotes the correlation between the human score on the integrated task and the human score on the independent task; the value of 0.69 toward the right of it denotes the correlation between the human score on the independent task and the e-rater score on the independent task; the number 0.57 below the first diagonal value denotes the correlation between the human score on the integrated task and the e-rater score on the integrated task. The reliability of e-rater scores is significantly higher than human scores (.70 vs. .51). The reliabilities of the grammar and fluency trait scores (.68 and

Table 9 Regression-Based Trait Scores: Within-Task Score Correlations and Cross-Task Reliabilities

Score	H	E	W	G	F	C
Human (H)	.51	.69	.25	.58	.54	.45
e-rater (E)	.57	.70	.44	.77	.79	.56
Word choice (W)	.23	.44	.07	.17	.11	.29
Grammar (G)	.46	.83	.32	.68	.39	.43
Fluency (F)	.53	.86	.22	.48	.64	.43
Content (C)	.27	.34	.26	.31	.25	.28

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

Table 10 Regression-Based Trait Scores: Cross-Task Score Correlations

	Independent				Integrated			
Score from other task	W	G	F	C	W	G	F	C
Word choice (W)	**.07**	.29	.17	.26	**.07**	.06	.17	.16
Grammar (G)	.06	**.68**	.37	.39	.29	**.68**	.41	.29
Fluency (F)	.17	.41	**.64**	.41	.17	.37	**.64**	.22
Content (C)	.16	.29	.22	**.28**	.26	.39	.41	**.28**
e-rater	.12	.63	.60	.47	.25	.57	.62	.32
Human	.23	.46	.38	.39	.27	.52	.49	.29
Reading	.28	.52	.39	.44	.32	.50	.48	.33
Listening	.25	.49	.37	.42	.27	.47	.49	.30
Speaking	.20	.49	.42	.43	.23	.48	.49	.28

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

.64) are not much lower than e-rater (.70), whereas the reliabilities of the word choice and content are very low (.07 and .28). Both the fluency and grammar trait scores show high correlations with e-rater scores (for fluency, .79 and .86), reflecting the high weights of the grammar and fluency features in e-rater scores. The highest correlations between trait scores is between grammar and fluency (.39 and .48), and the lowest is between word choice and fluency (.11 and .22). Overall, the corresponding correlations between scores in independent and integrated are similar.

In Table 10, we present 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 9. The question for each trait score is whether these reliabilities are higher than correlations between the trait score and other scores. In other words, the question is whether, in any column, the number in bold is the largest. The table shows that for grammar and fluency this is indeed the case for both tasks. For example, the highest correlation of independent grammar is with integrated grammar (.68), and the next highest correlation is with the total e-rater score (.63). For integrated fluency the difference between the correlation with independent fluency (.64) and independent e-rater (.62) is small but the differences are larger in the other three cases involving grammar or fluency. On the other hand, the very low reliabilities for word choice and content result in no added value. In fact, almost all scores from the other task have higher correlations (these cases are denoted with an underline). For example, to predict independent word choice it is better to use integrated fluency (.17) than integrated word choice (.07).

In Table 11, we show standardized mean scores for eight countries with the highest representation in the sample. To allow comparison across scores, all scores in the table were standardized. The patterns of average trait scores are, for the most part, similar for the different countries and across the two tasks. For example, for examinees from China, fluency scores (average standardized scores of .18 and .01) are somewhat higher than other trait scores. For examinees from India, grammar scores (average standardized scores of .16 and .23) are somewhat lower than other trait scores. The variability

