APPLIED MEASUREMENT IN EDUCATION
2019, VOL. 32, NO. 02, 116–137
https://doi.org/10.1080/08957347.2019.1577245

Routledge
Taylor & Francis Group

Prediction of Essay Scores From Writing Process and Product Features Using Data Mining Methods

Sandip Sinharay, Mo Zhang, and Paul Deane

Research and Development, Educational Testing Service

ABSTRACT

Analysis of keystroke logging data is of increasing interest, as evident from a substantial amount of recent research on the topic. Some of the research on keystroke logging data has focused on the prediction of essay scores from keystroke logging features, but linear regression is the only prediction method that has been used in this research. Data mining methods such as boosting and random forests have been found to improve over traditional prediction methods such as linear regression in various scientific fields, but have not been used in the prediction of essay scores from keystroke logging features. This article first provides a review of boosting, which is a popular data mining method. The article then applies boosting to predict essay scores from a large number of keystroke logging features and other predictor variables from two real data sets.

Writing assessments, particularly large-scale and high-stakes ones, have generally not been able to provide instructionally useful information for the classroom teacher (e.g., National Research Council, 2001, p. 222). One reason is that, typically, essays are only used to produce scores and not to provide any feedback. The introduction of technology now allows writing assessments to report not only a score but also a description of exactly how students wrote their essays (i.e., a description of the writing process). For example, one student may have written the whole essay at a stretch without any pause, while another may have started a minute late (due to planning in the beginning) and then written the essay in two stretches with one long pause in between. Teachers may be able to use this information to understand if students had difficulty fluently generating text, if they planned before or during writing, where and what they might have edited, whether they have difficulties with typing on the keyboard, and if they read what they wrote before submission (e.g., Berninger, 1999). Summarizing this type of information at the group or individual levels may provide teachers with a suitable starting point for instructional decision making. During instruction, the teachers would be able to show examples of typical writing processes by high performers that they want to encourage or examples of problematic writing behaviors that they want to discourage. In addition, students might benefit from seeing exactly how they composed their essays, helping them to become more reflective with respect to their own writing practice.

Naturally, there is an increasing interest in the study of writing processes, and several research methods have been recently developed to study the writing processes in educational settings. Keystroke logging (KL) is one of those methods (Leijten & Van Waes, 2013). In KL, keystroke activities are logged and time-stamped to reconstruct and describe the writing process. Table 1 gives a simple example of a log of keystroke activities of an examinee who ended up constructing the sentence "It is an apple."

The time-stamps allow investigators to compute several features that are henceforth referred to as KL or process features. For example, one feature is the length of the pause before the examinee

CONTACT Sandip Sinharay ✉ ssinharay@ets.org Research and Development, Educational Testing Service, 660 Rosedale Road, Princeton NJ 08541

Color versions of one or more of the figures in the article can be found online at www.tandfonline.com/hame.

Note: Any opinions expressed in this publication are those of the authors and not necessarily of Educational Testing Service.

Table 1. An example of a log of keystroke activities.

Time stamp	Inter-key interval	Position in text	Key	Operation	Text to date
0	NA	0	NA	Start session	—
0.56	0.56	1	I	Insert	"I"
1.55	0.99	2	t	Insert	"It"
2.39	0.84	3	Space	Insert	"It "
2.55	0.16	4	i	Insert	"It i"
3.22	0.67	5	s	Insert	"It is"
⋮	⋮	⋮	⋮	⋮	⋮
14.25	0.16	14	e	Insert	"It is an apple"
14.99	0.74	15	Period	Insert	"It is an apple."

started typing, with the hope that a short pause is the most ideal as that indicates an initial planning and yet little loss of time (e.g., Zhang, Zou, Wu, Deane, & Li, 2017). The length of the pause before the examinee corresponding to Table 1 started typing is 0.56 seconds.

Prediction of an essay score from KL features has been a topic of substantial interest. Almond, Deane, Quinlan, Wagner, and Sydorenko (2012), Deane (2014), and Zhang and Deane (2015) considered this prediction problem and used linear regression methods to predict essay scores from KL features and other features of the essay. Linear regression methods are well known, easy to fit, and well researched. However, computation-intensive prediction methods, which were impractical even a couple of decades ago but are practical now, may lead to better prediction than linear regression methods in predicting essay scores from KL features and other features.

Several computation-intensive prediction methods are available in *data mining* (e.g., Hastie, Tibshirani, & Friedman, 2009), a field that originated simultaneously in the statistical science and computer science communities. Data mining methods for classification and regression (DMMCR) are becoming increasingly popular in various scientific fields including social sciences, genetics, epidemiology, medicine, and psychology (see, e.g., Strobl, Malley, & Tutz, 2009, and the references therein). The DMMCR often provide more accurate prediction than traditional prediction methods (e.g., Fernandez-Delgado, Cernadas, Barro, & Amorim, 2014), especially in high-dimensional problems where application of traditional prediction methods is problematic (e.g., Strobl et al., 2009). In spite of the popularity of the DMMCR in other scientific fields, these methods are yet to be explored in educational measurement with exceptions such as Gierl (2007) and Sinharay (2016). Further, the availability of a large number of KL features (e.g., Zhang & Deane, 2015) makes the DMMCR likely to outperform linear regression in predicting essay scores. Leijten and Van Waes (2013) and Oranje, Gorin, Jia, and Kerr (2017) mentioned the possibility of applying data mining analyses to data from keystroke logs.

The purpose of this article is to explore the use of the DMMCR in predicting essay scores from KL features and other features of an essay. Research related to prediction of essay scores from KL features has several potential applications. The most important of them is to help identify the features that are the most relevant to writing quality and/or distinctive between subgroups of writers. Identification of such features would be useful in, for example, the writing assessment in the National Assessment of Educational Progress that involves the use of several indicators related to writing processes to report patterns of writing behavior at a group level (e.g., National Center for Education Statistics, 2011). There may also be applications where such features could help improve the prediction accuracy of writing quality, in cases where the assessed construct includes the writing process, or where differences in the writing process affect the predictive value of other features.

A review of writing processes and KL is provided in the next section. Two real data sets from a writing assessment are described in the following section; the data sets are described early in this article because they are used to describe important ideas in the following section that focuses on the description of one of the most popular DMMCR. It is demonstrated in the Application section that boosting, which is one of the most popular DMMCR, leads to some improvement over linear regression for the data sets from the writing assessment. Conclusions and recommendations are provided in the last section.

A Review of Writing Processes and Keystroke Logging

Even though there is a vast amount of literature on literacy and writing research from both the theoretical and practical perspectives (e.g., Medimorec & Risko, 2017), the study of writing processes appears to be much rarer. The analysis of writing process can be dated back to at least the 1970s (e.g., Emig, 1972; Flower & Hayes, 1981; Stallard, 1974). However, it is not until the recent decade or two when a growing number of studies including a considerable extent of theoretical work on the behavioral and cognitive processes of writing (e.g., Hayes, 2012; Kellogg, 1996; McCutchen, 1996) were published on writing processes. The study of writing processes has been on the upswing in recent years as digitally based writing has become popular, even ubiquitous, in modern life. With the advance in technology, the composition process can be directly observed and precisely reconstructed with the use of KL. Research on writing process is facilitated with this “new” KL information. InputLog (Leijten & Van Waes, 2013) is one example of KL programs/systems that has been widely used in a number of studies (Leijten, Janssen, & Van Waes, 2010; Leijten, Macken, Hoste, Van Horenbeeck, & Van Waes, 2012). In KL, the keyboard activities are recorded and time-stamped from which the entire process of text production can be precisely reconstructed. Essentially, a KL system records the type of a keystroke/action (e.g., insert, delete, cut, paste, replace) and information on how long the action lasts (e.g., continuous keystroke insertion for 10 seconds) and when and where it occurs in the writing process (see, e.g., Table 1). The recorded information is intended to provide useful information on the writer. For instance, (a) a pause between two paragraphs is likely to indicate idea planning whereas a pause in the middle of a word may be indicative of effort in spelling retrieval or word finding, and (b) the analysis of the processes of typo/spelling error corrections may reveal a writer’s editing style. Some writers may make the corrections immediately while some may correct the errors altogether in the end. These behavioral patterns and time-stamps allow investigators to extract and compute features/measures to describe a writing process.

In addition to precisely reconstructing the temporal and mechanical process of text production, another benefit of KL lies in the possibility of the quantification of separate theoretical components of a writing process. As an example, the four sub-processes of writing proposed in the cognitive model of Hayes (2012)—proposer, transcriber, translator, and evaluator—can potentially be estimated separately and, in turn, used to give instructional and learning feedback. For example, Zhang and Deane (2015) analyzed 29 features extracted from KL and found that the features can be meaningfully grouped into four process indicators that they labeled as “Fluency,” “Word and Local Level Editing,” “Phrasal and Chunk Level Editing,” and “Deliberation” and discussed the connections of these indicators to the Hayes theoretical model. In practice, students may have been struggling with one or more subprocesses of writing but proficient in others. The information extracted from KL may allow teachers to plan more focused instructions on the aspects that the students have trouble with. In essence, evaluations conducted on the quantified KL measures may lead to useful inferences on the efficiency and performance of one writer or a group of writers.

A few studies have been conducted to understand and evaluate the KL features in the assessment context. Previous research has suggested that writing time and number of keystrokes, which are indicative of general writing fluency and efforts, are related to writing quality (Allen et al., 2016; Zhang, Hao, Li, & Deane, 2016). Another feature that is of considerable interest is the length of the pause before the examinee started typing. Zhang et al. (2017) found that, under a certain timed-writing test condition, a shorter pre-writing pause is preferred as that indicates adequate understanding of the task requirements, more familiarity with the writing topic, and better task planning. In addition, writing burst, which is a sequence of rapid text production without interruptions (i.e., without long pauses), is one of the most important writing-process features discussed in the psycho-linguistic literature. Burst length and variations, among others, have been reported to show predictive power of essay quality in several timed-writing assessments (e.g., Deane & Zhang, 2015).

