Distinguishing Effective Writing Styles in the PERSUADE Corpus

Wesley Morris, Scott Crossley, Langdon Holmes & Joon Suh Choi

Peabody College, Vanderbilt University - Nashville, TN | USA

Abstract: Many linguistic studies of writing assume a single linear relationship between linguistic features in the text and human judgments of writing quality. However, writing quality may be better understood as a complex latent construct that can be constructed in a number of different ways through different linguistic profiles of high-quality writing styles as shown in Crossley et al. (2014). This study builds on the exploratory study reported by Crossley et al. by analyzing a representational corpus of 4,170 highly rated persuasive essays written by secondary-school students. The study uses natural language processing tools to derive quantitative representations for the linguistic features found in the texts. These linguistic features inform a k-means cluster analysis which indicates that a four-cluster profile best fits the data. By examining the indices most and least distinctive of each cluster, the study identifies a structured writing style, a conversational writing style, a reportive writing style, and an academic writing style. The findings support the notion that writers can employ a variety of writing profiles to successfully write an argumentative essay.

Keywords: Writing Styles, Natural Language Processing, Cluster Analysis, SALAT

journal of WRITING RESEARCH

Morris, W., Crossley, S., Holmes, L., & Suh Choi, J. (2025 - accepted for publication). Distinguishing Effective Writing Styles in the PERSUADE Corpus. *Journal of Writing Research*, *volume(issue)*, ##-##. DOI: xx

Contact: Wesley Morris, 230 Appleton Place Peabody Ste 552, Nashville, TN 37240 | USA - wesley.g.morris@vanderbilt.edu. ORCID: 0000-0001-6316-6479

1. Introduction

Writing is an essential skill for educational and professional success, but in 2011, only 27% of eighth and twelfth graders in the United States scored at or above proficient in writing on the National Assessment of Educational Progress (NCES, 2012). Considering these outcomes, an exploration of what constitutes proficient writing is essential to inform secondary school pedagogical practices and interventions. One way that educators and researchers have investigated writing proficiency is through the investigation of linguistic features in student texts. These features are used to predict writing quality and inform pedagogical interventions (Lu et al., 2021). The majority of these studies, however, presume a single linear relationship between linguistic features in the text and essay quality (e.g. Guo et al., 2013; McNamara et al., 2013, 2015).

Fewer studies focus on how observable linguistic features present in an essay may interact in complex ways to construct essay quality as a complex latent variable (i.e., there may be more than one way to write a high-quality essay, Crossley et al., 2014; Jarvis et al., 2003). That is to say, two highly rated essays within the same genre may contain very different patterns of linguistic features which, when seen in totality, construct different but equally effective writing styles. For example, Crossley et al. (2014) demonstrated that linguistic features identified four distinct profiles of highly successful writers. However, the data in Crossley et al. was exploratory and was not representative of developing writers, an important demographic to consider when designing earlier pedagogical interventions. Crossley et al.'s corpus comprised only a small collection of highly rated persuasive essays (N = 148) written by ninth graders, eleventh graders, and college freshmen, with the majority (76%) being college freshmen. The present study builds on this work using a larger corpus of high-quality persuasive essays (N = 4,170) written by secondary school students in the USA. Like Crossley et al. (2014), the current study uses quantitative indices to discern patterns in the linguistic features of highly rated essays by developing writers to explore the different ways in which students can write proficiently.

2. Writing Quality

While the words and language features of a text are manifest and directly observable, its quality is not. As a result, writing quality is a latent or unobserved construct. Traditionally, judgments about writing quality have been ascertained through human ratings, and these ratings of proficiency are considered the gold standard for measurements of writing quality. However, humans do not always agree on the absolute quality of a text. To ensure high inter-rater reliability, various forms of rubric-based rater training have been employed. These rubrics can be holistic, in which a single score is provided for the essay, or analytic, in which several scores are provided measuring different dimensions of writing performance such as grammar/syntax and organization (Moskal, 2000; Wiseman,

2012). Principled analytic rubrics have been found to improve inter-rater reliability significantly (Johnson et. al, 2000) but take significantly longer to utilize and may require more extensive rater training (Weigle, 2007). While rubrics are the most popular forms of writing assessment, other methods have been tested and utilized. These other methods include comparative judgements (Verhavert et al., 2019) in which raters are presented with two texts and asked to choose which text is preferable according to specified criteria. Multifaceted Rasch Measurement (Aryadoust et al., 2021) has also been used to control for inter-rater reliability by weighting scores from raters based on traits of the raters themselves (Crossley et al., 2023).