Table 11 Regression-Based Trait Scores: Country-Standardized Score Differences

					Independent						Integrated					
Country	%	R	L	S	H	E	W	G	F	C	H	E	W	G	F	C
China	24	.00	−.31	−.46	−.20	.04	−.07	−.17	.18	−.03	−.14	−.07	−.13	−.12	.01	−.10
India	7	.40	.57	.77	.55	.34	.50	.16	.30	.51	.66	.48	.62	.23	.46	.39
Korea	6	.10	−.12	−.41	−.19	−.09	.17	−.04	−.30	−.12	−.12	−.08	.30	−.04	−.22	.23
Japan	4	−.30	−.47	−.84	−.41	−.40	−.08	−.01	−.64	−.61	−.47	−.45	−.22	−.10	−.70	−.11
Taiwan	2	−.06	−.13	−.27	−.22	−.06	.05	.05	−.19	−.15	−.22	−.03	.01	−.02	−.06	−.03
France	2	.37	.21	.25	.19	.18	−.05	.33	.15	.15	.08	.12	.07	.25	.04	.09
Germany	2	.49	.69	1.05	.52	.42	−.03	.62	.27	.25	.51	.57	.08	.67	.43	.38
Turkey	2	.05	.14	−.04	.06	−.02	.03	.06	−.14	.15	−.01	.09	.30	.05	.00	.11

Note: R = reading; L = listening; S = speaking; H = human; E = e-rater; W = word choice; G = grammar; F = fluency; C = content.

of trait scores *across* countries is quite similar, with somewhat larger variability for fluency scores (across all scores in the table, speaking scores exhibit the highest variability across countries).

Reliability-Based Trait Scores

In Tables 12–14, we present the same results as in Tables 9–11 for reliability-based trait scores. Reliability-based trait scores show slightly higher cross-task reliability for grammar (.71 vs. .68) and dramatically higher reliability for word choice (.62 vs. .07). The reason for this difference lies in the negative weights of the word length feature for regression-based trait scores (see Table 8). Note, in Table 13, that reliability-based trait scores have lower within-task correlations with other trait scores (the median difference is .06), and in Table 14 the lower cross-task correlations with other trait scores (the median difference is .04). Similar country scores are shown in Table 15 for the grammar, fluency and content trait scores (the median difference is .00), but dramatically different scores for the word-choice trait, further evidence that the two weighting schemes produced different word-choice scores. The combination of higher reliabilities and lower correlations with other scores slightly increases the added value of reliability-based grammar scores, and dramatically increases the added value of reliability-based word-choice scores. The reliability-based word-choice scores are the best predictors of the word-choice scores from the other task.

Principal Component Analysis-based Trait Scores

In Tables 15–17, we present similar results for PCA-based trait scores. Reliabilities are lower than for the reliability-based scores, especially for word choice (.51 vs. .62), grammar (.66 vs. .71), and fluency (.58 vs. .63). This is probably due to the homogeneous weights of PCA-based scores that do not take into account differences in reliability across tasks. Consequently, PCA-based trait scores have lower added value for these traits compared to reliability-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 content scores. To compute

Table 12 Reliability-Based Trait Scores: Within-Task Score Correlations and Cross-Task Reliabilities

Score	H	E	W	G	F	C
Human (H)	.51	.69	.20	.58	.53	.43
e-rater (E)	.57	.70	.38	.76	.78	.52
Word choice (W)	.08	.05	.62	.05	.04	.23
Grammar (G)	.45	.82	−.07	.71	.39	.41
Fluency (F)	.50	.82	.03	.46	.63	.38
Content (C)	.27	.34	.15	.31	.25	.29

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

Table 13 Reliability-Based Trait Scores: Cross-Task Score Correlations

	Independent				Integrated			
Score from other task	W	G	F	C	W	G	F	C
Word choice (W)	**.62**	.06	.00	.15	**.62**	−.07	.11	.14
Grammar (G)	−.07	**.71**	.36	.39	.06	**.71**	.40	.28
Fluency (F)	.11	.40	**.63**	.36	.00	.36	**.63**	.22
Content (C)	.14	.28	.22	**.29**	.15	.39	.36	**.29**
e-rater	.05	.63	.58	.47	.23	.56	.61	.32
Human	.18	.45	.38	.38	.11	.51	.46	.29
Reading	.24	.51	.38	.44	.17	.49	.47	.33
Listening	.17	.48	.36	.42	.08	.46	.44	.30
Speaking	.11	.48	.40	.42	.02	.46	.43	.28

