Measuring the Power of Learning.®

Research Report
ETS RR–17-04

An Investigation of the *e-rater*® Automated Scoring Engine's Grammar, Usage, Mechanics, and Style Microfeatures and Their Aggregation Model

Jing Chen

Mo Zhang

Isaac I. Bejar

December 2017

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

An Investigation of the *e-rater*® Automated Scoring Engine's Grammar, Usage, Mechanics, and Style Microfeatures and Their Aggregation Model

Jing Chen, Mo Zhang, & Isaac I. Bejar

Educational Testing Service, Princeton, NJ

Automated essay scoring (AES) generally computes essay scores as a function of macrofeatures derived from a set of microfeatures extracted from the text using natural language processing (NLP). In the *e-rater*® automated scoring engine, developed at *Educational Testing Service* (ETS) for the automated scoring of essays, each of four macrofeatures (*grammar, usage, mechanics*, and *style* [GUMS]) is computed from a set of microfeatures. Statistical analyses reveal that some of these microfeatures might not explain much of the variance in human scores regardless of the writing tasks. Currently, the microfeatures in the same macrofeature group are equally weighted to produce the macrofeature score. We propose an alternative weighting scheme that gives higher weights to the microfeatures that are more predictive of human scores in each macrofeature group. Our results suggest that even though there is negligible difference between the proposed and the current equal weighting schemes and the current model in terms of the prediction of human scores and the correlation with external measures, our scheme improves the consistency of the resultant macrofeature scores across writing tasks to a considerable extent.

Keywords Automated scoring; feature aggregation; feature weighting; multiple linear regression; validity of machine scores

doi:10.1002/ets2.12131

Automated scoring is widely used to score constructed response items in educational tests given its advantages, such as low cost, real-time feedback, quick score turnaround, and consistency over time (Williamson, Bejar, & Hone, 1999). Automated scoring engines have been developed to score different types of constructed response, such as short responses (e.g., Attali, Powers, Freedman, Harrison, & Obetz, 2008; Leacock & Chodorow, 2003), essays (e.g., Attali & Burstein, 2006; Foltz, Laham, & Landauer, 1999; Shermis & Burstein, 2013), and speaking responses (e.g., Zechner et al., 2015). Automated essay scoring (AES)—the focus of our current study—is defined as the use of computer technology to evaluate and score written prose (Dikli, 2006). The scoring process of an AES system usually involves extracting and aggregating text features to predict a score that human raters would assign to a given response. Natural language processing (NLP) techniques are often used to extract text features (e.g., the average word complexity level in a text), and various statistical models can be used to weight and combine the extracted features to predict the human scores.

To establish the validity of automated essay scores, the features extracted should effectively and accurately evaluate the writing quality, and the statistical modeling of the features needs to be performed in substantively and technically defensible ways. The *e-rater*® automated scoring engine is the automated system used at *Educational Testing Service* (ETS) to score the writing quality of essays (Attali & Burstein, 2006; Burstein, Tetreault, & Madnani, 2013). This study investigates the performance of a set of microfeatures extracted in e-rater in evaluating the writing quality of essays and the statistical model that weights and combines these features in predicting human scores. In the following section, we introduce more details about e-rater, the microfeatures we evaluate, and their aggregation model.

e-rater GUMS Microfeatures

In the current practice, e-rater scores are generated by a linear combination of a set of high-level features computed for each essay with weights determined by regressing human ratings on the features. These features are also called *macrofeatures*. Most of these macrofeatures are composed of sets of lower-level features called *microfeatures* that are combined to

Corresponding author: J. Chen, E-mail: jchen@humrro.org

produce the macrofeature values. All of these macrofeatures and microfeatures are extracted by using NLP. The features of e-rater have been periodically updated and enhanced to improve engine performance. This study employs e-rater engine v13.1, which was used in operational scoring from July 2013 to June 2014. In this version of the engine, 10 macrofeatures are most commonly used to predict human scores. These 10 macrofeatures are *organization*, *development*, *grammar*, *usage*, *mechanics*, *style*, *word length*, *word choice*, *collocation and preposition*, and *sentence variety*. In addition to these 10 macrofeatures, two prompt-specific vocabulary usage features are used to predict human scores when the scoring model is custom built for each prompt.

Four of the 10 macrofeatures — grammar (G), usage (U), mechanics (M), and style (S) — are composed of sets of microfeatures. We refer to these microfeatures as GUMS microfeatures hereafter. Each GUMS microfeature detects a particular type of error in writing, and the value of the microfeature is the count of the respective errors. As an example, one microfeature detects spelling errors, and the value of this microfeature is the number of spelling errors found in an essay. There are 36 GUMS microfeatures in total, with 9 related to grammar, 9 related to usage, 12 related to mechanics, and 6 related to style. A brief description of each of these microfeatures is provided in the appendix. The grouping of the 36 GUMS microfeatures is based on the construct alignment of the microfeature to the more general linguistic aspects indicated by macrofeatures (Quinlan, Higgins, & Wolff, 2009).

The GUMS microfeatures are used in two ways in automated writing evaluation: producing macrofeature scores and providing diagnostic feedback of writing errors for a low-stakes online writing evaluation system (e.g., *Criterion*, ETS, 2008). For producing macrofeature scores, the GUMS microfeatures are aggregated to generate macrofeatures that are used to predict human scores directly. For providing diagnostic feedback on writing errors, the GUMS microfeatures are used to highlight the errors detected by these microfeatures in examinees' essays. These highlighted errors can be used by the examinees to guide the process of revision. In this study, we evaluate the GUMS microfeatures based only on how well they produce final e-rater scores rather than on how well they provide diagnostic feedback.

An aggregation model is used to combine the GUMS microfeatures to produce each macrofeature value. Currently, the aggregation model is in the form of

$$g = -\sqrt{\frac{\sum f_n}{Number\ of\ words}}$$

where g is the aggregated macrofeature score, and f_n are the n microfeature scores in the macrofeature group. *Number of words* is the total number of words in an essay. This aggregation model adopts an equal-weight scheme that assigns equal weights to all the microfeatures in the same macrofeature group. The rationale of the equal-weight approach is that each microfeature measures a single, but equally important, underlying construct. These microfeatures, therefore, should be weighted equally.

