What Is Successful Writing? An Investigation Into the Multiple Ways Writers Can Write Successful Essays

Scott A. Crossley, Rod Roscoe and Danielle S. McNamara

Written Communication 2014 31: 184
DOI: 10.1177/0741088314526354

Written Communication
2014, Vol. 31(2) 184–214
© 2014 SAGE Publications
Reprints and permissions:
sagepub.com/journalsPermissions.nav
DOI: 10.1177/0741088314526354
wcx.sagepub.com

What Is Successful Writing? An Investigation Into the Multiple Ways Writers Can Write Successful Essays

Scott A. Crossley[1], Rod Roscoe[2], and Danielle S. McNamara[2]

Abstract

This study identifies multiple profiles of successful essays via a cluster analysis approach using linguistic features reported by a variety of natural language processing tools. The findings from the study indicate that there are four profiles of successful writers for the samples analyzed. These four profiles are linguistically distinct from one another and demonstrate that expert human raters examine a number of different linguistic features in a variety of combinations when assessing writing proficiency and assigning high scores to independent essays (regardless of the scoring rubric considered). The writing styles in the four clusters can be described as *action and depiction style, academic style, accessible style*, and *lexical style*. The study provides empirical evidence that successful writing cannot be defined simply through a single set of predefined features, but that, rather, successful writing has multiple profiles. While these profiles may overlap, each profile is distinct.

[1]Georgia State University, Atlanta, GA, USA
[2]Arizona State University, Tempe, AZ, USA

Corresponding Author:
Scott A. Crossley, Department of Applied Linguistics/ESL, Georgia State University, 34 Peachtree St. Suite 1200, One Park Tower Building, Atlanta, GA 30303, USA.
Email: sacrossley@gmail.com

Keywords
writing quality, writing profiles, cluster analysis, corpus linguistics, computational linguistics

Traditionally, many studies of writing have approached the question of successful writing using a linear modeling approach (Attali & Burstein, 2006; McNamara, Crossley, & Roscoe, 2013). That is, some prior research seems to implicitly assume that there is a single definition of successful writing and that all "low" and "high" quality writers demonstrate similar patterns and strategies in developing an essay. Such assumptions can be seen in linear regression models of writing (Attali & Burstein, 2006; McNamara et al., 2013), readability formulas (Flesch, 1948; Kincaid, Fishburne, Rogers, & Chissom, 1975), writing guidelines (Strunk & White, 1968), additive checklists (Hayes, 1989), or primary trait scoring rubrics like those used by Educational Testing Service or SAT. However, this is likely a misleading approach because it presumes that all writing tasks and writers are alike and follow similar prescriptive processes (Schriver, 1989). It is reasonable to presume that the most successful essays and the most successful writers will not always share the same linguistic attributes (i.e., there are different ways to write a successful essay). Some successful essays might focus more on cohesion than structure, or they may focus more on lexical sophistication than content. Although cohesion, structure, lexical sophistication, and content are all important linguistic elements of essay writing, they may combine in different variations to create a successful essay. Thus, there may be multiple variable configurations available to writers that will lead to a successful essay.

The goal of this study is to identify multiple linguistic profiles of successful essays via a cluster analysis approach. Our definition of successful essays is limited to timed, independent, argumentative essays written by high school and college students and judged by human raters using a holistic scoring rubric. Thus, in this study, we examine essays judged as successful in a very particular context. The methodology used in this article groups successful essays written into dissimilar clusters using linguistic features reported by a variety of natural language processing tools such as Coh-Metrix (McNamara & Graesser, 2012), the Writing Analysis Tool (WAT; Crossley, Roscoe, & McNamara, 2013; McNamara et al., 2013), and Linguistic Inquiry and Word Count (LIWC; Pennebaker, Francis, & Booth, 2001). Our purpose is to provide empirical evidence for the argument that successful writing cannot be defined simply through a single set of predefined features, but that, rather, successful writing has multiple profiles. While these profiles may overlap, each profile is distinct. Providing evidence for such profiles would have important implications for better understanding successful writing, for understanding human ratings of essay quality, and for providing summative and formative feedback in the classroom and in automatic essay evaluation systems.

Successful Writing

The ability to produce a successful written text is thought to play an important role in the success of individuals in school and in the workplace (Geiser & Studley, 2001; Powell, 2009). However, attaining success in writing is often a difficult task (National Assessment of Educational Progress, 2011). As a result, there have been numerous research studies that focus on understanding less successful writing and developing pedagogical approaches to help advance less successful students' writing skills (Applebee, Langer, Jenkins, Mullis, & Foertsch, 1990; Ferrari, Bouffard, & Rainville, 1998; McNamara, Crossley, & McCarthy, 2010). These studies generally recognize that writing is a complex and demanding process that requires the coordination of a number of cognitive and knowledge skills, including discourse awareness, goal setting, sociocultural knowledge, and memory management strategies (Flower & Hayes, 1980; Kellogg & Whiteford, 2009).

Many of these skills (i.e., discourse awareness and memory management strategies) have a direct effect on the production of linguistic features in essays. Linguistic features such as cohesion, syntactic complexity, and lexical sophistication relate to higher cognitive functioning, and their use in an essay indicates a writer's ability to quickly and easily produce complex text, freeing the writer to focus on other cognitive resources that can be used to address rhetorical and conceptual concerns in the text (Deane, 2013). Thus, many researchers have examined the ability of these linguistic features to predict essay quality (e.g., Applebee et al., 1990; Ferrari et al., 1998; McNamara et al., 2010). Importantly, these studies often find that essay corpora oversample the worst strata of essays (i.e., essays scored as 1s, 2s, and 3s on a 6-point grading scale). For instance, McNamara et al. (2010) reported a mean score of 3.26 for a corpus of essays scored on a 1 to 6 rubric (where 3.5 would be the norm). These studies often are quite accurate at predicting and explaining poorly written essays, but less accurate for successful essays because the scoring models are so strongly influenced by the overrepresentation of lower strata essays. Thus, it becomes important and worthwhile to investigate more closely the properties of "successful writing" and "successful essays" in order to ascertain what constitutes the few essays that are rated highly by expert judges. Such analyses can help generate models of quality writing and help to define diverse profiles of successful writing that can be applied in pedagogical settings (e.g., strategy training and formative feedback).

Research demonstrates that successful writers (and successful essays) exhibit multiple characteristics related to knowledge of domains, discourse, and language. For example, successful writers tend to have high domain knowledge for the topic they are writing on, which allows them to produce a greater number of relevant ideas when writing (Graham & Perry, 1993; Kellogg, 1987). Strong domain knowledge has a powerful effect on writing performance (Ericsson, Charness, Feltovich, & Hoffman, 2006; Simon & Chase, 1973) leading to the use of more sophisticated strategies and the production of better-structured texts (Bereiter & Scardamalia, 1987). From a discourse perspective, successful writers have greater discourse knowledge, which allows them to freely switch between discourse modes such as found in differences among narrative, comparative, and argumentative essays (McCormick, Busching, & Potter, 1992; McCutchen, 1986). Linguistically, successful writers know more about written language and language in general. Thus, successful writers are more proficient with syntax, lexicon, punctuation, grammar, and spelling (Applebee et al., 1990). Such advantages allow successful writers to focus more on the argumentative and rhetorical structures in essays by reducing the cognitive workload associated with generating and organizing ideas (Deane, 2013; Kellogg, 1987; McCutchen, Covill, Hoyne, & Mildes, 1994).

Since our interest in this study lies in the linguistic properties of essays, we focus on this area of writing. Children began to master basic linguistic properties of writing at a relatively early age. By the second grade, children begin to demonstrate basic mastery of grammar, sentence structure, and discourse devices (Berninger et al., 1992). Linguistically, successful writers produce longer texts (Crossley, Weston, McLain Sullivan, & McNamara, 2011; Ferrari et al., 1998; Haswell, 2000; McNamara et al., 2010; McNamara et al., 2013) with fewer errors in spelling, grammar, and punctuation (Ferrari et al., 1998). Essay quality is also related to a writer's lexical knowledge. For instance, skilled middle school students show greater lexical generation than do less skilled middle school students (McCutchen, 1986; McCutchen et al., 1994). More specifically, essays scored as higher quality by expert raters contain more infrequent words (Crossley, Weston, et al., 2011; McNamara et al., 2010; McNamara et al., 2013), and as a function of grade level, more advanced writers produce essay with more infrequent words, more ambiguous words, fewer concrete words (Crossley, Weston, et al., 2011), and longer words (Haswell, 2000).