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

Table 14 Reliability-Based Trait Scores: Country-Standardized Score Differences

					Independent						Integrated					
Country	%	R	L	S	H	E	W	G	F	C	H	E	W	G	F	C
China	24	.00	−.31	−.46	−.20	.04	.09	−.17	.21	−.03	−.14	−.07	.13	−.11	.12	−.10
India	7	.40	.57	.77	.55	.34	.23	.15	.30	.57	.66	.48	.25	.19	.39	.39
Korea	6	.10	−.12	−.41	−.19	−.09	.40	−.01	−.31	−.08	−.12	−.08	.53	−.02	−.09	.23
Japan	4	−.30	−.47	−.84	−.41	−.40	.01	.01	−.61	−.55	−.47	−.45	.12	−.09	−.55	−.10
Taiwan	2	−.06	−.13	−.27	−.22	−.06	.07	.09	−.15	−.14	−.22	−.03	.18	.01	.02	−.03
France	2	.37	.21	.25	.19	.18	−.28	.33	.19	.06	.08	.12	−.24	.26	.08	.09
Germany	2	.49	.69	1.05	.52	.42	−.17	.58	.29	.20	.51	.57	−.31	.62	.40	.38
Turkey	2	.05	.14	−.04	.06	−.02	.07	.06	−.15	.14	−.01	.09	.06	.09	−.06	.11

Note: R = reading; L = listening; S = speaking; H = human; E = e-rater; W = word choice; G = grammar; F = fluency; C = content.

Table 15 PCA-Based Trait Scores: Within-Task Score Correlations and Cross-Task Reliabilities

Score	H	E	W	G	F	C
Human (H)	.51	.69	.22	.58	.49	.42
e-rater (E)	.57	.70	.41	.77	.73	.50
Word choice (W)	.14	.19	.51	.10	.10	.25
Grammar (G)	.45	.78	.10	.66	.36	.39
Fluency (F)	.46	.75	.13	.40	.58	.34
Content (C)	.27	.34	.20	.30	.23	.28

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

Table 16 PCA-Based Trait Scores: Cross-Task Score Correlations

	Independent				Integrated			
Score from other task	W	G	F	C	W	G	F	C
Word choice (W)	**.51**	.15	.08	.22	**.51**	.05	.15	.15
Grammar (G)	.05	**.66**	.31	.37	.15	**.66**	.36	.28
Fluency (F)	.15	.36	**.58**	.32	.08	.31	**.58**	.21
Content (C)	.15	.28	.21	**.28**	.22	.37	.32	**.28**
e-rater	.07	.61	.53	.46	.27	.55	.57	.32
Human	.20	.45	.35	.38	.18	.51	.42	.29
Reading	.26	.51	.37	.43	.25	.49	.45	.33
Listening	.20	.49	.33	.41	.15	.46	.40	.30
Speaking	.14	.49	.37	.40	.09	.46	.38	.28

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

augmented scores, the standardized observed trait scores were used, together with the cross-task reliability 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.

In Table 18, we show 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 independent task, a weight of .60 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 independent task,[3]

$$\text{Augmented word choice score} = .60 \times \text{Observed word choice score} - .01 \times \text{Observed grammar score} + .02 \times \text{Observed fluency score} + .20 \times \text{Observed content score}.$$

The weight of the corresponding observed score in any column is marked in bold font. In Table 18, we show that in the computation of any augmented trait score, the corresponding observed trait score receives the largest weight, except

Table 17 Principal Component Analysis (PCA)-Based Trait Scores: Country-Standardized Score Differences

					Independent						Integrated					
Country	%	R	L	S	H	E	W	G	F	C	H	E	W	G	F	C
China	24	.00	−.31	−.46	−.20	.04	.05	−.16	.23	−.04	−.14	−.07	.07	−.09	.18	−.10
India	7	−.35	−.14	.01	.55	.34	.32	.13	.30	.59	.66	.48	.41	.21	.33	.38
Korea	6	.40	.57	.77	−.19	−.09	.35	−.09	−.31	−.07	−.12	−.08	.54	−.03	−.02	.23
Japan	4	.10	−.12	−.41	−.41	−.40	−.02	−.04	−.56	−.51	−.47	−.45	.03	−.13	−.46	−.11
Taiwan	2	−.30	−.47	−.84	−.22	−.06	.07	.04	−.10	−.13	−.22	−.03	.15	−.03	.07	−.03
France	2	−.07	.13	.25	.19	.18	−.23	.37	.23	.02	.08	.12	−.18	.27	.14	.09
Germany	2	−.06	−.13	−.27	.52	.42	−.14	.63	.31	.18	.51	.57	−.24	.65	.39	.37
Turkey	2	.37	.21	.25	.06	−.02	.06	.06	−.17	.13	−.01	.09	.15	.01	−.11	.11

Note: R = reading; L = listening; S = speaking; H = human; E = e-rater; W = word choice; G = grammar; F = fluency; C = content.