While human ratings remain the gold standard for writing measurement, several automatic scoring mechanisms have been deployed to model human ratings based on computationally aggregated linguistic features manifest in the text. These linguistic features are calculated using natural language processing (NLP) tools that use syntactic parsers, part-of-speech taggers, word lists from reference corpora, lexicons and other components to compute numerical indices. An example NLP tool is Coh-Metrix (Graesser et al., 2004), a tool which generates over 200 indices of different linguistic features related to cohesion, lexical sophistication, and syntactic parsing. Building on the success of Coh-Metrix, Crossley and Kyle developed the Suite of Automatic Linguistic Analysis Tools (SALAT). SALAT consists of over ten tools that can be used to measure different linguistic features related to cohesion, lexical diversity, lexical sophistication, syntactic parsing, sentiment, cognition, morphology, and readability (Crossley et al., 2016, 2017; Kyle, 2016; Kyle et al., 2018, 2021).

The linguistic features derived from NLP tools can be used as features to train performant statistical models to predict essay scores assigned by humans (e.g. Attali & Burstein, 2006; Crossley & Kim, 2022; Kim & Crossley, 2018; Rudner et al., 2006; Shermis et al., 2010; Warschauer & Ware, 2006). The models used in past studies include linear multiple regression (McNamara et al., 2013), hierarchical classification (McNamara et al., 2015), or Bayesian conditional probabilities between linguistic features and human judgments of quality (McNamara et al., 2017). Such approaches are relatively successful at predicting essay quality. McNamara et al. (2013) developed a regression model from eight predictor variables related to text length, given information, narrativity, lexical sophistication, topicality, and discourse elements specific to conclusion and body paragraphs. The regression model accounted for 46% of the variance in human writing quality ratings and reported a perfect agreement (exact match of human and computer scores) of 44% and adjacent agreement (i.e., within 1 point of the human score) of 94%. Like most statistical models used to predict essay quality, the model reported in McNamara et al. (2013) provides a single linear interpretation of how linguistic features combine to produce a successful essay (McNamara et al., 2015).

3. Writing Strategies and Linguistic Profiles

It has long been understood that writers engage in diverse behavioral patterns while engaging in the writing and revision process, with Schwartz (1983) using classroom observations to posit nine distinct profiles of revision. This early theoretical work was followed by empirical examinations of writer behavior that used unsupervised machine learning methods to find groups of writers with shared characteristics. These methods included clustering algorithms, such as k-means or hierarchical clustering, which are statistical techniques in which responses are sorted into a predetermined number of clusters representing discrete profiles. Cluster analyses were performed on student responses to questionnaires about their behavior before (De Smedt et al., 2022), after (Hartley & Branthwaite, 1989; Torrance et al., 1994), and during a writing task (Torrance et al., 1999). For instance, a longitudinal study using a cluster analysis of questionnaires over time found that the majority of students have a most-used writing strategy that they use 69% of the time (Torrance et al., 2000). Since the development and widespread adoption of word processors, features derived from telemetry data have also been used to cluster writer behavior into distinct profiles (Van Waes & Schellens, 2003; Zhang et al., 2019). The studies above, however, identify writing profiles based on self-reports or observations of writer behavior, rather than the linguistic features observable in the text itself.