Almond et al. (2012), Deane (2014), and Zhang and Deane (2015) considered the prediction of essay scores and used linear regression to predict essay scores from KL features and other features on the essay. Almond et al. (2012) found that a few KL features related to writing bursts added some value over eight e-rater® (Attali & Burstein, 2006) features in one linear regression, but did not add any value in two other linear regressions. Deane (2014) performed linear regression of essay scores on both KL and e-rater features and found the regression coefficients corresponding to the KL features to be significant in most linear regressions. Zhang and Deane (2015) performed a linear regression of essay scores on (a) only the KL features and (b) both KL and e-rater features; they found that KL features added only a little value over e-rater features.[1] Zhang and Deane (2015) found linear regression using KL and e-rater features as predictors to explain only about 70% of variation of essay scores; potentially, a more complex prediction method might be able to explain a part of the 30% variation that is not explained by linear regression. Therefore, there is a scope of further research on prediction of essay scores using KL features and other essay features using methods more complex than linear regression.

Data From a Writing Assessment

Deane and Zhang (2015) and Zhang and Deane (2015) analyzed data that were collected from a multistate sample of 6th- to 9th-grade students who were part of a larger pilot test of CBAL® (e.g., Bennett, 2011) summative writing assessments in 2013. Schools were recruited for participation on a voluntary basis. Together, the students took six test forms. Each form started with a preliminary, or lead-in, section that required students to read, think, and respond to questions about a set of source documents on a scenario. In the second section, the students were required to complete an essay task on the same scenario using the same source reading documents as in the lead-in section. Each section was administered in a single 45-minute class session for a total of 90 minutes of testing time. The test scores were not used to make any decisions about the students, teachers, or schools, so the test can be considered low stakes for the examinees and the schools.

We consider data from two of the six test forms—the Generous Gift (GG) and Culture Fair (CF) forms, both of which have an emphasis on persuasive writing. The GG form presents a scenario asking students to recommend the best way for a school to spend a large sum of money provided by a generous donor; the CF form presents a scenario asking students to recommend the best theme for a school cultural fair. The sample sizes were 825 and 832, respectively, for the GG and CF forms. To ensure adequate quality of the data, a small percentage of records were excluded for reasons such as incomplete KL data and very short essay length (less than 25 words). The resulting sample sizes were 823 and 825, respectively.

Demographic data were available for 81% of the students. Among them, 49.4% were female and 50.6% were male; 68.4% were White, 22.9% were Hispanic, 4.6% were African American, 3.5% were Asian, and less than 1% belonged to any other group; 92.7% were initially English proficient, 4.3% were reclassified English proficient, and 2.9% were English language learners; 40.7% qualified for free or reduced school lunch programs.

All essays were scored by at least two human raters on a rubric evaluating basic text quality; each rater provided an integer-valued holistic score on a scale ranging from 0 to 5 or assigns a condition code of 9 for a blank response. The scoring rubric is provided in Appendix A. If the scores of the two raters disagreed by more than two points, a third rater was asked to adjudicate.

Table 2 shows some summary statistics of the human ratings for the two forms; it shows the average and standard deviation (*SD*) of the ratings and the quadratically weighted kappa (QWK), percent exact agreement (*Exact %*), and percent adjacent agreement (*Adjacent %*) between

[1] Both Deane (2014) and Zhang and Deane (2015) first performed factor analysis on the KL features to compute values of several factors from these features and used those factors in their regression analysis.

Table 2. Summary of the human ratings.

Form	Average	*SD*	QWK	Exact %	Adjacent %
GG	2.51	0.82	0.66	52	96
CF	2.56	0.96	0.64	47	94

the raters. The raters used all five points in the scale. The percentage of ratings that were equal to 1, 2, 3, 4, and 5 are 10, 45, 32, 11, and 2, respectively, for the GG form, and 9, 43, 33, 12, and 3, respectively, for the CF form.

The adjudicated essay score, which is the average of the score of two raters if they disagreed by up to two points and the score adjudicated by the third rater if they disagreed by more than two points, was used as a response variable in the analyses later in this article.

Keystroke logs on the essay tasks of all the students were collected. Several KL features were extracted from the keystroke logs using the KL engine developed at the Educational Testing Service. The features that were used by Almond et al. (2012), Li, Zhang, and Deane (2016), and Zhang and Deane (2015) were used as KL features or writing-process features (or *process features* in short) in our analyses. The *product features* on all essays were obtained using the e-rater automated scoring system (e.g., Attali & Burstein, 2006) and were used in our analysis.

The product and process features employed in this study are described in Tables 3 and 4, respectively.

Some of the process features are intended to be indicators of fluency in text production; these are primarily based on pause patterns (or, conversely, bursts of text production). Examples include the median value of the longest within-word pause time across all words (shown in the penultimate row of Table 4), indicating the extent to which there is a clear hesitation (or lack of burst); the median value of the interkey pause time across all keystrokes (13th row of Table 4), implicating an overall keyboarding fluency; and the mean value of between-word pause time across all pairs of words (first row), suggesting general typing fluency. A linear combination of several fluency features was found to predict the essay scores well in an application of linear regression by Zhang and Deane (2015). Some other KL process features shown in Table 4 measure the extent of editing and revision. These features include the proportion of corrected or uncorrected typographical errors among the total number of words typed in the composition process (rows 6 and 31 of Table 4), and the proportion of the total time spent on multiword deletion (row 18). Still other KL process features shown in Table 4 intend to provide measures of the extent of planning and deliberation. Examples include the proportion of time spent at the start of phrasal bursts (row 26) and the median value of pause length across all sentence boundaries (row 10); each of these measures might indicate the time devoted to planning and deliberation in between bursts and sentences, respectively.

Leijten and Van Waes (2013) commented that the main rationale behind studying KL is that writing fluency and tempo may reveal traces of the underlying cognitive processes, which explains the focus on pause and revision characteristics. Thus, inclusion of the above categories of KL features in our study is justified.

Table 3. Description of writing product features.

Product feature	Description
Grammar	Extent of absence of grammatical errors (e.g., pronoun, run-on)
Mechanics	Extent of absence of mechanical errors (e.g., capitalization, comma)
Usage	Extent of absence of word usage errors (e.g., missing article)
Style	Extent of absence of stylistic errors (e.g., word repetition)
Vocabulary sophistication	Median word frequency measured by standard frequency index
Word length	Average number of characters per word
Syntactic variety	Diversity of syntactic structure of the sentences
Development	Number of discourse units
Organization	Average length of the discourse units
Collocation-preposition	Extent of correct use of everyday English vocabulary

Table 4. Description of writing process features.

Process feature	Description
Between-word pause	Mean duration of pauses between words
Burst length	Mean duration of pauses between bursts
Burst number	Number of bursts
Burst variation	Standard deviation of durations of pauses between bursts
Characters in multiword deletion	Number of character deletions occurring in the process of deleting multiple words divided by the sum of inserted and deleted characters
Corrected typos	Number of corrections of mistyped words divided by the total number of words produced (including words that were later deleted)
Deleted characters	Number of deleted characters divided by total number of characters inserted or deleted
Discarded text	Number of deleted characters divided by the total number of characters in the final submission
Edited chunk	Length of deleted text replaced with edited text of similar content divided by the total number of keystroke actions
End sentence punctuation pause	Median pause time at sentence junctures
Event after last character	Number of keystrokes not occurring at the end of the text divided by the number of keystroke actions
In-sentence punctuation pause	Median time duration spent on the pauses occurring at a within-sentence punctuation mark
Interkey pause	Median duration of pauses between keystrokes
In-word pause	Mean duration of pauses inside of words
Long jump	Number of jumps to a different part of the text occurs divided by total number of keystrokes
Major edit	Number of words that are edited to make more than a one- or two-character change divided by the total number of words produced (including words that were later deleted)
Minor edit	Number of words that are subjected to minor editing (one- or two-character changes) divided by the total number of words produced (including words that were later deleted)
Multiword deletion	Number of multiword deletions divided by the total number of words produced (including words that were later deleted)
Multiword edit time	Time spent in deleting multiple words divided by total writing time
New content	Length of deleted text replaced with editing text with different content divided by the number of keystroke actions
Phrasal burst	Length of the longest string of fluent text production that is produced without interruption
Prejump pause	Median pause duration occurring just before jumping to a different part of the text
Retyped chunk	Length of deleted text replaced with essentially the same text divided by the number of keystroke actions
Start time	Time spent pausing before the first writing event divided by total writing time
Time on task	Duration of total writing time
Time spent at phrasal burst	Time spent on pauses occur at the beginning of a string of fluent text production divided by the total writing time
Time spent between phrasal burst	Time spent on pauses that are followed by other pauses (rather than by sequences of fluent text production) divided by the total writing time
Typing speed	Total number of keystrokes divided by total writing time
Typo corrected chunk	Length of text replaced with edited text that differs only in minor spelling correction divided by the number of keystroke actions
Typo correction rate	Number of typos that are corrected divided by the total number of uncorrected spelling errors
Uncorrected spelling errors	Number of times that a spelling error occurs that is not corrected before the writer starts typing another word divided by the total number of words produced (including words that were later deleted)
Word choice	Median duration of the longest pauses when words are edited to produce completely different words
Word choice event pause	Number of words that are edited to produce completely different words divided by the total number of words produced (including words that were later deleted)
Word edit pause	Longest pause occurring within words during text editing
Word final pause	Median duration of the pauses that occur just before typing the last character in a word
Word initial pause	Median duration of the pauses that occur just before typing the first character in a word
Word internal pause	Median duration of longest pause that occurs within words during text production
Word space pause	Median duration of pauses that occur before the space character that separates two words

Ten e-rater (Attali & Burstein, 2006) or product features that measure vocabulary complexity, discourse organization, accuracy of grammar and mechanics, and style in terms of syntactic variety and word use, among others, were used in our analyses. Note that in electronic essay scoring, including in applications of e-rater (e.g., Attali & Burstein, 2006), one fits a linear regression model to predict the essay score from the product features. Such an approach using the 10 features in Table 3 led to percent exact agreement between the human and predicted scores of 55% and 45% for the GG and CF forms, respectively, and QWK of 0.67 and 0.65, respectively.