Problems With GUMS Microfeatures and Their Aggregation Model

Results from previous studies have pointed to the need to critically and systematically review the e-rater GUMS microfeatures. Burstein, Chodorow, and Higgins (2007) evaluated the accuracy of 28 GUMS microfeatures. Their results showed that some microfeatures had low accuracy,[1] and the grammatical error microfeature tended to either falsely identify or miscategorize grammatical errors compared with human annotations. However, the researchers also pointed out that the low accuracy of the microfeatures might reflect either a deficiency in the microfeature or problems with human annotation that potentially undermine the evaluation of accuracy. Thus, they suspected that some microfeatures might be underperforming and suggested future research to determine the causes for the low performance of microfeatures. Quinlan et al. (2009) evaluated GUMS microfeatures by comparing the microfeature scores with human holistic scores and comparing the errors detected by the e-rater to those detected by the human raters. They concluded that some microfeatures perform poorly in detecting errors and differentiating essay quality. Their data were collected from 11,955 students across Grades 4, 6, 8, 10, and 12. The frequency analysis showed that two GUMS microfeatures, *run-on sentence error* and *preposition error*, identified no cases in this large sample. Their correlation analysis also revealed an unexpected relationship between some pairs of microfeatures, such as *fused word* and *spelling errors*. The pairs were expected to be relatively independent, but empirical results suggested that they were strongly correlated. All the above studies have noted that the poor performance of the individual microfeatures may impact the overall construct coverage of the resulting automated score and its correlation with human scores.

As mentioned previously, the GUMS microfeatures are used to provide diagnostic feedback on writing errors in low-stakes online writing evaluation services. The errors detected by the GUMS microfeatures are highlighted, so the writers can fix the errors and resubmit their essays as a practice to improve their writing. In this context, if a microfeature detects a lot of false-positive cases, it may impact the user's assessment of the quality and usefulness of the system. Thus, the microfeatures are developed with precision as a paramount consideration. This leads to a smaller number of flagged errors because the system is flagging cases only where it is very confident that an error has occurred. However, this approach might not work very well to predict human scores because a feature that only detects a small number of flagged errors is unlikely to explain much variance in human scores. Precision-oriented features also potentially overestimate the overall grammaticality of the essay because there might be more errors than in fact were detected. Work has been conducted at ETS (Heilman et al., 2014) to develop a grammaticality feature to achieve a better balance of precision and recall rates to replace or supplement some of the current grammar microfeatures. However, the GUMS microfeatures reviewed in this paper were developed with precision as a paramount consideration, so the number of errors they detect might be smaller than the number of actual errors that exist in essays. This is an inherent problem of the GUMS microfeatures because of the design choice related to the feedback application. Therefore, it is worth reviewing these GUMS microfeatures to determine which microfeatures detect very few errors and might need further research.

The aggregation model of the GUMS microfeatures needs to be reviewed as well. In AES, the most commonly used feature-weighting scheme is statistically driven. The determination of feature weights is based purely on statistical procedures to optimize the prediction accuracy of human scores (Burstein, Tetreault, & Madnani, 2013; Foltz, Streeter, Lochbaum, & Landauer, 2013). Most often, feature weights are obtained by a multiple linear regression of human scores on a set of features. In the e-rater, at the macrofeature level, the feature weights for the macrofeatures are obtained by regressing human scores on the macrofeature scores. However, at the microfeature level, equal weight is assigned to each microfeature that belongs to the same macrofeature group. In other words, the empirical weight of each macrofeature obtained through the multiple linear regression is split equally between all the microfeatures that belong to the same macrofeature group.

There are advantages of assigning equal weight to all the microfeatures that belong to the same macrofeature group. Because there are a large number of microfeatures, statistically optimal weights for each of these features will make the model susceptible to changes across datasets. Using statistically optimal weights of the microfeatures will make it difficult to establish the meaning of automated scores over time (Attali, 2007). It is well known that equal-weight models have greater robustness than least squares regression coefficients-based models. It is also well known that equal-weight linear models can perform as well as optimal weight models under circumstances when the predictor variables intercorrelate positively, and the empirical weights of the predictors are not very different (Wainer, 1976; Wilks, 1938).

However, there are several disadvantages of the equal-weight scheme. First, it often yields lower predictions of human scores compared to models that are based on empirical weights to optimize the prediction of human scores. Second, from a construct perspective, the equal-weight scheme is not optimal. This equal-weight approach treats each error in the same macrofeature group equally. However, some errors might be more problematic than others in writing (Napoles, Cahill, & Madnani, 2016). For example, missing an initial capital letter will probably not influence the reader's understanding of the sentence, but missing a verb might make it hard to understand the meaning of a sentence. So, from a construct perspective, it might be more reasonable to differentiate the weights of microfeatures.

Finally, the aggregation model for GUMS microfeatures was developed at an early stage of the e-rater engine development (i.e., early 2000). There have been many changes over the past decade that may require an adjustment of the aggregation models for new settings. For instance, the engine feature set has seen substantial changes. New features have been developed and added into the engine, and old features have been dropped, all of which can conceivably cause changes to the relative contribution of each microfeature to the essay score. Because of these reasons, we conducted this study to explore alternative microfeature aggregation models upon examination of the current model.

An Alternative GUMS Aggregation Model

A property of the GUMS microfeatures in a macrofeature group is that, often, a single microfeature or a small number of microfeatures are responsible for most of the variability in that macrofeature (Almond, 2008; Haberman, 2004; Haberman, 2007). For example, the *spelling errors* microfeature tends to explain most of the variance of the mechanics macrofeature score. One potential explanation for this phenomenon is that there are many more opportunities to make spelling errors

(each word) than the opportunities to make comma errors (each clause or list) or errors of missing initial capital letter in a sentence (each sentence). A similar property occurs with the style macrofeature where the *repeated words* microfeature accounts for most of the variance.