Writing can also be related to measures of text cohesion. At an early age, as young as 7, children begin to use cohesive devices such as referential pronouns and connectives (King & Rentel, 1979). The development of such cohesive devices continues until about the eighth grade (McCutchen & Perfetti, 1982). For instance, in comparisons of essays written by sixth and eighth graders, a greater use of local cohesion (i.e., more connectives) is found in eighth grade essays (McCutchen, 1986). However, as writers mature, they tend to develop coherence in text less through the use of explicit cohesive devices. For instance, when examining human judgments of essay quality, McNamara et al. (2010) found that less successful writers had less lexical diversity (i.e., more repetition of words) than more successful writers. In addition, Crossley, Weston, et al. (2011) found that ninth grade writers produced texts with more explicit cohesive devices than did eleventh grade writers and freshman college writers (e.g., greater incidence of connectives and greater word overlap). Haswell (2000) reported that as writers developed, they produced a greater number of coordinated nominal per clause as a function of time.

At the same time, advanced writers produce more complex syntactic structures. Thus, writing quality is also related to syntactic skills, with more advanced writers showing more refined sentence generation skills than beginning writers (McCutchen et al., 1994) with the development of syntactic complexity developing from first grade through college (Haswell, 2000; Stewart, 1978). As an example, in the Crossley, Weston, et al. (2011) study, more advanced writers at the freshman college level produced a greater number of syntactically complex sentences (as measured by the number of modifiers per noun phrase) than ninth grade writers. Haswell (2000) reported that more advanced writers produced longer sentences and longer clauses indicating syntactic growth over time. In similar studies that examined human judgments of essay quality, Crossley, Weston, et al. (2011) showed that higher rated essays had fewer base verb forms and McNamara et al. (2010) demonstrated that higher rated essays contained a greater number of words before the main verb phrases in sentence, both indicating that more successful essays were more syntactically complex.

Finally, essay quality can be predicted based on the use of rhetorical features. For instance, less successful essays contain more narrative features (i.e., more first person exploration of ideas). In contrast, more successful essays contain a greater number of rhetorical structures related to reported speech, contrasting ideas, strength of arguments, exemplification of arguments, and conditional structures (Crossley, Weston, et al., 2011; McNamara et al., 2013).

Taken together, these studies indicate that more successful writers (defined by human judgments and advanced grade level) produce longer essays that contain a greater number of linguistic features related to lexical sophistication, syntactic complexity, and rhetorical structures. Conversely, more successful writers produce fewer explicit cohesive devices. While these linguistic features do not measure elements of essay writing related to content and conceptual knowledge, they do indicate that successful writers can quickly and effectively write an essay of greater length with greater use of complex linguistic features, indicating that good writers are more proficient with language. This language proficiency likely reduces writers' cognitive workloads, freeing them to focus more on content, concepts, and argument structure.

Profiles of Successful Writing

Our argument is that successful writers can use many different combinations of linguistic features to produce a good essay. Such an approach is similar to combinatorial approaches to writing quality as found in analytic scoring (Huot, 1990) and concurrent testing (Schriver, 1989). Like these approaches, we hypothesize that there is more than one end result with respect to the success of an essay. However, this notion is not well represented in the literature on argumentative essays with many studies of essay quality taking a linear approach to predicting human scores of writing proficiency (i.e., predicting essay quality using linear models; Attali & Burstein, 2006; Crossley, Weston, et al., 2011; McNamara et al., 2013). In contrast to these linear approaches, we use a cluster approach that categorizes successful essays into groupings based on the incidence of linguistic features in the essays. Such an approach is relatively rare in studies of writing quality.

The one exception appears to be a study conducted by Jarvis, Grant, Bikowski, and Ferris (2003) that examined the multiple profiles of highly rated second language English compositions. Jarvis et al. examined two data sets (one corpus of 160 English placement tests and one corpus of 178 Test of English as a Foreign Language writing samples) for differences in their use of 21 linguistic features. Using a cluster analysis, they explored a five-cluster solution for the first data set. The uniqueness of the clusters was based on differences in word length, nouns, pronouns, adverbs, prepositions, present tense verbs, and stative verbs. For the second data set, they explored a three-cluster solution in which the clusters differed in terms of word length, nouns, impersonal pronouns, preposition, present tense, adverbial subordinators, and complementation. Jarvis et al. concluded that writing quality was dependent not on individual linguistic features, but rather on how the features were used in tandem with one another. Thus, highly rated texts were not isometric.

Cluster analyses have also been used to assess first language writers of English. Roid (1994) used a cluster analysis approach to examine difference between holistic, analytic, and primary-trait scoring methods. Using six analytic scores and 10,000 essays from third to eighth grade students, Roid identified subgroups of students based on similar patterns. However, Roid focused not on high quality essays but rather on low, medium, and high quality essays. The analytic traits he examined included idea, organization, voice, word choice, sentence fluency, and conventions. Overall, Roid found that high rated essays exhibited higher scores for all these traits, but medium and high essays showed four different profiles. For example, in one profile, medium-high essays earned negative scores for ideas, voice, and word choice. However, a different profile of medium-high essays demonstrated negative scores for sentence fluency and conventions.

These studies provide initial evidence that there are multiple profiles of successful writing that can be distinguished based on linguistic features. However, neither of these studies focused specifically on successful writing by first language writers using a linguistic feature analysis.

Method

The purpose of this study is to examine multiple profiles of successful writing. In particular, we focus on persuasive, prompt-based independent writing that is common in secondary writing classes and in high-stakes testing (e.g., the SAT). We first employ the computational tools Coh-Metrix, WAT, and LIWC to compute a variety of linguistic, affective, and rhetorical features for a corpus of higher quality essays. We then include these features as independent variables in a cluster analysis to examine how they might be used to categorize individual essays based on the incidence of language features. Cluster analysis can help develop profiles grounded in the linguistic features produced by writers using implicit patterns in the data. Like a factor analysis, cluster analysis is a grouping algorithm. However, while factor analysis groups variables together, cluster analysis groups cases (Antonenko, Toy, & Niederhauser, 2012). Such groups can be used represent profiles of higher quality writing.

Corpus

Our initial corpus comprised 1,631 argumentative essays scored by expert human raters. These essays provide a general representation of writing as found in secondary and early tertiary settings in the United States. The essays were written on 16 different prompts by students in three different grade levels (ninth grade, eleventh grade, and college freshman) and in four different temporal conditions (untimed essays, or essays written under a 10-, 15-, or 25-minute time limit). The majority of these essays have been used in previous studies that have focused on writing quality (Crossley & McNamara, 2010, 2011; Crossley, Weston, McLain Sullivan, & McNamara, 2011 Crossley, Roscoe, McNamara, & Graesser, 2011; McNamara et al., 2010; McNamara et al., 2013; Raine, Mintz, Crossley, Dai, & McNamara, 2011; Wolfe, Britt, & Butler, 2009).

Two or three expert raters independently scored each essay using a 6-point rating scale developed for the SAT. The rating scale (see online appendix) was used to holistically assess the quality of the essays with a minimum score of 1 and a maximum score of 6. Raters were first trained to use the rubric with a small sample of argumentative essays, and a Pearson correlation analysis assessed interrater reliability between raters. When the raters reached a correlation of $r \geq .70$, the ratings were considered reliable and the raters scored a larger subsection of the corpus. The final interrater reliability across all raters for all the essays in the corpus was $r > .70$.

For this analysis, we selected only the essays that received an average score of 4.5 or greater. This selection ensured that each essay received at least one rating of 5 or greater (i.e., scores of 4 and 5 on an essay would be averaged to 4.5) and was, thus, a more successful essay. Of the 1,631 essays in the original corpus, only 148 essays fit the criterion (i.e., 9% of the essays were rated highly). These 148 essays formed the final corpus used in this study. The essays in the final corpus were written on 11 different prompts. In reference to temporal conditions, 33% of the essays were untimed, 3% of the essays were written under 15-minute time constraints, and 64% of the essays were written under 25-minute time constraints. The majority of the essays (76%) were written by college freshmen, and the remaining essays were evenly split between ninth and eleventh grade students (12% each).