Table 5 provides the summary statistics of the adjudicated essay score and the features for the GG form— the statistics for the CF form were similar and are not shown. Rows 2–11 show summaries for the product features while the latter rows show summaries for the process features. Columns 2–5 show, for each feature, its mean, standard deviation, correlation coefficient with the adjudicated essay score, and partial correlation coefficient with the adjudicated essay score. The partial correlation coefficient is an important number here because it indicates the strength of a feature to predict the part of the essay score that is not predicted by any other feature. The table shows that most of the

Table 5. Summary statistics of the product and process features.

Feature	Mean	Standard deviation	Correlation	Partial correlation
Essay score	2.510	0.815	—	—
Grammar	−0.088	0.037	0.17	0.17
Mechanics	−0.221	0.091	0.40	0.18
Usage	−0.084	0.060	0.06	0.13
Style	−0.355	0.132	0.53	0.04
Vocabulary sophistication	−63.5	2.758	0.20	0.09
Word length	4.377	0.313	0.28	0.17
Syntactic variety	2.749	0.566	0.63	0.09
Development	3.641	0.377	0.27	0.29
Organization	1.474	0.508	0.66	0.37
Collocation-preposition	0.554	0.144	0.15	0.04
Between word pause	4.484	0.347	−0.36	−0.07
Burst length	1.452	0.413	0.38	−0.01
Burst number	172.9	94.07	0.49	−0.03
Burst variation	0.923	0.159	0.32	−0.03
Characters in multiword deletion	0.120	0.175	0.00	0.00
Corrected typos	0.007	0.004	0.11	−0.13
Deleted characters	0.342	0.093	0.09	−0.10
Discarded text	0.246	0.359	−0.02	0.03
Edited chunk	0.023	0.022	0.19	−0.05
End sentence punctuation pause	1.419	0.378	0.15	0.01
Event after last character	0.958	0.071	−0.11	0.00
In-sentence punctuation pause	1.652	1.032	0.27	−0.02
In-word pause	3.374	0.278	−0.36	0.02
Interkey pause	0.297	0.094	−0.32	0.05
Long jump	0.014	0.036	0.02	0.13
Major edits	0.000	0.001	0.04	−0.06
Multiword deletion	0.069	0.028	0.19	−0.01
Multiword edit time	0.048	0.040	0.14	−0.02
Minor edits	0.017	0.008	0.03	−0.04
New content	0.002	0.001	0.15	−0.07
Phrasal burst	15.79	6.87	0.37	0.03
Prejump pause	2.561	1.092	0.10	−0.01
Retyped chunk	3×10^{-5}	10^{-5}	0.06	0.17
Start time	0.346	0.106	−0.36	−0.01
Time spent at phrasal burst	0.280	0.101	0.01	0.01
Time spent between phrasal burst	0.330	0.120	0.07	0.02
Time on task	1051	501	0.52	0.07
Typing speed	1.310	0.470	0.39	0.03
Typo corrected chunk	2×10^{-4}	10^{-4}	0.12	−0.15
Typo correction rate	−0.638	0.660	0.16	0.02
Uncorrected spelling errors	0.014	0.007	−0.18	−0.16
Word choice	0.015	0.007	0.17	0.00
Word choice event pause	2.689	6.145	−0.13	−0.01
Word edit pause	0.799	0.135	−0.33	0.00
Word final pause	0.452	0.068	−0.40	−0.04
Word initial pause	0.658	0.128	−0.38	0.00
Word internal pause	0.575	0.091	−0.38	−0.04
Word space pause	0.440	0.056	−0.33	−0.06

product features have a large positive correlation with the essay score and the features that have the largest partial correlation with the essay score are product features (such as organization and development). Among the process features, the time on task has the largest correlation (0.52) with the essay score and several others have a correlation that is modest in absolute value (0.3 or larger).

Figure 1 shows a plot of the correlations among the features. The figure was created using the package *corrplot* (Wei & Simko, 2017) within the R software (R Core Team, 2017). The features were sorted to show clusters of correlated features. Abbreviated names of the features are shown at the left and the top of the figure. A larger square for a pair of features indicates a correlation between that pair that is large in absolute value. Black and white squares indicate positive and negative correlations, respectively. There are large black squares along the diagonal because each variable is correlated perfectly with itself. The cluster of large black squares above and to the left of the middle of the figure correspond to eight process features related to pauses (that are indicators of fluency according to an earlier discussion)—so the high positive correlation among them is expected. The cluster of large black squares below and right of that cluster include several process features (amount of discarded text, etc.) related to deletion. Toward the bottom right, a cluster of black circles correspond to five features, three of which are product features (organization, style, and syntactic variety) and two are process features (time on task and number of bursts). The cluster of large white squares (indicating negative correlations) toward the left of the figure—between eight pause-related process features and three other process features—indicate that, for example, those who type fast and write in long bursts typically use short pauses.

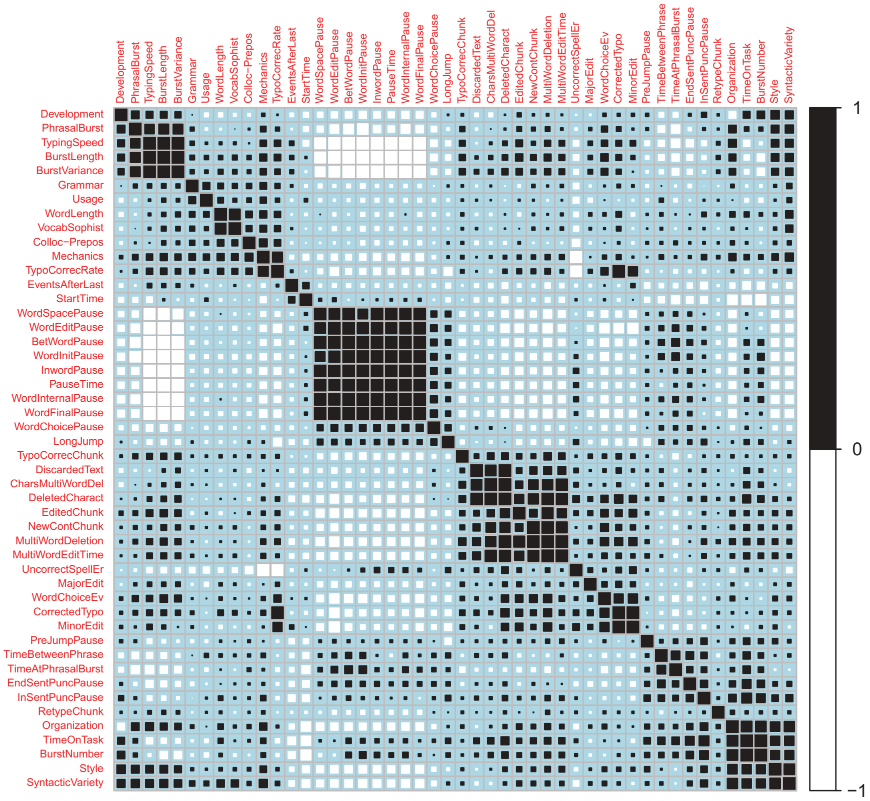

Figure 1. The correlations among the predictors.

Data Mining Methods for Classification and Regression

Several DMMCR including boosting are extensions of classification and regression trees (CARTs; Breiman, Friedman, Olshen, & Stone, 1984), which are described below.

Classification and Regression Trees

CARTs are nonparametric techniques in which no assumptions are made about the relationship between the response and the predictors (e.g., Strobl, 2013). CARTs include *regression trees* that are used in regression problems and *classification trees* that are used in classification problems. We focus on regression trees, which will be used in this article, in the following discussion. Regression trees have been applied in educational measurement by, for example, Sheehan and Mislevy (1994), for predicting the estimated difficulty of test items from the item characteristics.

Construction of a Regression Tree

In an application of a regression tree, the observations in the data set are repeatedly split into two subsets at a time so that the values of the response variable are as similar as possible for all observations within each subset, or equivalently, as different as possible between subsets. For example, a regression tree for predicting the essay scores only from the KL features for the GG test form is provided in Figure 2. The tree was drawn using the R packages *rpart* (Therneau, Atkinson, & Ripley, 2017) and *rpart.plot* (Milborrow, 2016).

In the figure, the symbols *StartTime, InSentPuncP, WordFinalPa, PhrasalBurs*, and *EventAfterL* respectively, denote the process features *Start time, In-sentence punctuation pause, Word final pause, Phrasal burst*, and *Event after last character* listed in Table 4. The figure shows that in the first step of the tree-building process, all the essays were split into two subsets, one subset including the essays with *Start Time* $\geq$ *0.42* constituting the branch on the left and another subset including the essays with *Start Time* $<$ *0.42* constituting the branch on the right. The numbers 1.9 (top) and 22% (bottom) inside the box on the left branch communicate that the average human score of the essays with *Start Time* $\geq$ *0.42* is 1.9 and these essays constitute 22% of all the essays in the data set. The right branch includes 78% essays and their average score is 2.7. The variable *Start Time* denotes the length of the pause before beginning writing, so this branching indicates that those who spend a short time before starting to write tend to receive a higher score on average. This finding agrees with a finding in Zhang et al. (2017). The essays on the left branch of the tree are further split into two subsets, one including the essays with *In-sentence Punctuation Pause* $<$ *0.27* (to the left) and the other including those with *In-sentence Punctuation Pause* $\geq$ *0.27* (to the right). The variable *In-sentence Punctuation Pause* denotes the extent to which pauses occur at a (natural) punctuation mark inside a sentence, which may suggest efforts spent on planning and deliberation. This branching suggests that those who spend longer on pausing at natural junctures tend to receive higher essay scores (2.1 vs. 1.5 on average).