Table 18 Weights for Computation of Augmented Trait Scores

	Independent augmented score				Integrated augmented score			
Observed score	W	G	F	C	W	G	F	C
Word choice (W)	**.60**	−.01	−.01	.11	**.61**	−.05	.02	.07
Grammar (G)	−.01	**.62**	.10	.12	−.04	**.60**	.16	.09
Fluency (F)	.02	.13	**.53**	.13	.02	.20	**.53**	.05
Content (C)	.20	.30	.25	**.02**	.13	.22	.10	**.18**

Note: Figures on the diagonal, which are in bold, indicate the weight of the corresponding observed score.

Table 19 Correlations of Observed and Augmented Trait Scores

Score	Independent	Integrated
Word choice	.97	.98
Grammar	.91	.92
Fluency	.92	.95
Content	.59	.90

for the case of the augmented content score for the independent prompt in which the content score receives the *lowest* weight. Although the observed content score provides direct information about the true content score, the former has low reliability. Therefore, the other trait scores, which provide only indirect information about the true content score, receive a larger weight due to their high reliability. Note also that in some cases observed trait scores contribute negatively to the computation of an augmented score (e.g., the word choice observed score has a negative weight in the computation of the grammar and fluency augmented scores). These negative weights, although small in magnitude, are problematic from a construct point of view.

In Table 19, we show the correlations between observed and augmented trait scores. The very high correlations for word choice, grammar, and fluency, for the most part, reflect the higher reliabilities of these scores (and to some extent, modest correlations with other scores). Interestingly, despite the low reliability of content, the augmented integrated score is also highly correlated with observed integrated score (.90), whereas for independent the correlation is much lower. This is due to the higher within-task correlations between independent content and other independent traits (compared to integrated, see Table 12), which allow us to borrow more information for the independent content score.

In Tables 19 – 21, we present the same analyses as in Tables 12 – 14 for observed reliability-based trait scores. The cross-task reliabilities are about the same for word choice and grammar, increase by .06 for fluency, and increase dramatically for content, from .29 to .58. Naturally, all within-task correlations are higher. Especially high within-task correlations of over .9 can be observed between integrated e-rater and augmented fluency (.91) and between independent e-rater and augmented content (.96). 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. Finally, augmented and observed scores show similar patterns across countries (Table 22).

Table 20 Reliability-Based Augmented Trait Scores: Within-Task Score Correlations and Cross-Task Reliabilities

Score	H	E	W	G	F	C
Human (H)	.51	.69	.28	.64	.61	.65
e-rater (E)	.57	.70	.46	.84	.85	.96
Word choice (W)	.13	.10	.62	.27	.25	.66
Grammar (G)	.52	.88	.02	.72	.80	.80
Fluency (F)	.56	.91	.16	.83	.69	.81
Content (C)	.44	.64	.52	.78	.71	.58

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

Table 21 Reliability-Based Augmented Trait Scores: Cross-Task Score Correlations

	Independent				Integrated			
Score from other task	W	G	F	C	W	G	F	C
Word choice (W)	**.62**	.13	.10	.40	**.62**	.08	.21	.35
Grammar (G)	.08	**.72**	.61	.56	.13	**.72**	.63	.56
Fluency (F)	.21	.63	**.69**	.64	.10	.61	**.69**	.49
Content (C)	.35	.56	.49	**.58**	.40	.56	.64	**.58**
e-rater	.15	.71	.68	.63	.27	.63	.68	.56
Human	.25	.51	.47	.51	.15	.56	.55	.47
Reading	.32	.57	.50	.58	.22	.55	.56	.51
Listening	.25	.54	.47	.51	.13	.52	.52	.45
Speaking	.20	.55	.51	.50	.06	.52	.50	.41

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

Table 22 Reliability-Based Augmented Trait Scores: Country-Standardized Score Differences