Selected Linguistic Features

We used the computational tool Coh-Metrix to calculate the majority of the language features examined in this study. Coh-Metrix represents the state of the art in computational tools and is able to measure text difficulty, text structure, and cohesion through the integration of lexicons, pattern classifiers, part-of-speech taggers, syntactic parsers, shallow semantic interpreters, and other components that have been developed in the field of computational linguistics. Coh-Metrix reports on linguistic variables that are primarily related to text difficulty. In addition, we used the WAT (Crossley et al., 2013) to compute specific linguistic features of student writing related to global cohesion, topic development, collocational accuracy, lexical sophistication, keyword use, and rhetorical features. Finally, we used LIWC tool to report on psychological (social, affective, and cognitive) and personal (leisure, work, religion, home, and achievement) variables. The various constructs measured by these features along with their prototypical features are presented in Table 1.[1] In total, we selected 212 indices from the three tools and placed them into 13 different constructs. More detailed descriptions for the tools and the features on which they report can be found in Graesser, McNamara, Louwerse, and Cai (2004), McNamara et al. (2013), McNamara and Graesser (2012), McNamara, Graesser, McCarthy, and Cai (2014), and Pennebaker et al. (2001).

Statistical Analysis

We initially conducted Pearson product-moment correlations between the selected indices to ensure that none of the indices demonstrated strong multicollinearity (i.e., we examined whether the indices were assessing the same construct). We then used the indices that did not demonstrate multicollinearity in an initial cluster analysis. We used an agglomerative hierarchical cluster analysis with squared Euclidean distances and Ward's method as the distance or similarity measure. This type of analysis uses a mathematical formula (squared Euclidean distances) to automatically classify cases (e.g., essay) into groups that contain shared similarities across features (e.g., the selected linguistic indices). Unlike other approaches (i.e., K-step clustering), agglomerative hierarchical cluster analysis assumes no prior knowledge of cluster membership. The analysis proceeds from a beginning stage in which each case constitutes its own cluster to an end stage in which all of the cases are combined into a single aggregate cluster (Burns, 2008; Norušis, 2011). Between these two stages, the process of forming clusters is iterative, wherein two or more clusters can be combined into a new aggregate cluster, reducing the overall number of clusters (see, e.g., Jarvis et al., 2003). From the initial cluster analysis, we selected the iteration that best represented the number of groups in the cluster by examining changes in the agglomeration coefficients. We than conducted a follow-up cluster analysis using the selected iteration solution. This cluster analysis grouped each essay in the corpus into a cluster.

After the cluster analyses, we conducted confirmatory statistical analyses to assess which linguistics features informed each cluster and to what degree these linguistic features could be used to predict group membership. To examine differences in linguistic features between the clusters, we conducted a multivariate analysis of variance (MANOVA). The MANOVA

Table 1. Selected indices from Coh-Metrix, Linguistic Inquiry and Word Count, and Writing Assessment Tool for Use in Cluster Analysis.

Affective processes	Causal features	Cognitive processes	Cohesion	Content features	Descriptors	Lexical indices
Positive emotions	Causal verbs	Cognitive mechanisms	Connectives	Nouns	Text length	Polysemy
Negative emotions	Causal particles	Tentativeness	Logical operators	Prompt relevance	Adjectives	Lexical diversity
Sadness		Certainty	Lexical overlap	Keywords	Adverbs	Word frequency
Anger		Inhibition	Semantic overlap		Prepositions	Familiarity
Anxiety		Inclusion	Paragraph cohesion			Imageability
		Exclusion	Spatial cohesion			Meaningfulness
			Temporal cohesion			Academic words
						N-gram indices
Perceptual processes		Personal concerns	Rhetorical features	Specificity	Syntax	Verbs
Seeing		Biology	Amplifiers	Hypernymy	Words before the main verb	Verb incidence
Hearing		Relativity	Private verbs	Concreteness	Modifiers per noun phrase	Verb phrases
Feeling		Home	Hedges	Determiners	Passives	Modal verbs
		Money	Indirect pronouns		Relative clauses	Present tense
		Achievement	Exemplification		Syntactic similarity	Past tense
		Religion	Downtoners		Minimal edit distance	Future time
		Death	Paragraph specific n-grams			Verb cohesion

was followed by a discriminant function analysis (DFA). The MANOVA was conducted to assess which variables demonstrated significant differences between the clusters. DFA, which is a common approach used to distinguish text types (e.g., Biber, 1993), generates a discriminant function that acts as an algorithm to predict group membership (i.e., the cluster to which each essay is assigned). We first conducted the DFA on the entire corpus of essays. Subsequently, the DFA model reported from the entire corpus was used to predict group membership of the essays using leave-one-out cross-validation (LOOCV). In LOOCV, a fixed number of folds (i.e., partitions of data) equal to the number of observations (i.e., texts) is selected. Each fold is then used to test the model such that one observation in turn is left out and the remaining instances are used as the training set (in this case the remaining essays). The accuracy of the model is tested on the model's ability to predict the proficiency classification of the omitted instance. This allows us to test the accuracy of the model on an independent data set. If the results of the discriminant analysis in both the entire set and the *n*-fold cross-validation set are similar, then the findings support the extension of the analysis to external data sets.

Analysis

Multicollinearity

We assessed multicollinearity between the 212 selected variables using a threshold of $r > .900$. Of the 212 variables, 15 demonstrated multicollinearity and were removed from the analysis, with 197 variables retained.

Cluster Analysis

We conducted an initial hierarchical cluster analysis using the *z*-scores from the 197 variables in order to determine if successful essays could be separated into individual groups according to shared factors. The initial hierarchical cluster analysis reported an optimal four-cluster solution. This solution was based on the agglomeration schedule, which reported a drop from a 500-point interval to a 300-point interval in the distance coefficient at the point where four clusters combined into five clusters. This drop reflected a discontinuity in the composition from the other clusters (one-, two-, three-, and four-cluster solutions). Such a discontinuity is commonly used to select the optimal cluster solution (Toms et al., 2001; Yuen, 2000). We followed our initial cluster analysis with a second cluster analysis using the selected four-cluster solution, which allowed us to allocate every case to a specific cluster (see Table 2 for clusters and essay assignment).

Table 2. Four-Cluster Solution.

Cluster	*n*
1	46
2	39
3	29
4	34

Multivariate Analysis of Variance (MANOVA)

A MANOVA was conducted to assess which linguistic indices demonstrated differences between the groups reported by the cluster analysis. The MANOVA uses the selected indices as the dependent variables and the essays in each cluster as the independent variables. The MANOVA reported significant differences for 120 of the 197 variables used in the cluster analysis. The *F, p*, and partial-eta squared values for each index, ordered by effect size, are reported in Appendix A.

Mean Score Analysis

For each of the 120 variables, we computed the mean score for each cluster in order to identify the clusters associated with the maximum and minimum scores reported for each variable. These variables thus provide indicators of the most distinctive properties of a given cluster. The variables that loaded positively and negatively in each cluster are discussed below.

Cluster 1. The first cluster was defined by high scores on indices related to *verbal terms, lexical features, rhetorical features, perceptual features*, and *descriptors*. We describe the writing style in this cluster as action and depiction. For instance, essays in the cluster contained higher scores for verb overlap and verb incidence (both part of speech tags and phrasal tags). Essays in this cluster also contained more frequent content words and more rhetorical features such as amplifiers and hedges. In addition, these essays contained more words related to perceptual processes such as seeing and hearing. Last, essays in this cluster contained more descriptive features such as adjectives and adverbs as well as longer paragraphs. Essays in this cluster reported low values for cohesion features such as lexical and semantic overlap. The essays in this cluster also had fewer content words (e.g., nouns and keywords) and fewer personal terms (e.g., words related to family, money, and socialness). An overview of this cluster is presented in Table 3.

Table 3. Cluster 1: Maximum and Minimum Indices.