At each step during the construction of the tree, all possible splits on the basis of all possible predictor variables are considered. Possible splits are compared to each other in terms of the residual sum of squares[2] (RSS; Hastie et al., 2009). The best split is the one that produces the largest difference between the RSS of the original set and the sum of the RSS in the two potential subsets, or, equivalently, minimizes the sum of the RSS in the two potential subsets. Typically, the predictor variable that is used in a split is the one that predicts the response variable the best and hence can separate the small values of the response from the large ones. Thus, for example, if a sample of 200 essays includes a subset of 100 essays with scores of 1 or 2 and *Start Time* larger than 1 minute and another subset of 100 essays with scores of 4 or 5 and *Start Time* smaller than 1 minute, then the best split would use *Start Time* as the predictor variable and would result in the essays being divided into these two subsets. The tree-building

[2] For a set of values $y_1, y_2, \ldots, y_n$ of a response variable, if $\hat{y}_1, \hat{y}_2, \ldots, \hat{y}_n$ denote the corresponding predicted values, then the residual sum of squares is equal to $\sum_i (y_i - \hat{y}_i)^2$.

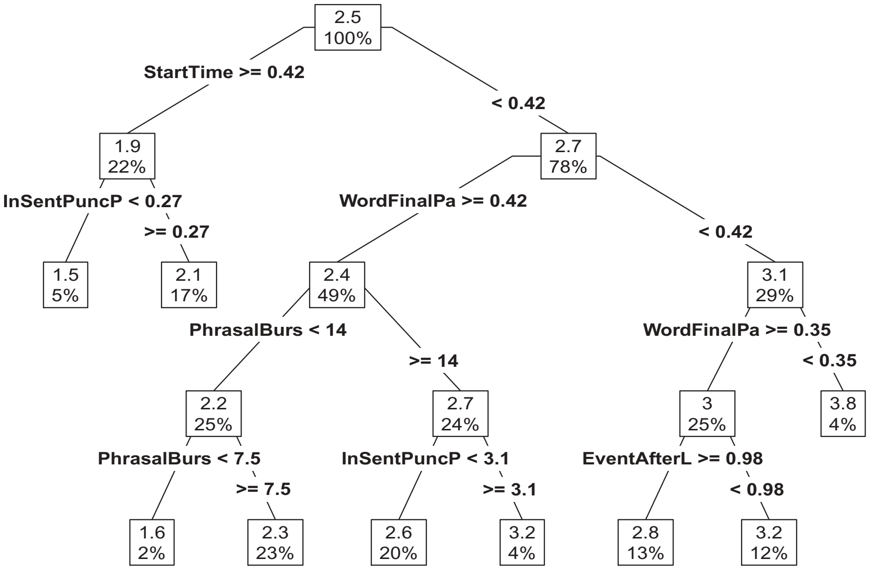

Figure 2. A regression tree for predicting essay score from process features.

algorithm is often referred to as *greedy* because at each step of the algorithm, the best split is made at that particular step rather than planning ahead; thus, the algorithm does not pick a split that would lead to a better tree in some future step (James, Witten, Hastie, & Tibshirani, 2013, p. 306); this aspect of the tree-building process is like forward stepwise linear regression in which, at any step, the procedure adds the predictor that leads to the largest improvement in prediction rather than planning ahead. For each split of the tree, the original set of observations is referred to as the *parent node* and the two subsets as the left and right child nodes. The node-splitting process continues until reaching a predetermined criterion that specifies when the process would stop. Examples of such criteria are (a) stopping after a predetermined number of splits, (b) stopping when the number of observations left in a node is smaller than a predetermined value, and (c) stopping when further splits would not substantially reduce the RSS (e.g., Strobl, 2013). In the final tree, the nodes without a child are called *terminal nodes*. The tree in Figure 2 includes nine terminal nodes that divide the essays into nine groups of varying average scores; for example, the leftmost group includes the essays with *Start Time* $\geq$ *0.42* and *In-sentence Punctuation Pause* $<$ *0.27* and have an average score of 1.5 and the rightmost group includes those with *Start Time* $<$ *0.42* and *Word final pause* $<$ *0.35* and an average score of 3.8; roughly, the average scores of the groups increase as one moves from the left of the tree to the right in Figure 2.

Computation of Predicted Values From a Regression Tree

To obtain the predicted value of the response of an individual from the final fitted tree, one first determines the terminal node that corresponds to the predictors of the individual. Then the predicted value for the individual from the regression tree is obtained as the average of the responses of the individuals belonging to that terminal node. For example, Figure 2 shows that the terminal node that corresponds to the predictors of an essay with *Start Time* $\geq$ *0.42* and *In-sentence Punctuation Pause* $<$ *0.27*, and any values of the remaining predictors is the first terminal node from the left and that the predicted value of the score of such an essay is 1.5.

The Need to Explore beyond Regression Trees

While the CARTs have the advantage of requiring no assumptions about the relationship between the response and the predictors or about the distribution of the variables, which often makes them advantageous for nonlinear problems, a problem with CARTs is their high variance or instability (e.g. Gey & Poggi, 2006). A small change in the data often leads to a very different series of splits and hence to a regression tree that looks completely different from the original one. This lack of stability may be overcome by an extension of CARTs that involves constructing several regression trees instead of one regression tree and combining their predictions into a single prediction. Boosting, which is one of the most popular DMMCR, implements this idea of using several trees. Boosting and other DMMCR that use several trees retain the trees' virtue of providing satisfactory prediction for nonlinear problems. An application of such a method is similar to employing a committee of experts for prediction; a committee of experts is more likely to make a correct prediction than a single expert.

Boosting

Tukey (1977) suggested a smoothing procedure referred to as *twicing*; in the first step of the procedure, a regression model is used to predict a response variable from several predictors and the residuals are computed from the regression; in the second step, a regression model is used to predict these residuals from the same set of predictors. The final prediction is obtained as the sum of the predictions from the two steps; that is, the final prediction is a combined prediction from two regressions. Tukey (1977) mentioned the possibility of iterating the process.

Boosting is a refined version of twicing and combines predictions from several trees. In boosting, the trees are constructed sequentially and each tree is constructed using information from previously grown trees, just like in *twicing*—the construction process of a tree places an increased emphasis on observations that were poorly modeled by the previously constructed tree. Boosting algorithms vary in how they determine the observations that are poorly modeled and how they select settings for the steps in the algorithm. Stochastic gradient boosting (Friedman, 2001, 2002) is arguably the most popular among the existing boosting algorithms and was used in our study.

In the context of regression, one starts the stochastic gradient boosting procedure with a regression tree that is rather small, with just a few terminal nodes (James et al., 2013, p. 322). Such a tree can be obtained by, for example, splitting the data into two subsets based on one predictor. Figure 2 shows that a small tree for the GG form can be obtained by splitting the examinees into two subsets based on the *Start Time* feature. Then one computes residuals from the small tree and fits a new regression tree, again small, to these residuals. The process of computing residuals and fitting small trees continues for a large number of times. By fitting small trees to the residuals, one slowly improves the prediction in areas where prediction till then has not been satisfactory. The final boosting model involves a linear combination of all the fitted trees (several hundreds to several thousands)—the model can be thought of as a linear regression model where each term in the regression equation corresponds to a tree (e.g., Elith, Leathwick, & Hastie, 2008).

To obtain a predicted value for a new observation from the fitted boosting model, each tree (that was fitted by the boosting algorithm) is used to obtain a predicted value for the observation. The final predicted value is then calculated as the average of these predicted values multiplied by a *shrinkage parameter* that is typically a positive number considerably smaller than 1—the multiplication by the shrinkage parameter effectively shrinks the contribution of each tree to the final predicted value and avoids over-fitting the observations (Elith et al., 2008).

The stochastic gradient boosting algorithm also includes a random/stochastic component by the way of using only a random subset of the data to construct each tree. Research has shown that random subsets that are about half as large as the data set lead to the most accurate prediction (Elith et al., 2008).

Thus, the application of boosting is like employing a committee of experts where Expert 2 learns from the mistakes of Expert 1 (or, to try harder to predict the observations incorrectly predicted by Expert 1), Expert 3 learns from the mistakes of Expert 2, and so on.

Stochastic gradient boosting can be implemented using the R packages (Kuhn, 2008) *tm* (Feinerer, Hornik, & Meyer, 2008) and *gbm* (Ridgeway, 2017). Among them, the *gbm* package has been found satisfactory in several disciplines by, for example, Elith et al. (2008), James et al. (2013), and Sinharay (2016)—so this package is used in this article.

To apply boosting to a data set, the investigator has to fix the values of three *tuning parameters*—the exact number of trees (n_t), the shrinkage parameter λ, and the interaction/tree depth (that is a measure of how large a tree is) d (e.g., James et al., 2013, p. 323). Each software package has a default value of these parameters. For example, the default values in the *gbm* package in R are 100, 0.001, and 1, respectively. However, the quality of the prediction obtained from boosting can vary over different choices of these parameters and there is no guarantee that the default values in a software package will lead to the best prediction. Therefore, a prudent strategy is to apply boosting with different combinations of values of these parameters and choose the combination that leads to the best prediction. Such an approach is used later in this article during the application of boosting to the data from the writing assessment.

Existing Applications of DMMCR for Predicting Essay Scores

Sinharay (2016) found that random forests and boosting did a better job at predicting essay scores from product features compared to linear and logistic regression models, especially for small sample sizes, for data from a state test. Chen, Fife, Bejar, and Rupp (2016) applied random forests, support vector regression (SVR), and k-nearest neighbor methods to predict human scores from e-rater (Attali & Burstein, 2006) product features and found that SVR marginally outperformed linear regression, but the other two DMMCR performed worse overall than linear regression.