Because of this property, an alternative aggregation model that we propose assigns empirical weights to the microfeatures in each macrofeature group that are most predictive of human scores. We introduce more details about how we determine the most predictive microfeatures in the Method section. These identified microfeatures will be treated as macrofeatures. They will receive empirical weights obtained by regressing human scores on these microfeatures and the other 10 macrofeatures introduced previously. However, in the multiple linear regression model, the GUMS macrofeatures are composed only of the remaining GUMS microfeatures that are not very predictive of human scores because the more predictive ones have already been included in the model at the macrofeature level. This means that the less-predictive GUMS microfeatures in each macrofeature group equally split the empirical weight of that macrofeature. Thus, this alternative model uses a combination of both equal and empirical weights to utilize the advantages of both weighting schemes.

In this study, we assume that the microfeatures that are the most predictive of human scores might be the high-performing microfeatures that deserve higher weights in the scoring model. Our assumption still needs to be verified with future research to evaluate the performance of microfeatures more comprehensively. The prediction of human scores is only one criterion to evaluate features. Other criteria, such as the precision and recall rates, can provide additional information about the microfeature performance as well as evaluations informed by investigating the relative gravity of the different errors.

We compare the machine scores generated from the current and the alternative aggregation models that we propose. Results from our study have implications for evaluating features and the models that aggregate low-level features into high-level features in AES. Two research questions and a set of subquestions for each guided this study.

1. How well do microfeatures perform in evaluating the writing quality of essays?
    a. What are the frequency distributions of the microfeatures' scores?
    b. How well do the microfeature scores associate with the human scores?
    c. How consistent are the microfeature scores across tasks?
2. Compared to the current microfeature aggregation model, does the alternative model produce automated scores that show
    a. better prediction of human scores?
    b. stronger associations with the other sections of the test?
    c. greater cross-task consistency of the resultant macrofeature scores?

Method

Data

In this study, we used much larger and more recent datasets from high-stakes assessments compared to previous studies (e.g., Quinlan et al., 2009), making the results potentially more useful in informing operational scoring. Our investigation was conducted using essays from the writing tasks of two large-scale high-stakes assessments. Assessment I is a graduate school admission test, and Assessment II is a college-level English proficiency test. The writing section in Assessment I includes two tasks, which we refer to as *Task A* and *Task B* hereafter. Task A requires examinees to critique an argument, whereas Task B requires examinees to construct an argument using examples and relevant reasoning. The writing section of Assessment II also has two tasks, which we refer to as *Task C* and *Task D*. Task C requires test takers to articulate and support an opinion on a topic. Task D requires test takers to respond in writing by synthesizing the information they have read with the information they have heard.

The essays from Assessment I and Assessment II were collected from the administrations between March 2012 and December 2012 and between January 2011 and December 2012, respectively. The sample size of the essays from Tasks A, B, C, and D were 199,966; 199,957; 197,112; and 195,811, respectively. Each essay was scored by at least one human rater and by the e-rater. Essays detected as unscorable by the e-rater filtering system were excluded from this study. The e-rater macro- and microfeature scores were generated for all the essays, including all the GUMS microfeatures. Because both

Assessment I and II had two separate writing tasks, an examinee usually wrote two essays responding to the two different writing tasks. Feature values of both essays from the same examinees were extracted, which allowed us to compare the consistency of the resultant GUMS macrofeature scores across writing tasks generated from different aggregation models.

In addition, we collected examinees' scores on the other section of the test as external variables to evaluate the validity of automated scoring generated from different aggregation models. Specifically, we obtained the verbal section scores for Assessment I and the reading, listening, and speaking sections' scores for Assessment II. The verbal score of Assessment I measures examinees' reading ability and reasoning skills. The reading score of Assessment II measures examinees' ability to read academic texts; the listening score measures examinees' listening comprehension of lectures, classroom discussions, and conversations; and the speaking score measures examinees' ability to express an opinion on a familiar topic or to speak based on reading and listening tasks.

Data Analysis

For the first research questions, we took statistical approaches to evaluate the performance of the GUMS microfeatures. Instead of evaluating the performance of the e-rater microfeatures against human annotation, we examined how microfeatures perform using simple descriptive statistics. Given that we had a large and representative sample of essays from both testing programs, the descriptive statistics should reflect how these microfeatures perform in terms of frequency and distribution. We analyzed frequency and percentage distributions of each microfeature score for all four writing tasks. If a microfeature has many zero scores across all the writing tasks, the microfeature might not explain much of the variance in the human scores regardless of the writing task.

We also evaluated each microfeature by analyzing whether its performance conformed or deviated from expected patterns. For example, because each GUMS microfeature is a count of a particular type of error, it should be negatively related to human scores that reflect the quality of the essays. Hence, we computed the correlations between individual microfeatures and human scores. By using simple descriptive statistics, we tried to identify GUMS microfeatures that were performing in unexpected ways. In addition to the frequency and correlation analyses, we also evaluated the performance of each microfeature by analyzing the correlation of the same examinee's microfeature scores across two writing tasks. A stronger correlation across tasks would arguably indicate higher stability of the microfeatures.

For the second research question, we first aggregated the GUMS microfeatures using the current equal-weight model and the proposed alternative model. To build the alternative aggregation model, we first identified the microfeatures in each macrofeature group that were the most predictive of human scores. We analyzed the relative importance of each microfeature in each macrofeature group in predicting human scores. The relative importance was determined by the proportionate contribution of each microfeature to R-squared, a number that indicates the proportion of the variance in the dependent variable that is predictable from the independent variables, when regressing human scores on all the microfeatures in the same macrofeature group. We used the LMG method, a method named after the authors Lindeman, Merenda, and Gold (1980), to determine the R-squared contribution averaged over all possible orderings of the entry of regressors (i.e., when regressors enter into the predictions). There are different methods for assessing relative importance in linear regression. Among these methods, the LMG method is one of the methods recommended by researchers (Grömping, 2006). In the scoring model, the identified microfeatures are treated as macrofeatures, and the GUMS macrofeatures are only aggregated by the remaining microfeatures that are not very predictive of human scores because the more predictive ones have already been included in the scoring model at the macrofeature level.