Indices with maximum score			Indices with minimum score	
Paragraph length	All adjectives (POS/phrases)	MED all words tagged	Argument overlap	Word meaningfulness
Verb overlap (LSA)	Attributive adjectives	Number words	Noun overlap	Word hypernymy
Verb overlap (WordNet)	Predicative adjectives	Perception words	Content word overlap	Word imageability
Third person verbs	Adverbs POS	Amplifiers	LSA sentence to sentence adjacent	Swear words
All verbs	All adverbs (POS/phrases)	Hedges	LSA paragraph to paragraph	Social words
Present tense verbs	All adjectives POS	Impersonal pronouns	LSA sentence to sentence	Family words
Frequency content words	Wh adverbs	Words of time	Words before main verb	Human words
Word familiarity	Hearing words	WHIZ deletion relatives	Keyword proportion	Causal words
Bigram correlation spoken 50,000				Money words

LSA = Latent Semantic Analysis; MED = Minimal Edit Distance; POS = Part of Speech; WHIZ = Wh word plus be deletion.

Cluster 2. The second cluster reported for this analysis was defined by higher incidences of *content words, descriptors, lexical features, personal terms, rhetorical features, specificity*, and *syntactic complexity*. We describe the writing style in this cluster as academic. For instance, the essays in the second cluster had higher scores for essay length, number of sentences, number of paragraphs, prepositions and prepositional phrases, and time adverbials, all of which are related to text descriptiveness. The essays also contained more content words such as nominalizations, nouns, and key types. In addition, the essays contained more lexical features such as greater word meaningfulness, more academic words, longer words, and more frequent n-grams. Linguistically, the essays were more specific as evidenced by higher hypernymy scores, more concrete words, and a greater incidence of determiners. Syntactically, essays in this cluster contained more passives and agentless passives as well as more words before the main verb and longer noun phrases. In addition, the essays contained a greater number of rhetorical features related to introduction and body paragraphs. Last, essays in the cluster reported a greater number of personal terms related to biology, home, relatives, and family. Essays in the second cluster also reported low scores for cohesion, casuality, pronouns, negations, verbal terms, and n-gram frequencies. For instance, essays in this cluster contained fewer causal, additive, and logical connectives. These essays also scored low in causal adverbs, causal verbs, and causal particles as well as in first and third person pronouns. In addition, essays in this cluster contained fewer negations and fewer verb phrases and modals. Last, essays in this cluster scored low in n-gram frequency indices indicating the use of less frequent n-grams (see Table 4 for an overview of Cluster 2).

Cluster 3. Cluster 3 was defined by linguistic features that scored high in *cohesion, causality, content words, lexical features, conclusion n-grams, negations, syntactic complexity, verbal terms*, and *affective words*. We describe the writing style in this cluster as accessible. For instance, essays in this cluster scored high in argument overlap, noun overlap, and content word overlap as well as in LSA semantic coreferentiality indices. The essays also scored high in incidences of connectives such as causal connectives, additive connectives, logical connectives, and all connectives. The essays also contained a greater incidence of causal terms such as causal verbs, particles, and adverbs. The essays also included more content words such as higher incidences of second and third person pronouns and plural nouns. In addition, the essays contained more words related to affect such as general affective words and positive emotion words. The essays in the third cluster also had greater scores for lexical features such as n-grams, word frequency, and polysemy. These essays also contained more negations, more verbal elements (e.g., be verbs, modal, and future terms), as well as greater syntactic complexity (e.g., more question forms and more s-bars). The essays in the third cluster were also defined by low scores in linguistic indices related to lexical features, descriptors, perceptual and personal terms, syntactic features, rhetorical features, and verbs. For instance, these essays contained fewer features related to text description such as fewer sentences, fewer and shorter paragraphs, fewer adjectives, and fewer prepositions. Lexically, the essays contained lower scores of lexical diversity (D and MTLD) and fewer academic words. The essays also contained fewer rhetorical features (introduction n-grams, amplifiers, and hedges), perceptual and personal terms (e.g., terms related to hearing, time, space, and biology), and syntactically complex items (e.g., passive structures). Last, the essays contained verbal properties including overall verbs, verb overlap, perfect verbs, and third person singular verbs (see Table 5 for an overview of Cluster 3).

Cluster 4. The fourth cluster was characterized by a higher incidence of *lexical features, syntactic similarity, content words*, and *personal terms*. We describe the writing style in this cluster as lexical in nature. For instance, lexically, the essays in the fourth cluster contained greater lexical diversity scores (TTR, D, and MTLD) and more imageable words. Syntactically, the essays

Table 4. Cluster 2: Maximum and Minimum Indices.

Indices with maximum score			Indices with minimum score		
Number of words	Number of modifiers per noun phrase	Nominalization	Causative connectives	Trigram proportion spoken	Affect words
Number of sentences	Number of words before main verb	Lexical density	Additive connectives	Trigram frequency spoken	Positive emotion words
Number of paragraphs	Agentless passives	Word concreteness	Logical connectives	Bigram frequency written	Exclusion words
Average sentence length	That subordinator	Word meaningfulness	Density negations	Trigram frequency written	Perception words
Body paragraph n-grams	Perfect aspect incidence	Noun hypernymy	All connectives	Trigram frequency logarithm written	Achievement words
Introduction paragraph n-grams	Prepositional phrases	Word hypernymy	Causative adverbial subordinators	MED all words tagged	Present tense verbs (non–third person)
Key type count	Quantity words	Academic words	Causal verbs and particles	MED all words mean	Incidence verb phrases
Time adverbials	Incidence determiners	Average word length	Causal links	TTR	Do proverbs
Biology words	Swear words	Bigram frequency spoken	First person pronouns	Incidence of all clauses	Wh adverbs
Relativity words	Family words	Bigram correlation written	Third person pronouns	Necessity modals	Analytic negation
Space words	Human words	Noun incidence			
Home words	Money words	Nouns and noun phrases			

MED = Minimal Edit Distance; TTR = Type Token Ratio.

Table 5. Cluster 3: Maximum and Minimum Indices.

Indices with maximum score			Indices with minimum score		
Argument overlap	Density negation	Bigram correlation spoken	Number of sentences	Lexical diversity	Attributive adjectives
Noun overlap	Analytic negation	Trigram proportion spoken	Number of paragraphs	D (lexical diversity)	All adjectives (POS/ phrase)
Content word overlap	Second person pronouns	Trigram frequency spoken	Paragraph length	MTLD (lexical diversity)	Adjective incidence (POS)
LSA sentence to sentence adjacent	Third person pronouns	Trigram frequency logarithm spoken	Sentence length	Academic words	Prepositional phrases
LSA paragraph to paragraph	Future words	Bigram frequency written	Verb overlap (WordNet)	Modifiers per noun phrase	Number words
LSA sentence to sentence all combination	Social words	Trigram frequency written	Verb perfect aspects	Noun incidence	Hearing words
LSA givenness	Affect words	Trigram frequency logarithm written	Verb third person singular	Determiner incidence	Biology words
Causative connectives	Positive emotion words	Frequency all words	All verbs (POS/phrase)	Past participial WHIZ deletion relatives	Relativity words
Additive connectives	Causative words	Word polysemy	Swear words	That subordinators	Space words
Logical connectives	Tentative words	Verb base form	Amplifiers	Agentless passives	Time words
All connectives	Exclusion words	Present tense verbs (non–third person)	Hedges		Work words
Causal verbs and particles	Plural nouns	Wh relative clauses subject	LSA to Prompt	Word concreteness	N-grams introduction paragraphs
Causal links	Keyword proportion	S bars	Key type count		
Causative adverbial subordinators	Conclusion paragraph n-grams	Incidence of all clauses			
Do proverb	Necessity modals				
Be as main verb	All modals				

LSA = Latent Semantic Analysis; MTLD = Measure Text Lexical Diversity; POS = Part of Speech; WHIZ = Wh word plus be deletion.

Table 6. Cluster 4: Maximum and Minimum Indices.