Prediction of Essay Scores for the Writing Assessment

Analysis

Boosting and linear regression were used to predict the adjudicated essay score using data from the GG and the CF forms separately. The adjudicated essay score is on a discrete scale—it can only take one of the nine values 1, 1.5, $\cdots$, 4.5, 5. However, the number of possible values (9) is modestly large; Agresti (2013, p. 327) noted that statistical models for predicting ordinal variables are approximated well by the linear regression model when the number of possible values of the ordinal variables becomes large. In addition, one goal of this article is to compare the results from DMMCR to those from linear regression (Deane, 2014; Zhang & Deane, 2015). Therefore, we employed a linear regression model and boosting using regression trees to predict the adjudicated essay score under the assumption that the latter is a continuous variable.[3]

For each test form, the following steps were replicated 100 times:

- The data set was randomly split into a larger subset of about two thirds of the total number of essays and a smaller subset with the remaining one third of the essays.
- The prediction model for boosting was built from the larger subset. The linear regression model was fitted using the stepwise procedure[4] on the larger subset.
- The prediction models for boosting and linear regression were used to predict the score of each essay in the smaller subset. Predicted values smaller than 1 and larger than 5 were reset to 1 and 5, respectively.
- Three accuracy/error measures were computed from the smaller subset. The first two accuracy measures are the *correlation coefficient* between the actual and predicted score and the *root mean squared difference* (RMSD) between the actual and predicted scores. The third accuracy measure is the *percent exact agreement* between the actual scores and the rounded predicted scores where the rounding was to the nearest half.[5]

[3] A cumulative logit model (e.g., Agresti, 2013) could instead be used to predict the discrete adjudicated essay score. A cumulative logit model and a linear regression model performed very similarly in predicting the adjudicated essay score in some limited analysis (not described here). The rest of this article does not consider the cumulative logit model.

[4] No use of stepwise procedure led to a worse prediction compared to the use of stepwise regression.

[5] Thus, for example, 2.21 is rounded to 2.0 and 2.57 is rounded to 2.5. Note that this rounding allows the possible values of the predicted scores (1, 1.5, 2, 4.5, 5.) to be the same as those of the actual scores.

The average of each accuracy/error measure was computed from the 100 replications. Boosting was performed using the R package *gbm* (Ridgeway, 2017). Linear (stepwise) regression was performed using the R function *stepAIC*—predictor variables were allowed to be included or excluded in each step.

The strategy of fitting the model on a subset of the data and computing the accuracy/error measure on another subset of the data, which is used here, is common in applications of data mining (e.g., James et al., 2013, p. 30). The two subsets are often referred to as *training/model-building set* and *test/holdout/evaluation set*, respectively. The idea is similar to that of cross-validation in the context of linear regression (e.g., Deane & Zhang, 2015; Kutner, Nachtsheim, Neter, & Li, 2004, p. 372) and is used because of the fact that DMMCR often overfit the data, leading to accurate prediction in the model-building data set, but make inaccurate prediction for new and unseen data. Thus, for the DMMCR, accurate prediction in the test set is a more desirable property than accurate prediction in the model-building set. In applications of DMMCR, the method that leads to the most accurate prediction in the test set is typically preferred over other methods (e.g., James et al., 2013, p. 30).

The following three sets of predictor variables were used in three separate analyses with each of boosting and linear regression:

- the KL or process features (described in Table 4)
- the e-rater or product features (described in Table 3)
- both the process and product features

The R codes for one replication of the above steps for the case when both the process and product features are included as predictors and for one form are included in Appendix B. The computation of only one accuracy measure—the RMSD—is shown in the codes. Note that the codes represent the case where the data set includes the adjudicated essay score and the process and product features for one form.

Results

Results on the choice of the tuning parameters, comparative performance of linear regression and boosting, and variable importance are discussed below.

Choice of the Tuning Parameters

We set the three tuning parameters in the *gbm* package to different combinations of values and computed the corresponding accuracy measures (average correlation coefficient, RMSD, and percent exact agreement) from test sets from 100 replications in the manner described above. Figure 3 shows the values of the RMSD for different combinations of values of the tuning parameters when both the process and the product features were used to predict the essay scores. Each panel in the figure corresponds to a number of tree (n_t), which was varied between the values 500, 1,000, 2,000, and 5,000. In each panel, the RMSD is plotted against the interaction/tree depth (d), where the latter was varied between 1, 2, 3, 5, 8, and 12. Each dotted line that joins a specific type of plotting symbol (a plus, a multiplication, or a diamond) corresponds to a fixed value of the shrinkage parameter (λ). For example, in the top left panel (that corresponds to $n_t = 500$), the dotted line that joins the plus symbols corresponds to the values of RMSD for different values of interaction/tree depth when the shrinkage parameter (λ) is fixed at 0.005 (see the legend in each panel). One more value of λ—0.0001—was also used in this analysis, but the RMSDs for this value were much larger than those shown in Figure 3, so the RMSDs for $\lambda = 0.0001$ are not shown in the figure. The values of RMSD for $\lambda = 0.001$ for 500 trees were all larger than 0.6, so they are not seen in the top left panel; in contrast, the values of RMSD for $\lambda = 0.001$ for 5,000 trees are all small—this result supports the result (e.g. Elith et al., 2008) that small shrinkage parameters, combined with a large number of trees, lead to satisfactory prediction. The figure shows that while the RMSD decreases with an increase in the interaction/tree depth for some values of n_t and λ (such as 500 and 0.005, respectively), the RMSD remains unaffected by some other values of these tuning parameters (such as 500 and 0.05, respectively). In general, no gain is achieved by increasing the tree depth beyond 5. The number of trees does not seem to affect the prediction quality on average. The figure also shows that the

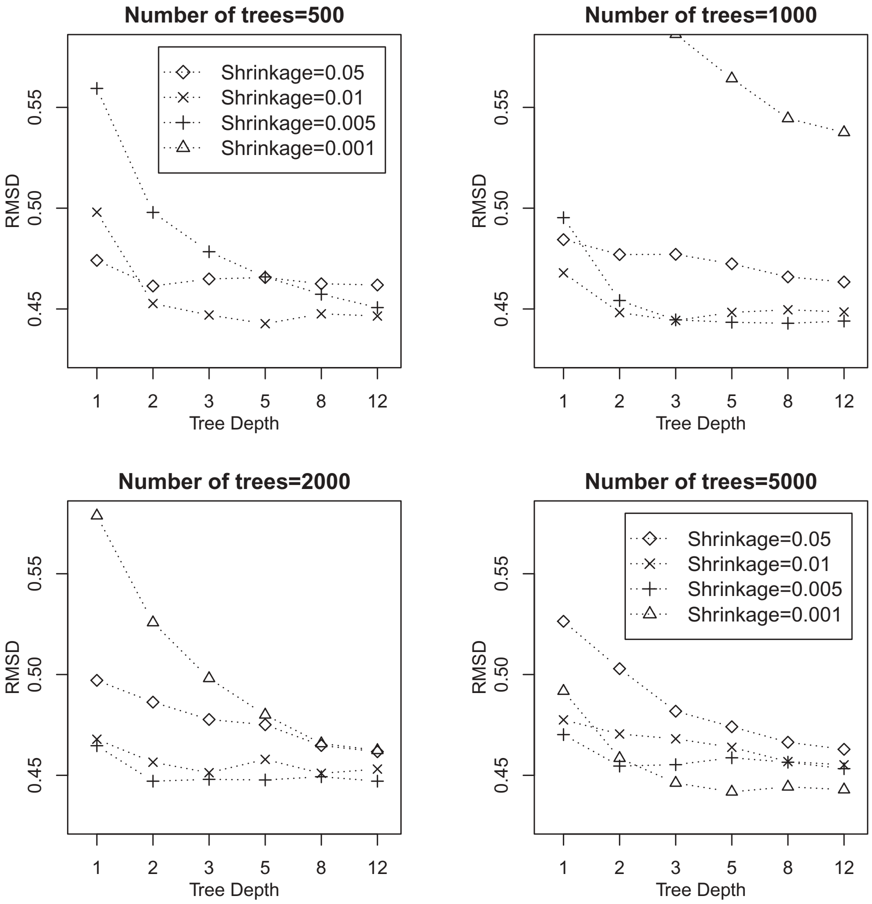

Figure 3. The values of RMSD for different values of the tuning parameter.

smallest values of the RMSD are achieved for different combinations of the tuning parameters such as, (a) $n_t = 2,000$, $\lambda = 0.005$, $d = 2$, (b) $n_t = 1,000$, $\lambda = 0.005$, $d = 5$, (c) $n_t = 500$, $\lambda = 0.01$, $d = 5$, and (d) $n_t = 5,000$, $\lambda = 0.001$, $d = 5$. For each set of predictors, we examined three figures—one like Figure 3 for RMSDs and two more for correlation and percent exact agreement and chose the combination of tuning parameters that led to values of the three accuracy measures that are better overall than those for any other combination. When both the process features and the product features were used to predict the essay scores, the combination $n_t = 5000$, $\lambda = 0.001$, $d = 5$ was chosen because it led to the smallest RMSD (see Figure 3) and to values of correlation and exact agreement that were very close to their largest values over all tuning parameters. The combinations $n_t = 2000$, $\lambda = 0.005$, $d = 3$ and $n_t = 1000$, $\lambda = 0.005$, $d = 5$ of the tuning parameters were chosen when only the product features and the process features, respectively, were used to predict the essay scores.

It is worth noting that the default values of the tuning parameters in the R package *gbm*, $n_t = 100$, $\lambda = 0.001$, $d = 1$ led to values of the accuracy measures that are much worse than those with the optimum values. For example, the RMSD with the default values was equal to 0.75 for prediction from both process and product features, which is much larger than the smallest value of the RMSD (0.44) in Figure 3. Thus, it is important to choose the optimum values of the tuning parameters using a search like that performed here rather than using the default values of the parameters of any statistical package.

Comparative Performance of the Two Methods

Table 6 shows the average correlation coefficients, RMSDs, and percent exact agreements (*% Exact*) resulting from the two prediction methods for each of the three sets of predictors and for each test form across 100 iterations. Note that the measures reported in Table 6 were computed from the test sets. Rows 1–2, 3–4, and 5–6 of the table, respectively, show the results for the process features, product features, and both process and product features. The main purpose of this article (and of essay scoring) is to obtain improved scoring or classification accuracy, so the numbers in Table 6 are of central interest in this article.