To build and validate the current and the proposed scoring models, essays from each writing task were randomly split into two halves: a training dataset and a validation dataset. The descriptive statistics (e.g., frequency, correlation analysis) for research question 1 were based on the training datasets only because the sample size was sufficiently large to provide information for the whole population. For the second research question about different aggregation models, all the optimal weights were determined using the training sample and were then applied to the validation sample separately for each writing task. We report results of the second research question based on the validation datasets.

The current and the proposed aggregation models were compared from three aspects. First, resultant e-rater scores from both models were evaluated based on their predictions of human scores according to the root mean squared error, which measures the differences between the model-predicted values (i.e., e-rater scores) and the actually observed values (i.e., human scores). Second, the aggregation models were compared using the association between resultant e-rater scores and external variables, such as examinees' scores on the other sections of the test. Previous studies evaluated automated scores

Table 1 Percentage of Essays That Received Score 0 for Each Microfeature in Each of the Four Datasets

Microfeatures	Task A ($N = 99982$)	Task B ($N = 99974$)	Task C ($N = 98963$)	Task D ($N = 98025$)
GUMS 101	85.2	85.9	79.3	88.0
GUMS 102	77.7	79.8	73.2	76.5
GUMS 103	94.3	94.8	89.3	91.0
GUMS 104	68.3	71.6	48.8	50.8
GUMS 105	76.9	76.4	61.3	73.1
GUMS 106	99.0	96.4	97.0	99.4
GUMS 107	81.7	78.5	85.6	88.9
GUMS 108	96.2	96.8	95.4	95.3
GUMS 109	83.5	81.7	76.6	83.2
GUMS 201	81.9	78.9	63.0	75.9
GUMS 202	12.7	18.0	9.5	13.5
GUMS 203	65.9	63.5	59.9	74.7
GUMS 204	96.1	96.1	94.1	92.9
GUMS 205	99.0	99.2	96.1	99.0
GUMS 206	69.0	67.1	61.8	73.1
GUMS 207	99.9	99.8	99.3	99.9
GUMS 208	99.7	99.7	99.6	99.7
GUMS 209	98.3	98.9	97.7	98.9
GUMS 301	7.0	7.8	2.6	3.4
GUMS 302	96.5	90.7	79.7	94.6
GUMS 303	88.5	89.6	77.7	81.6
GUMS 304	94.5	97.1	95.8	98.3
GUMS 305	91.7	91.3	86.0	87.8
GUMS 306	98.0	96.4	96.4	98.6
GUMS 307	41.6	39.1	40.5	58.0
GUMS 308	88.7	92.5	96.1	98.1
GUMS 309	99.6	99.2	97.7	99.8
GUMS 310	65.0	66.3	63.0	73.4
GUMS 311	90.1	90.5	91.5	91.1
GUMS 312	79.1	74.5	69.0	80.0
GUMS 401	42.0	40.6	22.3	15.7
GUMS 402	99.9	99.9	99.8	99.9
GUMS 403	98.6	98.4	96.2	97.8
GUMS 404	98.5	98.3	94.5	98.4
GUMS 405	86.5	87.0	88.4	92.0
GUMS 406	74.8	64.9	82.4	65.2

based not only on the consistency between automated scores and human scores but also on the relationship of automated scores with external criteria (Attali, Bridgeman, & Trapani, 2010; Petersen, 1997; Weigle, 2010). The assumption is that if human and automated scores reflect similar constructs, they are expected to relate to other measures of constructs in similar ways and should thus show similar correlation patterns (Bennett & Zhang, 2016). Finally, the aggregation models were compared based on the consistency of the resultant macrofeature scores (grammar, usage, mechanics, and style) across writing tasks. It is logical to expect the same student's GUMS scores be consistent to some extent across the writing tasks because these macrofeatures measure students' fundamental language abilities.

Results

Results for Research Question 1: Performance of GUMS Microfeatures

Frequency Analyses

The frequency analysis showed that some GUMS microfeatures had zero or very few cases across all four datasets. Table 1 shows the proportion of essays that received a score of 0 for each microfeature in each of the four datasets. For 10 of 36 microfeatures, over 95% of the essays received 0 across all four datasets. These features are GUMS 106, 108, 205, 207, 208, 209, 306, 309, 402, and 403. The result suggests that these microfeatures might lack effectiveness in differentiating the

quality of essays, and further research is needed to determine whether these microfeatures are capturing errors they are supposed to capture.

Correlational Analyses

Because the GUMS microfeatures are often associated with the errors detected in an essay, the expected human feature correlation was negative. The results suggested that the correlations are negative for most microfeatures, with some exceptions. For example, GUMS 307 (missing comma), GUMS 308 (hyphen error), and GUMS 406 (passive voice) were positively correlated with human scores across all four datasets. These microfeatures should be further analyzed due to the unexpected relationship between the microfeature scores and human scores. The positive correlation between the passive voice feature and human scores has already been revealed by Quinlan et al.'s study (2009). The authors explained that passive voice is often used in academic writing, which should not be considered a violation of style. As mentioned previously, in the most recent version of the e-rater engine, the *style* microfeatures were removed from the scoring models and used only to provide diagnostic feedback. Thus, the positive correlation between passive voice and human scores might not be problematic if the microfeature is not used to predict human scores. However, the other two microfeatures, missing comma and hyphen error, need to be further analyzed to find the cause of the unexpected correlation between human scores and these microfeatures.

Consistency Analysis

Table 2 presents the correlation between the microfeature scores of the two essays from the same examinee. In general, the correlations ranged from very weak (less than 0.2) to weak to moderate (between 0.2 and 0.5). A small subset of microfeatures had higher cross-task correlations than others, for example, GUMS 102 (run-on sentence), 104 (subject–verb agreement), 202 (article error), 301 (spelling error), 303 (missing initial capital letter), 307 (missing comma), and 405 (too many long sentences). GUMS 301 (spelling error) had the highest cross-task correlations of all microfeatures: 0.531 for Assessment I and 0.442 for Assessment II. All remaining microfeatures had very weak cross-task correlations. One possible reason for this phenomenon is that the microfeatures fail to identify errors because most essays receive a microfeature score of 0. Another possible reason is that the cross-task consistency is truly low, which might indicate that the same type of errors that a microfeature captures does not appear across writing tasks consistently. In addition, different tasks might afford different opportunities for making errors. For example, Tasks A and D have more extensive prompts than Tasks B and C because these two tasks provide test takers some texts to read and require them to respond to these texts in writing. Test takers can copy prompt texts, which could possibly alter the pattern of GUMS errors across essays from the same students.