					Independent						Integrated					
Country	%	R	L	S	H	E	W	G	F	C	H	E	W	G	F	C
China	24	.00	−.31	−.46	−.20	.04	.07	−.10	.11	.07	−.14	−.07	.11	−.09	.06	−.05
India	7	.40	.57	.77	.55	.34	.36	.37	.45	.39	.66	.48	.32	.34	.44	.47
Korea	6	.10	−.12	−.41	−.19	−.09	.34	−.10	−.27	.04	−.12	−.08	.54	−.01	−.01	.29
Japan	4	−.30	−.47	−.84	−.41	−.40	−.13	−.30	−.65	−.34	−.47	−.45	.08	−.25	−.48	−.17
Taiwan	2	−.06	−.13	−.27	−.22	−.06	.02	−.02	−.15	−.01	−.22	−.03	.17	−.01	.02	.04
France	2	.37	.21	.25	.19	.18	−.25	.29	.21	.10	.08	.12	−.22	.25	.14	.09
Germany	2	.49	.69	1.05	.52	.42	−.12	.53	.37	.33	.51	.57	−.24	.68	.52	.44
Turkey	2	.05	.14	−.04	.06	−.02	.10	.07	−.05	.00	−.01	.09	.07	.08	−.01	.11

Note: R = reading; L = listening; S = speaking; H = human; E = e-rater; W = word choice; G = grammar; F = fluency; C = content.

Table 23 Proportional Reduction in Mean Squared Errors (PRMSEs) of Augmented Trait Scores

	Independent		Integrated	
	Observed reliability	Augmented PRMSE	Observed reliability	Augmented PRMSE
Word choice	.62	.64	.62	.63
Grammar	.70	.74	.70	.75
Fluency	.63	.69	.63	.69
Content	.29	.94	.29	.56

We also computed the proportional reduction in mean squared errors (PRMSEs) of the augmented trait scores. In Table 23, we show 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. In Table 23, we show that all PRMSEs are higher than their corresponding reliabilities with especially large differences for content and very small differences for word choice.

Discussion

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, 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 subscores based on human analytic scoring rubrics can be very costly.

In this paper, we extend 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. In this paper we extend these results to the two TOEFL writing tasks and evaluate the content features as a fourth trait. Results from a PCA supported a four-component structure for the two TOEFL tasks. For three of the traits (word choice, grammar, and fluency), the cross-task reliabilities of these traits are relatively high (between .62 and .71 for reliability-based scores), but the reliability of the content trait score was very low (.29). The low reliability of content scores could be expected from the relatively large differences in definition of the two tasks—the independent task poses fewer restrictions on the ideas expressed in the essay. Therefore, from the perspective of the design of the two tasks, the content score should be least stable across the tasks.

Nevertheless, these results have implications for the psychometric value of reporting the different trait scores. In this paper, analyses of psychometric value added were based on predicting a particular trait score in one task from other scores in the other task. For the first three trait scores, the best predictor was always the same trait score in the other task. For the content trait score, other scores (notably the total e-rater score) were better predictors than the content score itself. This was especially true for the content score of the independent task—its correlation with the integrated e-rater score (.47) was much larger than that of the integrated content score (.29). In other words, these results support the value of three of four observed trait scores.

The use of augmented trait scores was explored as a way to improve the psychometric value of observed trait scores. Augmented scores borrow strength from other trait scores, depending on their reliability and intercorrelations. Augmented trait scores significantly improved the results of the content scores. The augmented independent content scores were the best predictors of integrated content scores, and the augmented integrated content scores were among the best predictors of independent content scores (together with e-rater and fluency scores). Augmented trait scores had an especially large effect on independent content scores. The weight of observed content scores was very low compared to the weights of other traits (Table 18), with a resulting correlation of only .59 between observed and augmented scores (Table 19). On the other hand, the large effect of other trait scores allowed the large improvement in performance of the independent content score, as is also evident in the large percent reduction in MSE (Table 23).