Indices with maximum score	Indices with minimum score		
TTR	Frequency content words	Verb overlap (LSA)	Second person pronouns
D (lexical diversity)	Frequency all words	Predicative adjectives	Impersonal pronoun
MTLD (lexical diversity)	Word familiarity	Adverbs	Future words
Word imageability	Word polysemy	Time adverbials	Quantitative words
First person pronouns	Noun hypernymy	Adverbs incidence (POS)	Tentative words
MED all words mean	Bigram frequency logarithms spoken	Be as main verb	Home words
MED all lemmas	Bigram correlation spoken	Present tense verbs	LSA givenness
LSA to prompt	Bigram correlation written	PRV_VB	Swear words
Work words	N-grams body paragraphs	Modal incidence	Number of words
Achievement words	N-grams conclusion paragraphs	S-bar incidence	
And incidence			
Noun phrase incidence			

LSA = Latent Semantic Analysis; MED = Minimal Edit Distance; MTLD = Measure Text Lexical Diversity; POS = Part of Speech; PRV_VB = Private verb; WHIZ = Wh word plus be deletion.

reported greater MED scores (related to syntactic similarity), and from a content perspective, the essays contained more noun phrases, more first person pronouns, and greater overlap between the prompt and the essay. Last, the essays contained more personal terms related to achievement and work. The cluster was also characterized by lower reported scores for linguistic features related to the lexical features, verbal terms, descriptors, and rhetorical features. For instance, essays in the fourth cluster reported more infrequent words and words that were less familiar, more ambiguous (i.e., lower polysemy scores), and less specific (i.e., lower hypernymy scores). The essays also contained less frequent n-grams. The essays were also less verbal as evidenced by lower incidences of verb forms, future, modals, and LSA verb overlap. Last, the essays had fewer incidences of rhetorical n-grams in body and conclusion paragraphs (see Table 6 for an overview of Cluster 4).

Discriminant Function Analysis (DFA)

We conducted a stepwise DFA to examine if the 120 linguistic indices that demonstrated significant differences between the clusters in the MANOVA could predict the cluster analysis groupings. We conducted a stepwise DFA

Table 7. Descriptive Analysis of Clusters.

Cluster	15 minutes	25 minutes	Unlimited time	Ninth grade	Eleventh grade	College freshman	Number of prompts
1	0.043	0.739	0.217	0.065	0.152	0.783	11
2	0.000	0.000	1.000	0.000	0.000	1.000	4
3	0.000	1.000	0.000	0.069	0.172	0.759	4
4	0.059	0.941	0.000	0.382	0.176	0.441	6

using the entire data set and also using cross-validation methods to assess the strength of the indices to classify essays into clusters.

The stepwise DFA retained 12 of the 120 indices (see Appendix B for the variables that were retained from each cluster). The classification results for the DFA on the entire data set correctly classified 87.8% of the essays as belonging to one of the four clusters as assigned by the cluster analysis, $\chi^2(9) = 316.796$, $p < .001$. The reported kappa of .837 indicated a substantial agreement between the assigned cluster and the predicted cluster. The classification results for the LOOCV set correctly classified 84.5% of the essays as belonging to one of the four clusters as assigned by the cluster analysis. The reported kappa of .792 indicated a substantial agreement. All results were well above chance (chance is 25% for the four groups) and baseline (baseline is 31%; see Appendix C for confusion matrix). These results demonstrate that 12 of the 120 variables that were characteristic of the clusters were able to successfully classify 85% of the essays as falling into the assigned clusters.

Post Hoc Analysis

We conducted a post hoc descriptive analysis of the reported clusters to ensure that the clusters reported by the cluster analysis were not related to temporal conditions, grade level, or prompt. The descriptive analysis is located in Table 7. In general, the time constraints and grade levels were evenly represented in each cluster (recalling that the majority of the essays were written in 25 minutes by college freshman). The exception appears to be Cluster 2, which consists of untimed essays written by college freshman. All clusters included essays written on at least four prompts. All but two prompts were shared by two or more clusters. These results indicate that prompt-based effects were not present.

Discussion

We have demonstrated an approach to assessing successful writing based on cluster analysis and computational linguistics. The approach reveals the variety and linguistic nature of the multiple profiles of successful writing available in the corpus sampled. The findings from the study indicate that there are four profiles of successful writers (or essays) for the samples analyzed. These four profiles are linguistically distinct from one another and demonstrate that expert human raters examine a number of different linguistic features in a variety of combinations when assessing writing quality and assigning high scores to independent essays (regardless of the scoring rubric considered). The writing styles in the four clusters can be described as *action and depiction style, academic style, accessible style*, and *lexical style*. The findings from the study provide evidence that a linear approach to predicting essay quality may not be optimal and may not provide reliable and valid models of writing quality for a diverse population of writers. Overall, the findings from this study have important implications for understanding writing quality, human essay scoring, automatic essay scoring (AES) systems, and automatic writing evaluation (AWE) systems. Below, we provide a summary of the four writing styles identified in the clusters. We then briefly discuss how the writing styles from these clusters could inform the teaching of writing, the training of human raters, and automatic scoring and evaluation systems.

Our cluster analysis provided a four-cluster solution. On average, each cluster contained about 35 essays, with Cluster 1 containing 46 essays and Cluster 3 containing 29 essays. 120 linguistic indices taken from Coh-Metrix, WAT, and LIWC demonstrated significant differences between the four clusters. These indices were used to examine the linguistic elements that distinguished the clusters from one another. A follow-up DFA using these features correctly classified 85% of the essays as belonging to their assigned cluster. Thus, we have strong empirical support that helps confirm the cluster distinctions and strong qualitative support based on the mean difference analysis that helps explain the linguistic properties of each cluster. An overview of the positive and negative features for each cluster is presented in Table 8.

The first writing profile yielded by the cluster analysis (Cluster 1: action and depiction) was strongly verbal and descriptive and tended to be low in cohesive devices and personal terms. Thus, the style of these essays can be described as an action and depiction writing style. For instance, the essays in this cluster contained an increased number of verbs and overlap between verbs throughout the essays providing an indication that these essays are probably more action-oriented than other essays. The essays are also very descriptive with an increased number of adjectives, adverbs, and rhetorical devices likely helping to characterize, depict, or modify the ideas discussed in the essays. Last, the essays were low in the use of cohesive devices. This is not to say that the essays are incoherent because coherence does not always

Table 8. Overview of Linguistic Features for Each Cluster.

Cluster	Positive features	Negative features
1	Verbal terms	Cohesion
	Lexical features	Content words
	Rhetorical features	Personal terms
	Perceptual terms	
	Descriptors	
2	Content words	Cohesion
	Descriptors	Causality
	Lexical features	Pronouns
	Personal terms	Negations
	Rhetorical features	Verbal terms
	Specificity	N-gram frequencies
	Syntactic complexity	
3	Cohesion	Lexical features
	Casuality	Descriptors
	Content words	Perceptual terms
	Lexical features	Personal terms
	Conclusion n-grams	Syntactic complexity
	Negations	Rhetorical features
	Syntactic complexity	Verbal terms
	Verbal terms	
	Affective words	
4	Lexical features	Lexical features
	Syntactic similarity	Verbal terms
	Content words	Descriptors
	Personal terms	Rhetorical features

rely on cohesive devices (e.g., coherence could be the result of background knowledge or reading skill; McNamara, 2013; McNamara, Kintsch, Songer, & Kintsch, 1996), but rather that the essays lack semantic and lexical cohesive devices that help connect ideas from one sentence and paragraph to another by the repetition of words and semantically similar terms. However, as noted previously, there is a tendency for these essays to have strong verbal connections from sentence to sentence. Thus, it is likely, that coherence in these essays is maintained through verbal properties. Last, the essays demonstrated lower incidences of personal terms related to social, family, and human concerns.

The second writing profile yielded by the cluster analysis (Cluster 2: academic) contained the hallmarks of academic writing including strong structural components, strong rhetorical choices, specific word choices, syntactically complex sentences, and the use of infrequent n-gram patterns (Biber, 1988; Brown & Yule, 1983). Thus, we describe this cluster as containing academic style writing. The cluster also provided little causal cohesion, contained few explicit cohesion devices, and used few pronominal references, all indicators of academic writing. Tellingly, all of the essays in this cluster were untimed essays, meaning the writers had time to develop longer essays, select more specific and academic words, develop and use more rhetorical strategies, use more infrequent word combinations, and produce more complicated syntactic structures. Such syntactic structures likely contained embedded cohesive features obviating the need for explicit cohesive devices (Crossley, Weston, et al. 2011; Haswell, 2000). Because the essays were more academic, they likely contained few causal verbs and particles along with pronominal forms, which are more common in narrative writing (Graesser, McNamara, & Kulikowich, 2011; Pearson, 1974-1975). Also, unlike the first cluster, the profile for this cluster reported decreases in verbal terms (verb phrases, present tense verbs). Thus, quality is not defined based on action terms. In addition, unlike the first cluster, the essays contained more terms related to family and human concerns. These indices potentially relate to the use of more specific words.