Results for Research Question 2: Comparisons Between Two Aggregation Models

Relative Importance of Microfeatures in Each Macrofeature Group

To build the alternative aggregation model using both equal and empirical weights, we first identified the microfeatures in each macrofeature group that were the most predictive of human scores. We calculated the proportionate contribution of each microfeature in each macrofeature group to the R-squared in the linear regression model in predicting human scores. In each macrofeature group, we picked one to three microfeatures that contributed around or above 15% of the R squared averaged across four datasets as the most predictive microfeatures. In total, we identified two grammar microfeatures, three usage microfeatures, three mechanics microfeatures, and one style microfeature as the most predictive microfeatures.

Prediction of Human Scores

There is negligible difference between the proposed model and the current model in terms of the prediction of human scores. Table 3 presents the agreement statistics between human scores and e-rater scores generated from the two aggregation models. Model 1 is the current equal-weight aggregation model, and Model 2 is the proposed model. The root

Table 2 Correlation Between Microfeature Scores Across Writing Tasks for the Same Examinees

	Assessment I Tasks A and B ($N = 99931$)	Assessment II Tasks C and D ($N = 97517$)
GUMS 101	0.091	0.142
GUMS 102	**0.201**	**0.258**
GUMS 103	0.152	0.095
GUMS 104	**0.274**	**0.245**
GUMS 105	0.188	0.183
GUMS 106	0.016	0.009
GUMS 107	0.101	0.041
GUMS 108	0.060	0.039
GUMS 109	0.073	0.062
GUMS 201	0.164	0.124
GUMS 202	**0.310**	**0.262**
GUMS 203	0.174	0.124
GUMS 204	0.029	0.025
GUMS 205	0.030	0.029
GUMS 206	0.069	0.054
GUMS 207	0.073	0.068
GUMS 208	0.004	0.024
GUMS 209	0.043	0.026
GUMS 301	**0.531**	**0.442**
GUMS 302	**0.246**	0.150
GUMS 303	**0.398**	**0.477**
GUMS 304	0.087	0.121
GUMS 305	0.188	0.199
GUMS 306	0.158	0.146
GUMS 307	**0.329**	**0.299**
GUMS 308	0.041	0.023
GUMS 309	0.027	0.014
GUMS 310	0.163	0.126
GUMS 311	0.079	0.055
GUMS 312	0.137	0.175
GUMS 401	0.167	0.183
GUMS 402	0.010	0.005
GUMS 403	**0.222**	0.199
GUMS 404	0.119	**0.211**
GUMS 405	**0.290**	**0.321**
GUMS 406	0.070	0.070

Note. The sample sizes for the consistency analysis are slightly smaller than the sample sizes in previous analyses because some examinees only wrote one essay, and those data were excluded from the consistency analysis. Correlations greater than 0.20 are given in bold.

mean squared errors (RMSEs) indicated negligible differences between Model 1 and Model 2 in terms of the prediction of human scores. Thus, Model 2 is not preferable to Model 1 from this perspective.

Correlation With External Measures

Table 4 presents the correlations of human or e-rater scores with examinees' scores on the other sections of the test. Overall, e-rater scores generated from both models related to the scores on the other sections of the test in ways that are similar to the way human scores are related to the other sections of the test. If the human and e-rater scores reflect a similar construct, they should relate to examinees' scores on the other test parts in a similar fashion, thereby providing further validity evidence for the e-rater scores. The difference of correlations between Model 2 and Model 1 presented in Table 4 is marginal. So, Model 2 is not preferable to Model 1 when they are evaluated based on whether scores from these models relate to the scores on the other sections of the test in similar ways as human scores do.

For Task D, human scores had a considerably higher correlation with examinees' listening scores than the machine scores generated from both models. The correlation between human scores and examinees' listening scores was around 0.67; however, the correlation between e-rater scores and examinees' listening scores was only 0.58. This is the largest

Table 3 Agreements Between Human and e-rater Scores Resulting from Different Aggregation Models[a]

Writing tasks	Sample size (*N*)	RMSE (Model 1)	RMSE (Model 2)
Task A	99,984	0.6266	0.6269
Task B	99,983	0.5429	0.5414
Task C	98,419	0.5852	0.5893
Task D	97,786	0.9043	0.9069

Notes. Model 1: current equal-weight model. Model 2: empirical and equal-weight model. RMSE = root mean squared error.
[a]In this study, the e-rater scores generated were slightly different from the e-rater scores generated in operational scoring. In operational scoring, the final e-rater scores were adjusted so that the distribution of the scores (e.g., mean and standard deviation) matched up with the distribution of human scores. In this study, we did not include this adjustment step for both models. Thus, the results from both models are comparable, but the human – machine agreement level is slightly lower than that obtained from operational scoring.

Table 4 Correlations Between Human or e-rater Scores and External Variables

	N	External variables	Human	e-rater (Model 1)	e-rater (Model 2)
Task A	99,984	Verbal	0.6153	0.6256	0.6220
Task B	99,983	Verbal	0.5753	0.6015	0.5966
Task C	98,419	Speaking	0.6058	0.6269	0.6258
		Reading	0.5392	0.6118	0.6047
		Listening	0.5573	0.5724	0.5670
Task D	97,786	Speaking	0.6031	0.6027	0.5989
		Reading	0.6244	0.5935	0.5891
		Listening	0.6650	0.5788	0.5748

Table 5 Correlations of Resultant Macrofeatures Across Two Writing Tasks

	Assessment I (Tasks A and B) (*N* = 99939)		Assessment II (Tasks C and D) (*N* = 93173)	
	Model 1	Model 2	Model 1	Model 2
Grammar	0.4038	0.4939	0.3910	0.4078
Usage	0.4092	0.4767	0.2816	0.3151
Mechanics	0.6456	0.6911	0.6040	0.5728
Style	0.2724	0.2549	0.2767	0.3173

difference between the correlations from human scores and those from machine scores as shown in Table 4. This writing task requires test takers to respond in writing by synthesizing the information that they had read and heard. Human raters might perform better in evaluating whether examinees appropriately synthesized information they had heard in their essays. Thus, human scores have a stronger correlation with examinees' listening scores than machine scores. Currently, researchers at ETS are developing features that intend to capture the use of information from external sources to improve *e-rater* performance. See Beigman Klebanov, Madnani, Burstein, and Somasundaran (2014) for details.