Table 6 shows that for each form and for each set of predictors, boosting slightly outperforms linear regression; boosting leads to a larger correlation coefficient and percent exact agreement and smaller RMSD than linear regression. The gain in prediction quality from the use of boosting over the use of linear regression is modest; for example, the maximum gain (reduction) in RMSD over all the cases is 0.034 (for Form CF and process features). Table 6 also shows that the process features do not add much above and beyond the product features to the prediction of the essay scores—the accuracy measures are about the same when only the product features are used as predictors and when both the product and process features are used as predictors. Zhang and Deane (2015) described a similar result for linear regression.

Another important characteristic of Table 6 is that the process features by themselves predict the essay scores only slightly worse than the product features. This result, keeping in mind the fact that the product features are designed to predict the essay scores well (e.g., Attali & Burstein, 2006), suggests validity of writing process features in that they are related to essay quality; on the other hand, the result may be an indication of the fact that the majority of the process features (such as those based on pause patterns) are more or less indicators of general writing fluency that are correlated with product features. Cognitive theories of writing (e.g., Kellogg, 1996, 2001; Kellogg, Whiteford, Turner, Cahill, & Mertens, 2013; McCutchen, 1996, 2011) attach considerable importance to fluency of linguistic and orthographic processing and indicate that problems in any part of the writing process will tend to place a load on working memory, thereby reducing overall writing fluency. An association between writing process features, fluency, and writing quality is therefore to be expected.

Variable Importance

We were also interested in finding the process and product features that are the most important in predicting the essay scores. A variable-importance measure referred to as the *relative influence* can be computed for each predictor in boosting and is available in the R package *gbm* (Ridgeway, 2017). The relative influence of a predictor is based on the number of times the predictor is selected for splitting, weighted by the squared improvement to the model as a result of each split, and is averaged over all the trees. The relative influence of each predictor is scaled so that the sum of the measure for all predictors is 100 with a larger value indicating stronger influence on the response.

Columns 1 and 2 of Table 7 show the relative influence measure, computed using the R package *gbm*, of the predictors with the 10 largest values of relative influence for either of the two forms in

Table 6. Average correlation coefficient and RMSD for the test set for the four prediction methods.

		Form GG			Form CF		
Predictor set	Method	Correlation	RMSD	% exact	Correlation	RMSD	% exact
Process	*Lin reg*	.761	.526	36.8	.678	.642	30.7
features	Boosting	.784	.500	37.3	.712	.608	31.4
Product	*Lin reg*	.829	.451	44.0	.770	.551	34.3
features	Boosting	.838	.442	44.9	.772	.549	34.8
Process &	*Lin reg*	.824	.457	43.0	.762	.561	35.0
product features	Boosting	.838	.442	45.0	.783	.550	35.2

Note. Lin reg refers to linear regression.

the prediction using both process and product features. The features are sorted by their relative influence measure for Form CF. For the Form CF (column 1), the 10 predictors with the largest relative influence included six product features (development, mechanics, etc.) and four process features (time on task, typing speed, burst number, and burst length). For the form GG (column 2), the 10 predictors with the largest relative influence included seven product features and three process features. With the addition of the process features in the prediction model, it is interesting that three product features—usage, vocabulary sophistication, and collocation-preposition—are *not* shown to have the largest relative influence. It appears that persistence in writing process (measured by time on task) and translation and transcription fluency in text generation (measured by burst number and typing speed) outperform these product features in predicting the quality of the final product. Especially given that the human scores are based on the submitted essays (and not on the process features), these results speak to the value of the process features, in particular the three mentioned above, in predicting writing quality.

Columns 3 and 4 of Table 7 show the relative influence measure for the 10 predictors for which the relative influence was the largest for either of the the two forms when only the process features were used as predictors. The features that have the largest relative influence include time on task, typing speed, burst number, burst length, uncorrected spelling error, word initial pause, multiword edit time, start time, phrasal burst, end sentence punctuation pause, and word final pause. These results generally indicate that writers who are more persistent and efficient in producing texts tend to produce responses of higher quality as perceived by human raters. Note that Li et al. (2016) also found time on task to be an important predictor of essay scores in their application of random forests to KL data. Even though the time on task was an important predictor in our study, the feature alone does not predict essay scores as well as do all the KL features. When only the time on task was used as a predictor of the adjudicated essay scores, the correlation coefficient between the actual and predicted essay scores on a test set was 0.51 for linear regression and 0.53 for boosting. These values are substantially smaller than the corresponding numbers in the first two rows of Table 6 and show that the KL variables other than time on task have a substantial amount of added value in prediction of the essay scores.

The values of the relative influence when only the product features are used as predictors were roughly in the same order as in columns 1 and 2 and are not shown here.

In applications of boosting to regression, the influence of a predictor on the response is also examined using a *partial dependence plot* that shows the average predicted value of the response for

Table 7. Relative influence of the predictors.

	Process and product features		Process features	
Predictor	Form CF	Form GG	Form CF	Form GG
Development	25.9	30.1	—	—
Mechanics	11.0	6.1	—	—
Syntactic variety	10.7	12.7	—	—
Organization	6.4	8.9	—	—
Time on task	5.0	5.5	18.3	19.3
Style	4.7	3.9		
Typing speed	3.3	1.8	15.5	7.6
Grammar	2.8	1.9	—	—
Burst length	2.7	1.4	6.5	9.5
Burst number	2.2	3.0	14.7	18.3
Word length	1.2	4.4	—	—
Uncorrected spelling errors	—	—	4.8	3.7
Word initial pause	—	—	3.5	4.8
Multiword edit time	—	—	2.7	2.4
Start time	—	—	2.5	2.5
Phrasal burst	—	—	2.0	2.7
End sentence punctuation pause	—	—	1.8	—
Word final pause	—	—	—	3.4

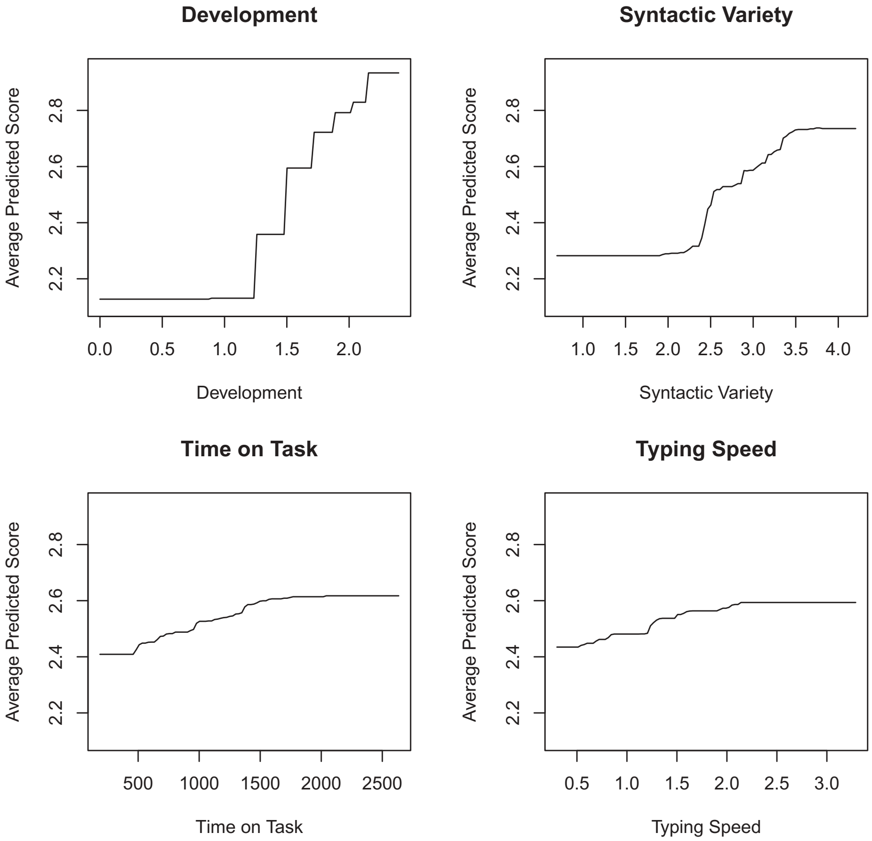

Figure 4. Partial dependence plots for four features.

different values of the predictor (e.g., Elith et al., 2008). Figure 4 shows the *partial dependence plots* produced by the R package *gbm* for four features—development, syntactic variety, time on task, and typing speed—that were found to have large relative influence in Table 7 in prediction from both the process and product features. Note that the first two of these features are product features and the latter two are process features. Each panel of Figure 4 shows, for one feature, the average predicted value of the adjudicated essay score (averaged over the values of the other predictors) for different values of the feature. The range of the Y-axis is the same in all the panels so as to enable a comparison of the effect of the features on the adjudicated essay score. While these plots are not a perfect representation of the effect of the predictors, especially if the predictors are highly correlated, they provide a useful basis for interpretation (Friedman, 2001). Figure 4 shows that the predicted score increases with an increase in the value of each feature, but the rate of increase is different; the rate is the largest for development, which is expected given that the relative influence of that feature is the largest in Table 7. The two bottom panels indicate that the average score increases with an increase of the two process features until the process features reach certain values (1800 and 2.2, respectively)

Conclusions and Recommendations

This article focuses on the application of boosting, one of the most popular DMMCR, to predict the essay scores on a writing assessment from writing process and product features using freely available R software packages. Boosting slightly outperformed linear regression, which has been the method of choice in predicting essay scores from process features (e.g., Deane, 2014; Zhang & Deane, 2015). The computation times of boosting were very short for these data sets. For example, when both the process and product features were used as predictors, the time required in one replication of the aforementioned analysis[6] was about two seconds for stepwise regression and three seconds for boosting on a standard computer. Thus, the improved prediction from boosting does not come at much of a cost. It should be noted that the search for the optimum values of the tuning parameters in boosting took a couple of additional hours for our data and may take some time in any application of boosting.

The choice of the tuning parameters is a major issue in applications of the DMMCR including boosting. Results for the data from the writing assessment show that setting the tuning parameters to their default values from a software package may lead to poor prediction. A preliminary search was used to find optimum values of the tuning parameters, which were used in the final analysis.