The third writing profile (Cluster 3: accessible) was strongly set apart from the first two by its higher incidence of cohesion indices, causal features, pronominal referents, and frequent words and n-grams. Such a profile hints at an accessible style of writing that provides a narrative quality, expected word combinations, and a text that is linked through the use of explicit cohesion devices. For instance, the third profile contained almost every index of cohesion available including connectives, lexical overlap, semantic overlap, and givenness. From a narrative perspective, the profile contained a greater incidence of causal connectives, verbs, and particles common to narrative writing (Pearson, 1974-1975) as well as more second and third person pronouns related to narration. Similarly, the cluster contained more social and affective words, which should provide for a more accessible text. The cluster also contained more common words and n-grams (i.e., more high frequency words and n-grams) that would make the essays more lexically accessible. Unlike the first cluster, the third cluster reported lower scores for descriptive indices (adjectives and prepositional phrases) and verbal indices (all verbs and verb overlap). Unlike the profile for Cluster 2, the profile for the third cluster was less academic with fewer structural elements (number of paragraphs and sentences), lower syntactic complexity (agentless passives and length of noun phrases), and lower lexical sophistication (fewer academic words and fewer long words).

The final cluster (Cluster 4: lexical) reported a profile that is best characterized as lexically fluid. Essays in this cluster yielded higher scores for lexical diversity indices, which are often used as proxies for the number of unique words produced. Thus, essays in this cluster contain a greater number of unique words. However, the type of words produced are more imageable and specific while at the same time less sophisticated (as indicated by the incidence of frequent and familiar words that are more polysemous) and occurring in less common word combinations (as attested to by the n-gram indices). Syntactically, unlike the second cluster, the essays provide greater syntactic similarity. The essays are also set apart from the first cluster because they are less descriptive and less verbal.

The results of this study not only help us to better understand successful writing and the multiple profiles associated with successful writing, but also have important implications for the teaching of writing, the training of human experts for essay scoring using standardized rubrics, and informing automatic scoring and essay evaluation systems. In reference to teaching writing, the results of this study could be used to guide instruction about the multiple approaches to writing a successful essay. Such an approach would diverge from the classic academic approach characterized by the linguistic features revealed in Cluster 2 (i.e., writing essays that are syntactically complex and contain sophisticated words). Such instruction could provide teachers with a number of pedagogical choices in terms of writing styles that may better fit a student's perspective, interest, and writing ability.

In addition, the results from this study call into question the use of standardized essay scoring rubrics like those produced by the SAT. Such rubrics generally use a linear approach to writing quality and assume that successful writing samples contain a number of predefined features. For instance, the SAT rubric for persuasive writing (College Board, 2011) includes six levels that consider writers' organization and coherence, language and vocabulary, sentence structure, and mechanics. For example, high scoring essays that receive a score of 6 are classified as "well organized and clearly focused, demonstrating clear coherence and smooth progression of ideas" and exhibiting "skillful use of language, using a varied, accurate, and apt vocabulary" along with "meaningful variety in sentence structure." In contrast, low scoring essays receiving a score of 1 are "disorganized or unfocused, resulting in a disjointed or incoherent essay" and display "fundamental errors in vocabulary" and "severe flaws in sentence structure." Such rubrics do not consider the notion that a combination of features that may or may not include cohesion, syntactic complexity, and lexical sophistication may be used to develop a successful essay. Instead, these rubrics assume a linear approach to essay quality that may lock writers and raters into a single method of developing and scoring a successful writing sample that may not represent the breadth and depth of successful writing samples available.

Linear approaches to writing quality can also be observed in AES systems such as e-rater developed at Educational Testing Service (Burstein, Chodorow, & Leacock, 2004). e-rater uses a linear approach to essay scoring (Attali & Burstein, 2006) that might not capture all the intricacies of successful writing and may instead force test takers into a rigid and systematic approach to writing that may belie naturalistic writing and the individual writing strategies used by said test takers. In our own previous research with an AWE system, we took a linear approach to modeling writing quality for users of the intelligent tutoring system Writing Pal (W-Pal; McNamara et al., 2013; Roscoe & McNamara, 2013). However, more recent scoring algorithms in W-Pal have taken a hierarchical approach that affords more substantial formative feedback to writers in the system. The results of this study demonstrate that AES and AWE systems would benefit from scoring algorithms that can evaluate multiple profiles of writing quality simultaneously.

Conclusion

We present an analysis of successful essays that examines the potential for such essays to exhibit multiple profiles of quality. We find that four distinct profiles of successful writing are available for the sampled essays. These profiles include an action and depiction writing profile, an academic writing profile, an accessible writing profile, and a lexical writing profile. The potential existence of these profiles has important implications for understanding how successful writers write and how the knowledge of such writing can inform classroom pedagogy, the training of expert raters, and the automatic assessment and evaluation of writing samples.

Future extensions of this approach should focus on larger data samples of successful essays from a wider variety of writing genres. The current study was limited by the number of successful essays available and the focus on independent essays only. Future studies should also consider features of writing that may be nonlinguistic. Such features may include strength of arguments, developments of point of view, the use of appropriate examples, grammar, and mechanics or other trait or analytic evaluation. Such extensions would provide a greater understanding of the potential for successful writing samples to exhibit a variety of profiles.

Appendix A

MANOVA Results: Linguistic Difference Among Clusters

Index	*F*	*p*	η^2_p	Index	*F*	*p*	η^2_p
Number of words	66.956	< .001	.582	Adverb phrases	7.938	< .001	.142
Number of sentences	44.237	< .001	.480	And incidence	7.857	< .001	.141
Bigram frequency logarithm spoken	35.000	< .001	.422	Word polysemy	7.868	< .001	.141
Positive emotion words	33.650	< .001	.412	Space words	7.593	< .001	.137
Argument overlap	31.645	< .001	.397	Noun phrase length	7.139	< .001	.129
Affective words	30.897	< .001	.392	Time words	7.098	< .001	.129
TTR	26.796	< .001	.358	Keyword proportion score	6.994	< .001	.127
Word imageability	26.502	< .001	.356	That subordinators	6.889	< .001	.126
Trigram frequency logarithm spoken	25.377	< .001	.346	All adverbs (POS and phrases)	6.878	< .001	.125
Content word overlap	22.143	< .001	.316	Word words	6.847	< .001	.125
Trigram frequency logarithm written	22.126	< .001	.316	Plural nouns	6.509	< .001	.119
Noun overlap	21.671	< .001	.311	Money words	6.437	< .001	.118
Nouns and noun phrases	20.381	< .001	.298	Paragraph length	6.398	< .001	.118
Causal connectives	18.918	< .001	.283	Agentless passives	6.381	< .001	.117
Key type counts	18.417	< .001	.277	Modals	6.306	< .001	.116
Word concreteness	18.273	< .001	.276	Verb overlap WordNet	6.222	< .001	.115
Body paragraph n-grams	17.711	< .001	.270	Human words	6.179	< .001	.114
Logical connectives	17.620	< .001	.269	Family words	5.937	< .001	.110
Word hypernymy	17.382	< .001	.266	Frequency all words	5.954	< .001	.110
LSA givenness	16.289	< .001	.253	LSA to prompt	5.756	< .001	.107
LSA sentence to sentence adjacent	16.268	< .001	.253	Analytic negation	5.686	< .001	.106
Quantity words	15.624	< .001	.246	Causal verbs and particles	5.604	< .001	.105
Word meaningfulness	15.298	< .001	.242	Number words	5.566	< .001	.104
All connectives	14.420	< .001	.231	Social words	5.577	< .001	.104
Adverbs	14.215	< .001	.228	Do verbs	5.507	< .001	.103
Lexical diversity (MTLD)	14.136	< .001	.228	Time adverbials	5.527	< .010	.103
Lexical density	13.325	< .001	.217	Hearing words	5.332	< .010	.100
Number of paragraphs	13.021	< .001	.213	Perception words	5.328	< .010	.100
Lexical diversity (*d*)	12.674	< .001	.209	Introduction paragraph n-grams	5.210	< .010	.098
Causal links	12.633	< .001	.208	Academic words	5.154	< .010	.097
Incidence of nouns	11.956	< .001	.199	Third person singular	5.145	< .010	.097
Trigram proportion spoken	11.350	< .001	.191	All adjectives (POS and phrases)	5.072	< .010	.096
Non–third person singular	11.018	< .001	.187	Adjective incidence	4.839	< .010	.092
Verb phrase incidence	11.060	< .001	.187	Home words	4.635	< .010	.088