Consistency of Resultant Macrofeatures

For both models, we calculated the consistency of macrofeature scores across two writing tasks. For Model 1, each macrofeature was computed by aggregating all its contained microfeatures with equal weight. For Model 2, each macrofeature was computed by aggregating all its contained microfeatures with empirical weights for the identified more-predictive microfeatures and equal weight for the less-predictive microfeatures. In general, we found that the macrofeature scores resulting from Model 2 had greater cross-task consistency than the ones resulting from Model 1, as shown in Table 5. In only two cases did Model 1 show slightly greater consistency than Model 2 (i.e., 0.2724 vs. 0.2549 for style in Assessment I and 0.6040 vs. 0.5728 for mechanics in Assessment II).

In Assessments I, the macrofeature scores for grammar, usage, and mechanics generated from Model 2 were notably more consistent across writing tasks than those generated from Model 1. For example, the cross-task correlation between

the same examinee's grammar scores was 0.49, considerably higher than the 0.40 resulting from Model 1. In Assessment II, the differences of consistency between Model 1 and Model 2 were smaller than those from Assessment I, but Model 2 generated more consistent grammar, usage, and style macrofeature scores than Model 1.

Discussion

This study investigated e-rater GUMS microfeatures and the model used to aggregate these GUMS microfeatures into macrofeatures. Two research questions were addressed. The first question mainly concerned the performance of the microfeatures in differentiating the writing quality of essays. The second question compared the performance of two weighting schemes for microfeatures aggregation.

For the first research question, we applied different statistical approaches to evaluate the performance of individual microfeatures. First, frequency analysis suggests that some microfeatures do not discriminate the essay quality well. Among 36 features, there were 10 microfeatures for which, overwhelmingly, more than 95% of the essays received a zero score, meaning that no respective error was detected. These microfeatures should be further analyzed from the NLP perspective, possibly in combination with qualitative analyses, in order to determine if they indeed capture the errors as intended. For instance, GUMS 207 intends to detect various nonwords commonly used in oral language, such as *gonna*, *kinda*, etc. It is worthwhile examining whether it is true that over 95% of the essays do not have any nonwords or if this microfeature might fail to capture some nonwords.

Second, correlational analysis showed that most of the microfeatures negatively correlated with human scores, as expected. However, there were some exceptions, such as GUMS 307 (missing comma), 308 (hyphen error), and 406 (passive voice), which were positively correlated with human scores across all four datasets. It is worth revisiting these microfeatures to see whether they accurately capture comma or hyphen mistakes (e.g., check the precision and recall rates of these microfeatures) as intended. Meanwhile, it is worth considering the exclusion of these microfeatures from the aggregation model due to the unexpected correlations with human scores.

Third, cross-task consistency analysis provided additional information about the performance of these GUMS microfeatures. Given that these microfeatures intend to measure examinees' basic writing skills, which should be relatively stable across writing tasks, we would expect the microfeature scores of the same students' two essays to be related to some extent. However, the results showed that only a small number of microfeatures had weak to moderate cross-task associations (between 0.20 and 0.50), whereas most microfeatures had very weak or no cross-task associations (lower than 0.20). Further analyses (e.g., qualitative annotation) are needed to determine the root cause of this. The generally low cross-task consistency might be due to the fact that the microfeature fails to detect any errors, or the errors that a microfeature captures do not appear consistently across writing tasks. In addition, some tasks might provide more or fewer opportunities to make GUMS errors. For example, if test takers are required to write an essay in response to a reading passage, they might copy many words form the passage, thus making fewer spelling errors than they would in a task where no reading material is given. Therefore, the low cross-task consistency might be due to influence from writing tasks. In sum, results from different kinds of statistical analyses on the microfeatures all point to the fact that some microfeatures are potentially more problematic than others. Microfeatures such as GUMS 106, 108, 205, 207, 208, 209, 306, 307, 308, 309, 402, 403, and 406 should be further analyzed to find the causes of unexpected patterns in order to ensure the validity of the resultant automated scores.

For the second research question, we compared and evaluated two aggregation models for the GUMS microfeatures. We evaluated these two models from three perspectives: the prediction accuracy of human scores, the correlation between the resultant e-rater scores and external variables, and the consistency of resultant macrofeature scores across writing tasks. Our results revealed that the proposed model and the current model perform similarly in terms of the prediction of human scores and the correlation with external measures. However, the proposed model produced macrofeature scores with greater cross-task consistency than the current model for both assessments.

Higher consistency of GUMS macrofeatures across writing tasks is a desirable property. In the *Criterion* online writing evaluation service of ETS, test takers will get e-rater scores that evaluate the overall quality of their essays and trait scores that evaluate certain aspects of their writing, such as their language control and their word choice. These trait scores are generated based on a subset of the macrofeatures. The language control trait score is generated based on the grammar, usage, and mechanics scores. Higher consistency of the macrofeatures will make this trait score more reliable across tasks, which is a reasonable assumption because the language ability is relatively stable across writing tasks.

Two limitations of this study are worth noting. One limitation is that we evaluated the microfeatures' performance and their aggregation model in only two assessments. The results may not be generalizable to other assessments with different stakes, examinee population, and test purpose. A second limitation is that because we have taken a statistical approach to evaluate the performance of GUMS microfeatures, we do not have further detailed information to determine the root cause of some extremely skewed microfeature score distributions. Qualitative approaches such as evaluating the performance of GUMS microfeatures against human annotation can provide more details about the identified issues.