Note that the extent of improvement that boosting provides over the linear regression method is modest for the data considered in this article—this result may inspire practitioners to continue using linear regression because of its simplicity in similar problems. This modest gain for boosting in this article and similar results in Sinharay (2016) and in applications of DMMCR in other fields (e.g., Fernandez-Delgado et al., 2014) indicate that measurement practitioners should not always expect huge improvements from the DMMCR. Also, the above results do not guarantee that boosting will outperform linear regression for all data sets related to essay scoring. If a practitioner is interested in applying boosting in a real application, it would be prudent to apply boosting with different values of the tuning parameters and examine if the method leads to improved prediction over linear regression. It is possible that boosting does not improve over linear regression; for example, Konig et al. (2008) found several DMMCR including boosting, to not improve over logistic regression in predicting patient prognosis.

The DMMCR including boosting have several additional limitations. First, they are more time-consuming than traditional prediction methods and can take hours for large data sets compared to seconds for traditional methods. Second, while the DMMCR may account for the true functional form of the relationship between the response and the predictors, the relationship is usually impossible to interpret because of complicated interactions. Thus, the DMMCR are like black boxes so that the investigator feeds the data and obtains predictions, but does not obtain any insight on the exact form of the relationship between the response and the predictors (e.g., James et al., 2013, pp. 24–26, p. 319) although relative influence measures and partial dependence plots may provide some such insight. Some experts suggested that because of their improved prediction at the expense of interpretability, DMMCR can be used where prediction is the only goal (e,g,, in predicting the stock market) and should not be used where inference is the main goal (e.g., James et al., 2013, pp. 24–25). Third, while some theoretical results are available on the accuracy of prediction of the DMMCR, the number of such results is much smaller than that for linear regression; for example, theoretical results are available on the standard error of the predicted values in linear regression but not for the values predicted by boosting. Thus, the practitioners who intend to apply the DMMCR have to weigh these limitations against the (possibly) improved prediction provided by the DMMCR.

From a substantive viewpoint, the results, especially those on the relative influence measures of the process features, support the (theoretically well-grounded) conclusion that efficiency, fluency, and persistence during the writing process are important contributors to writing quality. It is important to note, however, the limitation that the writing assessment considered here was administered under low-stakes conditions and required writers to generate text in a single sitting. On an extended writing task, in contrast, writers may be given time separately dedicated for task planning, drafting, reviewing, and editing, as well as finalizing their essays for submission. The type of direct, timed writing assessment analyzed in this study places a premium emphasis on fluency and efficiency, where writers, constrained by time, are required to retrieve their knowledge on the

[6] Where one replication involves the fitting of the model to one subset of the data and the prediction from another subset.

topic, understand the writing goal, generate a plan, translate the plan into well-formed standard English language, and transcribe the text on a computer accurately, in a single writing session. As a result, deficiencies in linguistic fluency and/or orthographic skills (e.g., difficulty with keyboarding), for example, will majorly interfere with overall writing performance by demanding more cognitive resources that can potentially be allocated for such other tasks as idea generation, reviewing, and revising. Content knowledge is less of a concern in the assessment studied here as all writers completed a series of pre-writing tasks on the topic and hence they were presumably equally well-prepared on the writing subject. All in all, the particular writing assessment, administered in a low-stakes testing environment and analyzed in this study, may lead to less-motivated students devoting less time and effort to the task than they would in a high-stakes context or if they had been able to complete the task in multiple sessions. It may also discourage students from attempting major rewrites or revisions. The relationships between KL features and writing quality may be mediated by task requirements and the conditions under which writing takes place.

The present article has several practical implications for researchers and practitioners. The study serves as a precursor if a *writing process score* or subscore (e.g., fluency) is to be reported or writing processes are considered when a general writing score is reported. The features identified to have a high level of importance, including the time on task and typing speed, may be of most interest. Another implication is that the process features can be used as validation evidence, thereby supporting analysis of differences between writing tasks with different requirements. For example, for a copy-typing task, typing speed is most likely to manifest itself as a variable of high importance, whereas for an editing task, a long pause before a jump may exhibit high importance. Researchers are also encouraged to develop profiles of students' writing processes using the identified high-importance features as a starter; such profiles may have important practical usefulness for writing instruction. Finally, the data analyses in this study suggest a complex structure of the timing and process data that are obtained from writing processes, which may not be readily dealt with using traditional psychometric methods. For instance, classification of writing processes using the temporal data is not a traditional field in educational measurement. Hence, this study calls for practitioners and researchers to develop or apply nontraditional methods that can help explain the complex writing processes and validate the variables extracted from the writing processes for educational purpose.

Several additional related topics can be examined further. First, more data sets may be used to compare boosting and linear regression in the context of predicting essay scores from process and product features. Of special interest would be data from high-stakes examinations that may show different patterns, either with respect to the quality of prediction or to the process features that are important predictors of the essay score. Similarly, more DMMCR in addition to boosting can be used to predict essay scores. Some such methods are generalized additive models, multivariate adaptive regression splines, nearest neighbor methods, neural networks, random forests, and support vector regression (e.g., Hastie et al., 2009). In a preliminary analysis, the last two methods were used with the data from the writing assessment, but they did not improve over boosting. Second, although we briefly looked at variable importance measures, more research can be performed on the examination of variable importance in applications of boosting. Third, it is possible to consider more process features and apply them in prediction of essay scores. Finally, it is possible to repeat the analysis performed in this article with demographic subgroups to examine whether the quality of prediction varies over subgroups.

Disclosure statement

No potential conflict of interest was reported by the authors.

ORCID

Sandip Sinharay http://orcid.org/0000-0003-4491-8510
Paul Deane http://orcid.org/0000-0002-6838-2856

References

Agresti, A. (2013). *Categorical data analysis* (*3rd* ed.). New York, NY: Wiley.

Allen, L. K., Jacovina, M. E., Dascalu, M., Roscoe, R. D., Kent, K., Likens, A. D., & McNamara, D. S. (2016). {ENTER}ing the time series {SPACE}: Uncovering the writing process through keystroke analyses. In T. Barnes, M. Chi, & M. Feng (Eds.), *Proceedings of the 9th international conference on educational data mining*. Raleigh, NC: International Educational Data Mining Society.

Almond, R., Deane, P., Quinlan, T., Wagner, M., & Sydorenko, T. (2012). *A preliminary analysis of keystroke log data from a timed writing task* (ETS Research Report No. RR-12-23). Princeton, NJ: ETS.

Attali, Y., & Burstein, J. (2006). Automated essay scoring with e-rater V.2. *Journal of Technology, Learning and Assessment*, *4*(3), 1–29.

Bennett, R. E. (2011). *CBAL: Results from piloting innovative K-12 assessments* (ETS Research Report No. RR-11-23). Princeton, NJ: ETS.

Berninger, V. W. (1999). Coordinating transcription and text generation in working memory during composing: Automatic and constructive processes. *Learning Disability Quarterly*, *22*, 99–112. doi:10.2307/1511269

Breiman, L., Friedman, J. H., Olshen, R. A., & Stone, C. J. (1984). *Classification and regression trees*. Belmont, CA: Wadsworth.

Chen, J., Fife, J., Bejar, I. I., & Rupp, A. A. (2016). *Building e-rater scoring models using machine learning algorithm* (ETS Research Report No. RR-16-04). Princeton, NJ: ETS.

Deane, P. (2014). *Using writing process and product features to assess writing quality and explore how those features relate to other literacy tasks* (ETS Research Report No. RR-14-03). Princeton, NJ: ETS.

Deane, P., & Zhang, M. (2015). *Exploring the feasibility of using writing process features to assess text production skills* (ETS Research Report No. RR-15-26). Princeton, NJ: ETS.

Elith, J., Leathwick, J. R., & Hastie, T. (2008). A working guide to boosted regression trees. *Journal of Animal Ecology*, *77*, 802–813. doi:10.1111/j.1365-2656.2008.01390.x

Emig, J. (1972). *The composing processes of twelfth graders*. Urbana: National Council of Teachers of English.

Feinerer, I., Hornik, K., & Meyer, D. (2008). Text mining infrastructure in R. *Journal of Statistical Software*, *25*(5), 1–54.

Fernandez-Delgado, M., Cernadas, E., Barro, S., & Amorim, D. (2014). Do we need hundreds of classifiers to solve real world classification problems?. *Journal of Machine Learning Research*, *15*, 3133–3181.

Flower, L. S., & Hayes, J. R. (1981). A cognitive process theory of writing. *College Composition and Communication*, *32*, 365–387. doi:10.2307/356600

Friedman, J. H. (2001). Greedy function approximation: A gradient boosting machine. *The Annals of Statistics*, *29*, 1189–1232. doi:10.1214/aos/1013203451

Friedman, J. H. (2002). Stochastic gradient boosting. *Computational Statistics and Data Analysis*, *38*, 367–378. doi:10.1016/S0167-9473(01)00065-2

Gey, S., & Poggi, J.-M. (2006). Boosting and instability for regression trees. *Computational Statistics & Data Analysis*, *50*, 533–550. doi:10.1016/j.csda.2004.09.001

Gierl, M. J. (2007). Making diagnostic inferences about cognitive attributes using the rule space model and attribute hierarchy method. *Journal of Educational Measurement*, *27*, 325–340. doi:10.1111/j.1745-3984.2007.00042.x

Hastie, T., Tibshirani, R., & Friedman, J. H. (2009). *The elements of statistical learning: Data mining, inference, and prediction*. New York, NY: Springer.

Hayes, J. R. (2012). Modeling and remodeling writing. *Written Communication*, *29*, 369–388. doi:10.1177/0741088312451260

He, Q., & von Davier, M. (2016). Analyzing process data from problem-solving items with n-grams:. In Y. Rosen, S. Ferrara, & M. Mosharraf (Eds.), *Handbook of research on technology tools for real-world skill development* (pp. 750–777). Hershey, PA: IGI Global.

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). *An introduction to statistical learning with applications in R*. New York, NY: Springer.