(continued)

Appendix A (continued)

Index	*F*	*p*	η^2_p	*Index*	*F*	*p*	η^2_p
Be as main verbs	10.975	< .001	.186	Noun phrases	4.600	< .010	.087
Causative adverbial subordinators	10.650	< .001	.182	Predicative adjectives	4.583	< .010	.087
Prepositional phrases	10.550	< .001	.180	Determiner incidence	4.521	< .010	.086
Third person pronouns	10.301	< .001	.177	First person pronoun	4.453	< .010	.085
Relativity words	10.301	< .001	.177	Attributive adjectives	4.209	< .010	.081
LSA sentence to sentence all	10.144	< .001	.174	Verbs (POS and phrases)	4.222	< .010	.081
Impersonal pronoun	10.003	< .001	.172	Conclusion paragraphs n-grams	4.106	< .010	.079
Average word length	9.864	< .001	.170	Noun hypernymy	3.982	< .010	.077
Bigram frequency written	9.834	< .001	.170	Tentative words	3.943	< .010	.076
Necessity modals	9.677	< .001	.168	Wh relative clauses subject	3.630	< .050	.070
MED all lemmas	9.491	< .001	.165	Hedges	3.461	< .050	.067
Wh adverbs	9.422	< .001	.164	Past participial WHIZ deletion	3.366	< .050	.066
Trigram frequency spoken	9.417	< .001	.164	Words before main verb	3.418	< .050	.066
Bigram correlation written	9.312	< .001	.162	Verb overlap LSA	3.344	< .050	.065
Words of achievement	9.288	< .001	.162	Perfect verbs	3.295	< .050	.064
All clauses incidence	9.236	< .001	.161	Verb base form	3.278	< .050	.064
Content word frequency	9.188	< .001	.161	Word familiarity	3.295	< .050	.064
LSA paragraph to paragraph	9.088	< .001	.159	Density negation	3.218	< .050	.063
Trigram frequency spoken	8.982	< .001	.158	Word length	3.234	< .050	.063
Nominalizations	9.002	< .001	.158	Causal words	3.163	< .050	.062
S-bar incidence	8.906	< .001	.156	Present tense verbs	3.014	< .050	.059
MED all words	8.578	< .001	.152	Second person pronouns	2.915	< .050	.057
Amplifiers	8.366	< .001	.148	Additive connectives	2.919	< .050	.057
MED all words tagged	8.205	< .001	.146	Biology words	2.883	< .050	.057
Exclusion words	8.094	< .001	.144	Swear words	2.753	< .050	.054
Bigram correlation spoken	8.052	< .001	.144	Future words	2.700	< .050	.053

LSA = Latent Semantic Analysis; MED = Minimal Edit Distance; MTLD = Measure Text Lexical Diversity; POS = Part of Speech; TTR = Type Token Ratio; WHIZ = Wh word plus be deletion.

Appendix B

Indices Retained in Discriminant Function Analysis and Their Related Assigned Clusters

Cluster 1	Cluster 2	Cluster 3	Cluster 4
Adverb incidence	Number of words	Argument overlap	Word imageability
Content word frequency	Word length	Positive emotion words	
	Quantity words	Causal connectives	
	Bigrams introduction paragraphs	Bigram frequency logarithm spoken	
		Logical connectives	

Appendix C

Confusion Matrix for Discriminant Function Analysis Classification of Clusters

		Cluster 1	Cluster 2	Cluster 3	Cluster 4
Total set	Cluster 1	**37**	3	2	4
	Cluster 2	2	**37**	0	0
	Cluster 3	1	0	**26**	2
	Cluster 4	3	0	1	**30**
		Cluster 1	Cluster 2	Cluster 3	Cluster 4
Cross-validated set	Cluster 1	**34**	5	4	3
	Cluster 2	2	**37**	0	0
	Cluster 3	1	0	**25**	3
	Cluster 4	3	0	2	**29**

Acknowledgments

We thank various members of the SoLET lab at Arizona State University (Jianmin Dai, Erica Snow, and Laura Varner specifically) and members of the Institute for Intelligent Systems at the University of Memphis (Carl Cai and Art Graesser specifically). We would also like to thank all of our human raters at Mississippi State University and at the University of Memphis.

Declaration of Conflicting Interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: This research was supported in part by the Institute for Education Sciences (IES R305A080589 and IES R305G20018-02). Ideas expressed in this material are those of the authors and do not necessarily reflect the views of the IES.

Note

1. The prototypical features are the most common features measured within each construct. Each feature can contain a number of related indices. For instance, Coh-Metrix calculates four lexical diversity indices (TTR, MTLD, D, and M) and numerous word frequency indices.

Supplemental Material

The online appendix is available at https://sat.collegeboard.org/scores/sat-essay-scoring-guide

References

Antonenko, P. D., Toy, S., & Niederhauser, D. S. (2012). Using cluster analysis for data mining in educational technology research. *Education Technology Research Development*, *60*, 383-398. doi:10.1007/s11423-012-9235-8

Applebee, A. N., Langer, J. A., Jenkins, L. B., Mullis, I., & Foertsch, M. A. (1990). *Learning to write in our nation's schools*. Washington, DC: U.S. Department of Education, Office of Educational Research and Improvement.

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

Bereiter, C., & Scardamalia, M. (1987). *The psychology of written communication*. Hillsdale, NJ: Lawrence Erlbaum.

Berninger, V., Yates, C., Cartwright, A., Rutberg, J., Remy, E., & Abbott, R. (1992). Lower-level developmental skills in beginning writing. *Reading and Writing: An Interdisciplinary Journal*, *4*, 257-280.

Biber, D. (1988). *Variation across speech and writing*. Cambridge, UK: Cambridge University Press.

Biber, D. (1993). Representativeness in corpus design. *Journal of Literary and Linguistic Computing*, *8*(4), 243-257.

Brown, G., & Yule, G. (1983). *Discourse analysis*. Cambridge, UK: Cambridge University Press.

Burns, R. A. (2008). *Business research methods and statistics using SPSS*. Thousand Oaks, CA: Sage.

Burstein, J., Chodorow, M., & Leacock, C. (2004). Automated essay evaluation: The Criterion online writing system. *AI Magazine*, *25*, 27-36.

College Board. (2011). *Essay scoring guide: A framework for scoring SAT essays*. Retrieved from http://professionals.collegeboard.com/testing/satreasoning/scores/essay/guide

Crossley, S. A., & McNamara, D. S. (2010). Cohesion, coherence, and expert evaluations of writing proficiency. In S. Ohlsson & R. Catrambone (Eds.), *Proceedings of the 32nd Annual Conference of the Cognitive Science Society* (pp. 984-989). Austin, TX: Cognitive Science Society.

Crossley, S. A., & McNamara, D. S. (2011). Text coherence and judgments of essay quality: Models of quality and coherence. In L. Carlson, C. Hoelscher, & T. F. Shipley (Eds.), *Proceedings of the 29th Annual Conference of the Cognitive Science Society* (pp. 1236-1241). Austin, TX: Cognitive Science Society.

Crossley, S. A., Roscoe, R. D., McNamara, D. S., & Graesser, A. (2011) Predicting human scores of essay quality using computational indices of linguistic and textual features. In G. Biswas, S. Bull, J. Kay, and A. Mitrovic (Eds.), *Proceedings of the 15th International Conference on Artificial Intelligence in Education*. (pp. 438-440). New York: Springer.

Crossley, S. A., Roscoe, R., & McNamara, D. S. (2013). Using automatic scoring models to detect changes in student writing in an intelligent tutoring system. In P. M. McCarthy & G. M. Youngblood (Eds.), *Proceedings of the 26th international Florida Artificial Intelligence Research Society (FLAIRS) conference* (pp. 208-213). Menlo Park, CA: AAAI Press.