Given the results of this study, we have several recommendations on future directions. One is to implement the alternative microfeature aggregation model for the two assessments in operational practice. The alternative model is easy to implement. Compared to the current model, the proposed model requires an additional step to identify the most predictive microfeatures. However, this step is not always necessary if the most predictive microfeatures are always the same set of microfeatures across different datasets so that they do not need to be identified again. After this step, the proposed model can be built in the same way as the current model is built in operational scoring. The alternative model still keeps all the GUMS microfeatures in the model and does not lose any construct coverage compared to the current model. The second recommendation is to further investigate the microfeatures that have low variance or show unexpected patterns (i.e., GUMS 106, 108, 205, 207, 208, 209, 306, 307, 308, 309, 402, 403, and 406) using NLP techniques and qualitative analyses. Finally, we recommend replications of the analyses on more assessments to determine the generalizability of the findings from this study.

Acknowledgments

The authors of this paper thank Shelby Haberman for his helpful suggestions and feedback for this study. We also thank Patrick Houghton for his careful edits to an earlier version of this paper.

Note

1 Accuracy is defined as the ratio of correctly classified observations to all observations.

References

Almond, S. (2008). *Exploratory analysis of GRE® e-rater® data*. Unpublished manuscript.

Attali, Y. (2007). *Construct validity of e-rater in scoring TOEFL essays* (ETS Research Report No. RR-07-21). Princeton, NJ: Educational Testing Service. http://dx.doi.org/10.1002/j.2333-8504.2007.tb02063.x

Attali, Y., Bridgeman, B., & Trapani, C. (2010). Performance of a generic approach in automated essay scoring. *The Journal of Technology, Learning, and Assessment*, *10*(3), 1 – 15.

Attali, Y., & Burstein, J. (2006). Automated essay scoring with e-rater® v. 2. *The Journal of Technology, Learning, and Assessment, 4*(3), 2 – 29.

Attali, Y., Powers, D. E., Freedman, M., Harrison, M., & Obetz, S. (2008). *Automated scoring of short-answer open-ended GRE Subject Test items* (GRE Board Research Report No. 04 – 02). Princeton, NJ: Educational Testing Service. http://dx.doi.org/10.1002/j.2333-8504.2008.tb02106.x

Beigman Klebanov, B., Madnani, N., Burstein, J., & Somasundaran, S. (2014). Content importance models for scoring writing from sources. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics, Volume 2: Short Papers* (pp. 247 – 252). Baltimore, MD: Association for Computational Linguistics.

Bennett, R. E., & Zhang, M. (2016). Validity and automated scoring. In F. Drasgow (Ed.), *Technology and testing: Improving educational and psychological measurement*. New York, NY: Taylor & Francis.

Burstein, J., Chodorow, M., & Higgins, D. (2007). *Evaluation of* Criterion *feedback codes for sentence checking in FMI's ProofWriter.* Unpublished manuscript.

Burstein, J., Tetreault, J., & Madnani, N. (2013). The e-rater® automated essay scoring system. In Shermis, M. D., & Burstein, J. (Eds.), *Handbook of automated essay scoring: Current applications and future directions* (pp. 55 – 67). New York, NY: Routledge.

Dikli, S. 2006. An overview of automated scoring of essays. *The Journal of Technology, Learning, and Assessment*, *5*(1) 1 – 36.

Educational Testing Service. (2008). Criterion℠ *online writing evaluation service*. Retrieved from http://www.ets.org/s/criterion/pdf/9286_CriterionBrochure.pdf

Foltz, P. W., Laham, D., & Landauer, T. K. (1999). Automated essay scoring: Applications to educational technology. In B. Collis & R. Oliver (Eds.), *Proceedings of World Conference on Educational Multimedia, Hypermedia and Telecommunications* (pp. 939 – 944). Chesapeake, VA: Association for the Advancement of Computing in Education.

Foltz, P. W., Streeter, L. A., Lochbaum, K. E., & Landauer, T. K. (2013). Implementation and applications of the Intelligent Essay Assessor. In M. Shermis & J. Burstein (Eds.). *Handbook of automated essay evaluation* (pp. 68 – 88). New York, NY: Routledge.

Grömping, U. (2006). Relative importance for linear regression in R: The package relaimpo. *Journal of Statistical Software, 17*, 1 – 27.

Haberman, S. J. (2004). *Statistical and measurement properties of features used in essay assessment* (Research Report RR-04-21). Princeton, NJ: Educational Testing Service. http://dx.doi.org/10.1002/j.2333-8504.2004.tb01948.x

Haberman, S. J. (2007*).* Electronic essay grading. In C. R. Rao & S. Sinharay (Eds.), *Handbook of statistics, Vol. 26* (pp. 205 – 233). Amsterdam, The Netherlands: North-Holland.

Heilman, M., Cahill, A., Madnani, N., Lopez, M., Mulholland, M., & Tetreault, J. (2014, June). Predicting grammaticality on an ordinal scale. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Short Papers)* (pp. 174 – 180). Baltimore, MD: Association for Computational Linguistics.

Leacock, C., & Chodorow, M. (2003). C-rater: Automated scoring of short-answer questions. *Computers and the Humanities, 37*(4), 389 – 405.

Lindeman, R. H., Merenda, P. F., & Gold, R. Z. (1980). *Introduction to bivariate and multivariate analysis.* Glenview, IL: Scott, Foresman.

Napoles, C., Cahill, A., & Madnani, N. (2016). The effect of multiple grammatical errors on processing non-native writing. In *Proceedings of the Eleventh Workshop on Innovative Use of NLP for Building Educational Applications* (pp. 1 – 11). San Diego, CA: Association for Computational Linguistics.

Petersen, N. S. (1997, March). *Automated scoring of writing essays: Can such scores be valid?* Paper presented at meeting of the National Council on Education, Chicago, IL.