In the context of TOEFL writing assessment, the advantages of using cross-task correlations as a basis for evaluating the value of trait scores are that all examinees write in response to both tasks and that TOEFL treats the tasks as two items contributing to a single writing score. However, since the two tasks are different in their design and requirements for writing, it is reasonable to assume that the alternate-form *within-task* reliability of the different trait scores would be higher than the cross-task reliabilities found in this study. Consequently, the value of trait scores would likely be higher when evaluated in the context of a within task design. In other words, the results of this study can be seen as a lower bound for the reliability and value-added of e-rater trait scores.

Altogether, the PCA and value-added results can be seen as further evidence for the validity of e-rater as an alternative method for scoring essays. A fundamental issue in supporting the validity of automated essay scoring systems is establishing the meaning of features in terms of the writing skills they are supposed to measure. This is because features based on extracting computable elements from the text bear an indirect relation with linguistic or writing qualities we value. Therefore, it is not sufficient to develop a logical argument that connects a feature with a writing quality. In this respect, PCA can further clarify what individual features measure by establishing that logically related features are also statistically related. PCA can also be useful in supporting (or refuting) the intuitive rationale for new candidate features. For example, content analysis in the context of automated essay scoring has always been based on prompt-specific vocabulary analysis (Attali & Burstein, 2006; Landauer, Laham, & Foltz, 2003). Attali (2011a) proposed a task-level feature, on the assumption that specific tasks (such as the GRE issue or the GRE argument tasks) elicit distinct types of vocabularies that have commonalities across prompts. However, such a feature can also be interpreted as a general word-choice vocabulary feature. Nevertheless, PCA confirmed that the new feature clearly loads on the content component and not on the word-choice component.

In this paper, we explored another issue with implications for the validity of automated essay scores. As Bennett and Bejar (1998) noted, an advantage of automated scoring is that it makes it possible to control (to some degree) what Embretson (1983) called the construct representation (the meaning of scores based on internal evidence) and nomothetic span (the meaning of scores based on relationships with external variables) of automated scores. Construct representation can be controlled by the weights (or importance) of the different features. However, automated essay scoring applications traditionally give up this control by using weights that are based on optimal prediction of human essay scores. Since the computer does not evaluate essays as humans do (e.g., it does not understand the essay!), these weights will not necessarily reflect the relative importance of *human* measured traits. Nevertheless, performance issues have dominated the choice of human prediction weights. 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 PCA) had slightly higher value than those based on prediction of human scores, apart from possessing more homogeneous sets of feature weights.

Finally, although we document in this paper possible informational benefits of trait scores, it remains to be seen whether reporting of such trait scores can have beneficial effects on student learning, teacher instruction, or program decisions. For example, it would be interesting to see if students are able to respond to feedback about their essays that is based on trait scores, either in revising their essay or in writing a new essay.

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 GRE 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. (2011a). *A differential word use measure for content analysis in automated essay scoring* (Research Report No. RR-11-36). Princeton, NJ: Educational Testing Service.

Attali, Y. (2011b). *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 GRE writing tasks* (Research Report No. RR-15-15). Princeton, NJ: Educational Testing Service.

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

Bennett, R. E., & Bejar, I. I. (1998). Validity and automated scoring: It's not only the scoring. *Educational Measurement: Issues and Practice*, *17*(4), 9–17.

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

Embretson, S. E. (1983). Construct validity: Construct representation versus nomothetic span. *Psychological Bulletin*, *93*, 179–197.

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.

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

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

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: Erlbaum.

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.

Roid, G. H. (1994). Patterns of writing skills derived from cluster analysis of direct writing assessments. *Applied Measurement in Education*, *7*, 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., White, K. P. (1999). Using generalizability theory to estimate the reliability of writing scores derived from holistic and analytic scoring methods. *Educational and Psychological Measurement*, *59*, 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.

Widaman, K. F. (2007). Common factors versus components: Principals and principles, errors and misconceptions. In R. Cudeck, & R. C. MacCallum (Eds.), *Factor analysis at 100: Historical developments and future directions* (pp. 177–203). Mahwah, NJ: Erlbaum.

Suggested citation:

Attali, Y., & Sinharay, S. (2015). *Automated trait scores for TOEFL® writing tasks* (Research Report No. RR-15-14). Princeton, NJ: Educational Testing Service. http://dx.doi.org/10.1002/ets2.12061

Action Editor: James Carlson

Reviewers: Shelby Haberman and Isaac Bejar