Crossley, S. A., Weston, J., McLain Sullivan, S. T., & McNamara, D. S. (2011). The development of writing proficiency as a function of grade level: A linguistic analysis. *Written Communication*, *28*(3), 282-311.

Deane, P. (2013). On the relation between automated essay scoring and modern views of the writing construct. *Assessing Writing*, *18*, 7-24.

Ericsson, K. A., Charness, N., Feltovich, P. J., & Hoffman, R. R. (2006). *The Cambridge handbook of expertise and expert performance*. New York, NY: Cambridge University Press.

Ferrari, M., Bouffard, T., & Rainville, L. (1998). What makes a good writer? Differences in good and poor writers' self-regulation of writing. *Instructional Science*, *26*, 473-488. doi:10.1023/A:1003202412203

Flesch, R. (1948). A new readability yardstick. *Journal of Applied Psychology*, *32*, 221-233.

Flower, L. S., & Hayes, J. R. (1980). The dynamics of composing: Making plans and juggling constraints. In L. W. Gregg & E. R. Steinberg (Eds.), *Cognitive processes in writing* (pp. 31-50). Hillsdale, NJ: Lawrence Erlbaum.

Geiser, S., & Studley, R. (2001). *UC and SAT: Predictive validity and differential impact of the SAT I and SAT II at the University of California*. Oakland: University of California.

Graesser, A. C., McNamara, D. S., & Kulikowich, J. (2011). Coh-Metrix: Providing multilevel analyses of text characteristics. *Educational Researcher*, *40*, 223-234.

Graesser, A. C., McNamara, D. S., Louwerse, M. M., & Cai, Z. (2004). Coh-Metrix: Analysis of text on cohesion and language. *Behavior Research Methods, Instruments, and Computers*, *36*, 193-202.

Graham, S., & Perry, M. (1993). Indexing transitional knowledge. *Developmental Psychology*, *29*, 779-788.

Haswell, R. H. (2000). Documenting improvement in college writing: A longitudinal approach. *Written Communication*, *17*, 307-352.

Hayes, J. R. (1989). *The complete problem solver*. Hillsdale, NJ: Lawrence Erlbaum.

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

Jarvis, S., Grant, L., Bikowski, D., & Ferris, D. (2003). Exploring multiple profiles of highly rated learner compositions. *Journal of Second Language Writing*, *12*(4), 377-403.

Kellogg, R. T. (1987). Effects of topic knowledge on the allocation of processing time and cognitive effort to writing processes. *Memory and Cognition*, *15*(3), 256-266.

Kellogg, R., & Whiteford, A. (2009). Training advanced writing skills: The case for deliberate practice. *Educational Psychologist*, *44*, 250-266.

Kincaid, J. P., Fishburne, R. P., Rogers, R. L., & Chissom, B. S. (1975). *Derivation of new readability formulas (Automated Readability Index, Fog Count and Flesch Reading Ease Formula) for Navy enlisted personnel* (Research Branch Report 8-75). Millington, TN: Naval Technical Training, U.S. Naval Air Station, Memphis, TN.

King, M., & Rentel, V. (1979). Toward a theory of early writing development. *Research in the Teaching of English*, *13*, 243-253.

McCormick, C. B., Busching, B. A., & Potter, E. F. (1992). Children's knowledge about writing: The development and use of narrative criteria. In M. Pressley, K. R. Harris, & J. T. Gutherie (Eds.), *Promoting academic competence and literacy in school* (pp. 331-336). San Diego, CA: Academic Press.

McCutchen, D. (1986). Domain knowledge and linguistic knowledge in the development of writing ability. *Journal of Memory and Language*, *25*, 431-444.

McCutchen, D., Covill, A., Hoyne, S. H., & Mildes, K. (1994). Individual differences in writing: Implications of translating fluency. *Journal of Educational Psychology*, *86*, 256-266. doi:10.1037/0022-0663.86.2.256

McCutchen, D., & Perfetti, C. (1982). Coherence and connectedness in the development of discourse production. *Text*, *2*, 113-139.

McNamara, D. S. (2013). The epistemic stance between the author and the reader: A driving force in the cohesion of text and writing. *Discourse Studies*, *15*, 1-17.

McNamara, D. S., Crossley, S. A., & McCarthy, P. M. (2010). The linguistic features of quality writing. *Written Communication*, *27*(1), 57-86.

McNamara, D. S., Crossley, S. A., & Roscoe, R. (2013). Natural language processing in an intelligent writing strategy tutoring system. *Behavior Research Methods*, *45*(2), 499-515.

McNamara, D. S., & Graesser, A. C. (2012). Coh-Metrix: An automated tool for theoretical and applied natural language processing. In P. M. McCarthy & C. Boonthum-Denecke (Eds.), *Applied natural language processing and content analysis: Identification, investigation, and resolution* (pp. 188-205). Hershey, PA: IGI Global.

McNamara, D. S., Graesser, A. C., McCarthy, P., & Cai, Z. (2014). *Automated evaluation of text and discourse with Coh-Metrix*. Cambridge, UK: Cambridge University Press.

McNamara, D., Kintsch, E., Songer, N., & Kintsch, W. (1996). Are good texts always better? Interactions of text coherence, background knowledge, and levels of understanding in learning from text. *Cognition and Instruction*, *14*, 1-43.

National Assessment of Educational Progress. (2011). *The nation's report card: Writing 2011*. Retrieved from http://nces.ed.gov/nationsreportcard/writing/

Norušis, M. J. (2011). *IBM SPSS statistics 19 statistical procedures companion*. Upper Saddle River, NJ: Pearson.

Pearson, P. D. (1974-1975). The effects of grammatical complexity on children's comprehension, recall, and conception of certain semantic relationships. *Reading Research Quarterly*, *10*, 155-192.

Pennebaker, J. W., Francis, M. E., & Booth, R. J. (2001). *Linguistic Inquiry and Word Count (LIWC): LIWC2001*. Mahwah, NJ: Lawrence Erlbaum.

Powell, P. (2009). Retention and writing instruction: Implications for access and pedagogy. *College Composition and Communication*, *60*, 664-682.

Raine, R. B., Mintz, L., Crossley, S. A., Dai, J., & McNamara, D. S. (2011). Text box size, skill, and iterative practice in a writing task. In R. C. Murray & P. M. McCarthy (Eds.), *Proceedings of the 24th international Florida Artificial Intelligence Research Society (FLAIRS) conference* (pp. 537-542). Menlo Park, CA: AAAI Press.

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

Roscoe, R. D. & McNamara, D. S. (2013). Writing Pal: feasibility of an intelligent writing strategy tutor in the high school classroom. *Journal of Educational Psychology*, 105(4), 1010-1025.

Schriver, K. A. (1989). Evaluating text quality: The continuum from text-focused to reader-focused methods. *IEEE Transactions on Professional Communication*, *32*(4), 238-255.

Simon, H. A., & Chase, W. G. (1973). Skills in chess. *American Scientist*, *61*, 394-403.

Stewart, M. F. (1978). Syntactic maturity from high school to university: A first look. *Research in the Teaching of English*, *12*(1), 37-46.

Strunk, W., & White, E. B. (1968). *The elements of style*. New York, NY: Macmillan.

Wolfe, C. R., Britt, M. A., & Butler, J. A. (2009). Argumentation schema and the myside bias in written argumentation. *Written Communication*, *26*, 183-209.

Author Biographies

Scott A. Crossley is an associate professor of applied linguistics and ESL at Georgia State University. His research focuses on natural language processing and the application of computational tools in language assessment, second language learning, and text comprehensibility.

Rod Roscoe is an assistant professor of cognitive science and engineering and a research scientist at the Learning Sciences Institute at Arizona State University. His research examines self-regulated learning processes in formal and informal contexts, and examines how these processes can be facilitated via adaptive technology, instruction, and peer support.

Danielle S. McNamara is a senior research scientist at the Learning Sciences Institute and a professor of psychology at Arizona State University. The overarching theme of her research is to better understand cognitive processes involved in comprehension, writing, knowledge acquisition, motivation, and memory, and to apply that understanding to educational practice by developing and testing educational technologies.