Quinlan, T., Higgins, D., & Wolff, S. (2009). *Evaluating the construct-coverage of e-rater® scoring engine.* (Research Report No. RR-09-01). Princeton, NJ: Educational Testing Service. http://dx.doi.org/10.1002/j.2333-8504.2009.tb02158.x

Shermis, M.D., & Burstein, J. (Eds.) (2013). *Handbook of automated essay evaluation: Current applications and new directions*. New York, NY: Routledge.

Wainer, H. (1976). Estimating coefficients in linear models: It don't make no nevermind. *Psychological Bulletin, 83*, 213 – 217.

Weigle, S. C. (2010). Validation of automated scores of TOEFL iBT tasks against test indicators of writing ability. *Language Testing*, *27*(3), 335 – 353.

Wilks, S. S. (1938). Weighting systems for linear functions of correlated variables when there is no dependent variable. *Psychometrika*, *3*, 23 – 40.

Williamson, D. M., Bejar, I. I., & Hone, A. S. (1999). Mental model comparison of automated and human scoring. *Journal of Educational Measurement*, *36*, 158 – 184.

Zechner, K., Chen, L., Davis, L., Evanini, K., Lee, C., Leong, C., Wang, X., & Yoon, S. (2015). *Automated scoring of speaking tasks in the Test of English-for-Teaching (TEFT™)*. (Research Report No. RR-15-31). Princeton, NJ: Educational Testing Service. http://dx.doi.org/10.1002/ets2.12080

Appendix. Description of GUMS Microfeatures

Macrofeature	Microfeature code	Brief description of microfeature
Grammar (NSQG) — errors in pronouns, run-ons, missing possessives, etc.	GUMS 101 Fragment	A sentence-like string of words that does not contain a tensed verb or that is lacking an independent clause
	GUMS 102 Run-on sentence	A sentence-like string of words that contains two or more independent clauses without a conjunction or four or more independent clauses joined together or two independent clauses joined with a comma
	GUMS 103 Garbled sentence	A sentence-like string of words that contains five or more errors, or that has an error-to-word ratio > 0.1, or that is unparseable by the Santa module, which organizes words into clauses
	GUMS 104 Subject – verb agreement	A singular noun with a plural verb or a plural noun with a singular verb
	GUMS 105 Ill-formed verb	A mismatch between the tense of a verb and the local syntactic environment; also, use of *of* for *have*, as in *could of*
	GUMS 106 Pronoun error	An objective case pronoun where nominative pronoun is required, or vice versa
	GUMS 107 Possessive error	A plural noun where a possessive noun should be; usually the result of omitting an apostrophe
	GUMS 108 Wrong or missing word	An ungrammatical sequence of words that is usually the result of a typographical error or an omission of a word
	GUMS 109 Proofread this!	An error that is difficult to analyze; often the result of multiple, adjacent errors
Usage (NSQU) — errors in missing or wrong articles, nonstandard verbs, etc.	GUMS 201 Determiner noun agreement	A singular determiner with a plural noun or a plural determiner with a singular noun; use of *an* instead of *a* or vice versa
	GUMS 202 Articles (wrong, missing, extraneous)	An article where none should be used or a missing article where one is required
	GUMS 203 Confused words	Confusion of homophones: words that sound alike or nearly alike
	GUMS 204 Wrong form of word	A verb used in place of a noun
	GUMS 205 Faulty comparisons	Use of *more* with a comparative adjective or *most* with a superlative adjective
	GUMS 206 Preposition error	Use of incorrect preposition, omitting a preposition, or using an extraneous one
	GUMS 207 Nonstandard word form or verb	Various nonwords commonly used in oral language, such as gonna, kinda, etc.
	GUMS 208 Negation error	Instances of *not* or its contracted form, *n't,* followed by negatives such as *no*, *nowhere*, *nohow*
	GUMS 209 Wrong part of speech	Instances with high probability of speech-related grammatical errors, such as "electrically devices"
Mechanics (NSQM) — errors in capitalization, punctuation, commas, hyphens, etc.	GUMS 301 Spelling error	A group of letters not conforming to a known orthographic pattern
	GUMS 302 Capitalize proper nouns	Comparison of words to lists of pronouns that should be capitalized (e.g., names of countries, capital cities, male and female proper nouns, and religious holidays)
	GUMS 303 Missing initial capital letter	Missing initial capital letter in a sentence
	GUMS 304 Missing question mark	An unpunctuated interrogative
	GUMS 305 Missing final punctuation	A sentence lacking a period
	GUMS 306 Missing apostrophe	Detects missing apostrophes
	GUMS 307 Missing comma	Detects missing commas
	GUMS 308 Hyphen error	Missing hyphen in number constructions, certain noun compounds, and modifying expressions preceding a noun

Appendix. Continued

Macrofeature	Microfeature code	Brief description of microfeature
	GUMS 309 Fused words	Fused: an error consisting of two words merged together
	GUMS 310 Compound words	Detection of errors consisting of two words that should be one
	GUMS 311 Duplicates	Two adjacent identical words or two articles, pronouns, modals, etc.
	GUMS 312 Extraneous comma	Comma that is not needed
Style (NSQStyle)—errors in repetition of words, inappropriate words, etc.	GUMS 401 Repetition of words	Excessive repetition of words
	GUMS 402 Inappropriate words or phrases	Inappropriate words. Various expletives
	GUMS 403 Sentences beginning with coordinate conjunction	Too many sentences beginning with a coordinate conjunction
	GUMS 404 Too many short sentences	More than four short sentences (short sentence is defined as fewer than seven words per sentence)
	GUMS 405 Too many long sentences	More than four long sentences (long sentence is defined as more than 55 words per sentence)
	GUMS 406 Passive voice	By-passives: the number of times there occur sentences containing a form of the verb *to be* + a past participle verb form, followed somewhere later in the sentence by the word *by*

Suggested citation:

Chen, J., Zhang, M., & Bejar, I. I. (2017). *An investigation of the e-rater® scoring engine's grammar, usage, mechanics, and style microfeatures and their aggregation model* (Research Report No. RR-17-04). Princeton, NJ: Educational Testing Service. https://dx.doi.org/10.1002/ets2.12131

Action Editor: Beata Beigman-Klebanov

Reviewers: Shelby Haberman and Daniel McCaffrey