Keller, L., Zenisky, A. L., & Wang, X. (2016). Analyzing process data from technology-rich tasks. In Y. Rosen, S. Ferrara, & M. Mosharraf (Eds.), *Handbook of research on technology tools for real-world skill development* (pp. 725–749). Hershey, PA: IGI Global.

Kellogg, R. T. (1996). A model of working memory in writing. In C. M. Levy & S. Ransdell (Eds.), *The science of writing: Theories, methods, individual differences, and applications* (pp. 57–71). Hillsdale, NJ: Lawrence Earlbaum Associates.

Kellogg, R. T. (2001). Competition for working memory among writing processes. *The American Journal of Psychology*, *114*, 170–191. doi:10.2307/1423513

Kellogg, R. T., Whiteford, A. P., Turner, C. E., Cahill, M., & Mertens, A. (2013). Working memory in written composition: A progress report. *Journal of Writing Research*, *5*, 159–190. doi:10.17239/jowr-2013.05.02.1

Konig, I. R., Malley, J. D., Pajevic, S., Weimar, C., Diener, H. C., & Ziegler, A. (2008). Patient-centered yes/no prognosis using learning machines. *International Journal of Data Mining and Bioinformatics*, *2*, 289–341. doi:10.1504/IJDMB.2008.022149

Kuhn, M. (2008). Building predictive models in R using the caret package. *Journal of Statistical Software*, *28*, 340–352. doi:10.18637/jss.v028.i05

Kutner, M. H., Nachtsheim, C. J., Neter, J., & Li, W. (2004). *Applied linear statistical models* (*5th* ed.). New York, NY: McGraw-Hill/Irwin.

Leijten, M., Janssen, D., & Van Waes, L. (2010). Error correction strategies of professional speech recognition users: Three profiles. *Computers in Human Behavior*, *26*, 964–975. doi:10.1016/j.chb.2010.02.010

Leijten, M., Macken, L., Hoste, V., Van Horenbeeck, E., & Van Waes, L. (2012). From character to word level: Enabling the linguistic analyses of inputlog process data. In M. Piotrowski, C. Mahlow, & R. Dale (Eds.), *Proceedings of the second workshop on computational linguistics and writing* (pp. 1–8). Avignon, France.

Leijten, M., & Van Waes, L. (2013). Keystroke logging in writing research: Using Inputlog to analyze and visualize writing processes. *Written Communication*, *30*, 358–392. doi:10.1177/0741088313491692

Li, C., Zhang, M., & Deane, P. (2016, April). *Investigating the relations of writing process features and the final product.* Paper presented at the Annual meeting of the National Council on Measurement in Education, Washington, DC.

McCutchen, D. (1996). A capacity theory of writing: Working memory in composition. *Educational Psychology Review*, *8*, 299–325. doi:10.1007/BF01464076

McCutchen, D. (2011). From novice to expert: Implications of language skills and writing-relevant knowledge for memory during the development of writing skill. *Journal of Writing Research*, *3*, 51–68. doi:10.17239/jowr-2011.03.01.3

Medimorec, S., & Risko, E. F. (2017). Pauses in written composition: On the importance of where writers pause. *Reading and Writing*, *30*, 1267–1285. doi:10.1007/s11145-017-9723-7

Milborrow, S. (2016). *rpart.plot: Plot ‘rpart’ models: An enhanced version of ‘plot.rpart’.* (R package version 2.1.0).

National Center for Education Statistics. (2011). *The National Assessment of Educational Progress at grades 8 and 11 (NCES-2012-470).* Washington, DC: Institute for Educational, US Department of Education.

National Research Council. (2001). *Knowing what students know: The science and design of educational assessment.* Washington, D.C.: National Academies Press.

Oranje, A., Gorin, J. S., Jia, Y., & Kerr, D. (2017). Collecting, analyzing and interpreting response time, eye-tracking, and log data. In K. Ercikan & J. Pellegrino (Eds.), *Validation of score meaning for the next generation of assessments: The use of response processes* (pp. 39–51). New York, NY: Routledge.

R Core Team. (2017). *R: A language and environment for statistical computing.* Vienna, Austria: R Foundation for Statistical Computing.

Ridgeway, G. (2017). *gbm: Generalized boosted regression modeling.* (R package version 2.1.3).

Sheehan, K., & Mislevy, R. J. (1994). *A tree-based analysis of items from an assessment of basic mathematics skills* (ETS Research Report No. RR-94-14). Princeton, NJ: Educational Testing Service.

Sinharay, S. (2016). An NCME instructional module on data mining methods for classification and regression. *Educational Measurement: Issues and Practice*, *35*(3), 38–54. doi:10.1111/emip.12115

Stallard, C. K. (1974). An analysis of the writing behavior of good student writers. *Research in the Teaching of English*, *8*, 206–218.

Strobl, C. (2013). Data mining. In T. Little (Ed.), *The Oxford handbook of quantitative methods in psychology* (Vol. 2, pp. 678–700). New York, NY: Oxford University Press.

Strobl, C., Malley, J., & Tutz, G. (2009). An introduction to recursive partitioning: Rationale, application, and characteristics of classification and regression trees, bagging, and random forests. *Psychological Methods*, *14*, 323–348. doi:10.1037/a0016973

Therneau, T., Atkinson, B., & Ripley, B. (2017). *rpart: Recursive partitioning and regression trees.* (R package version 4.1-11).

Tukey, J. W. (1977). *Exploratory data analysis*. Reading, MA: Addison-Wesley.

Wei, T., & Simko, V. (2017). *R package “corrplot”: Visualization of a correlation matrix.* (Version 0.84).

Zhang, M., & Deane, P. (2015). *Process features in writing: Internal structure and incremental value over product features* (ETS Research Report No. RR-15-27). Princeton, NJ: ETS.

Zhang, M., Hao, J., Li, C., & Deane, P. (2016). Classification of writing patterns using keystroke logs. In L. A. van der Ark, D. M. Bolt, W.-C. Wang, J. A. Douglas, & M. Wiberg (Eds.), *Quantitative psychology research* (pp. 299–314). New York, NY: Springer International Publishing.

Zhang, M., Zou, D., Wu, A. D., Deane, P., & Li, C. (2017). An investigation of the writing processes in timed task condition using keystrokes. In B. D. Zumbo & A. M. Hubley (Eds.), *Understanding and investigating writing processes in validation research* (pp. 321–339). New York, NY: Springer.

Appendix A. The Scoring Rubric

EXEMPLARY (5). An exemplary response meets all of the requirements for a score of 4 but distinguishes itself by skillful use of language, precise expression of ideas, effective sentence structure, and/or effective organization, which work together to control the flow of ideas and enhance the reader's ease of comprehension.

CLEARLY COMPETENT (4). A clearly competent response typically displays the following characteristics:

- It is adequately structured.
  - Overall, the response is clearly and appropriately organized for the task.
  - Clusters of related ideas are grouped appropriately and divided into sections and paragraphs as needed.
  - Transitions between groups of ideas are signaled appropriately.

- It is coherent.
  - Most new ideas are introduced appropriately.
  - The sequence of sentences leads the reader from one idea to the next with few disorienting gaps or shifts in focus.
  - Connections within and across sentences are made clear where needed by the use of pronouns, conjunctions, subordination, and so on.

- It is adequately phrased.
  - Ideas are expressed clearly and concisely.
  - Word choice demonstrates command of an adequate range of vocabulary.
  - Sentences are varied appropriately in length and structure to control focus and emphasis.

- It displays adequate control of Standard Written English (SWE).
  - Grammar and usage follow SWE conventions, but there may be minor errors.
  - Spelling, punctuation, and capitalization follow SWE conventions, but there may be minor errors.

DEVELOPING HIGH (3). A response in this category displays some competence but differs from Clearly Competent responses in at least one important way, including limited development; inconsistencies in organization; failure to break paragraphs appropriately; occasional tangents; abrupt transitions; wordiness; occasionally unclear phrasing; little sentence variety; frequent and distracting errors in SWE; or relies noticeably on language from the source material.

DEVELOPING LOW (2). A response in this category differs from Developing High responses because it displays serious problems such as, marked underdevelopment; disjointed, list-like organization; paragraphs that proceed in an additive way without a clear overall focus; frequent lapses in cross-sentence coherence; unclear phrasing; excessively simple and repetitive sentence patterns; inaccurate word choices; errors in SWE that often interfere with meaning; or relies substantially on language from the source material.

MINIMAL (1). A response in this category differs from Developing Low responses because of serious failures such as, extreme brevity; a fundamental lack of organization; confusing and often incoherent phrasing; little control of SWE; or can barely develop or express ideas without relying on the source material.

NO CREDIT (0). Not enough of the student's own writing for surface-level characteristics to be judged; not written in English; completely off topic; or random keystrokes.

Appendix B. Annotated R Code Used in the Computations

```
>library(gbm)
>Data=read.csv('Data.csv',header=TRUE) #Read the input file
>Y=Data$scores #Set the adjudicated essay score as the response variable >n=round(0.67*nrow(Data))
>s=sample(1:nrow(Data),n)
>ModelBuild=Data[s,]#Assign a subset of the sample as the 'model-building set'
>Test=Data[-s,]#Assign the other subset of the data as the 'Test set'
# LINEAR REGRESSION
>reg=lm(Y˜.,data=ModelBuild)
>step=stepAIC(reg,direction="both",trace = FALSE)
>predReg=predict(step,newdata=Test)
>RMSDReg=sqrt(mean((Test$Y-predReg)**2)))#Compute RMSD from the test set #BOOSTING
>gbm1=gbm(Y˜.,data=ModelBuild,distribution="gaussian",n.trees=500,interaction.depth=3, shrinkage=0.01)
>predgbm= predict(gbm1,newdata=Test,type = "response",n.trees=500)
>RMSDgbm=sqrt(mean((Test$Y-predgbm)**2)))#Compute RMSD from the test set for boosting
#RELATIVE INFLUENCE FROM BOOSTING
>summary(gbm1,method=relative.influence)
#PARTIAL DEPENDENCE PLOT
>plot(gbm,i.var=1)#plot for one predictor—'time on task'
```