Despite the insights derived over decades of studies on profiles of writing behavior, most studies that examine the relationship between linguistic features and language quality have used a single linear statistical model in which certain linguistic features correlate positively or negatively to human judgements. This method presumes that there is only one combination of linguistic features that can explain writing success. However, as seen in the literature on writing behavior, there are many different approaches and constraints used when writing an essay. These may result in linguistic features working together in various ways to construct meaning and argumentation and, as a result, two proficient writers may use different writing strategies on the same task, resulting in different linguistic profiles. While the writing strategies are internal to the writer and can only be revealed through questionnaires or inferred through process data, the linguistic profiles are manifested in the text itself and can be investigated by analyzing textual features.

Specific patterns of linguistic features have long been known to typify language of different modalities (Biber, 1991) and in writing for specific social communicative purposes (Swales, 1990). Thus, linguistic competence can be explained, at least in part, as proficiency with specific genres (Devitt, 2015). This observation has been borne out by research indicating that different lexical patterns are predictive of writing quality in different genres (Olinghouse & Wilson, 2013; Uccelli et al., 2013). Different writing tasks also appear to elicit different linguistic resources. For example, in a study examining writing quality in text-dependent and text-independent essays taken from the Test of English as a Foreign Language, Guo et al. (2013) found that syntactic features are stronger

predictors of success in independent writing tasks as compared to dependent writing tasks while cohesion features are stronger predictors of success in text-dependent writing tasks. Similar differences have been reported for lexical features in text-dependent and text-independent essays (Tywoniw & Crossley, 2019).

In addition to different linguistic profiles for different genres and task types, writers may also produce idiosyncratic linguistic profiles based on preferred writing strategies and differential background knowledge. As a result, two different texts that are judged to be of equal quality may address the same task through different linguistic profiles. Early work on within-task linguistic profiles was reported by Jarvis et al. (2003), who used text characteristics such as text length and average word length, as well as lexical and grammatical features to perform a cluster analysis of two datasets of 178 and 150 highly rated essays by adult English Language Learners (ELLs). The first dataset comprised essays on a single prompt and included a full range of text quality, while the second dataset included essays on two different prompts and included only essays that scored a 3, 4, or 5 on a 6-point scale. The goal of the analysis was to determine whether meaningful writing profiles could emerge from a cluster analysis. Jarvis et al. reported five clusters in the first dataset and three clusters in the second dataset. A limitation of this study was that the clusters correlated strongly with the learner's first language (L1), indicating that the clusters may represent cross-linguistic interference rather than writing profiles. Additionally, Jarvis et al. found that the topic may affect the choice of linguistic features, as one cluster in the second dataset consisted entirely of essays on a single prompt. As a result of these interactions, as well as the small sample sizes and low number of tasks, Jarvis et al. (2003) interpreted their results carefully, refraining from labeling the clusters that emerged.

Crossley et al. (2014) examined the potential to develop linguistic profiles for native speakers of English enrolled in a college composition course. They used the computational tool Coh-Metrix to derive language features from 148 highly rated, independent persuasive essays (i.e., essays that required no source integration) on 11 different prompts from high-school students in ninth and eleventh grade and first-year college students. The Coh-Metrix indices were used to perform a cluster analysis examining the emergence of different writing profiles. The results indicated that high quality essays could be discriminated by their linguistic features into four clusters, each representing a different writing profile. The 'Action and Depiction' profile was typified by present tense verbal terms. The 'Academic' profile included more passive voice and greater phrasal complexity. The 'Accessible' profile integrated a greater number of affective words and demonstrated greater cohesion. The 'Lexical' profile was typified by greater lexical diversity. The findings indicated that there were multiple linguistic profiles observable in successful persuasive essays.

4. Current Study

The current study begins by replicating Crossley et al. (2014) with a larger sample of 4,170 highly rated persuasive essays written by students in middle and high school. Expanding on Crossley et al. (2014), we assess whether the same number of clusters emerge and whether they exhibit similar characteristics on a larger corpus comprising writing samples from a different demographic of writers. We then expand this analysis by assessing whether the derived clusters vary across text-independent and text-dependent writing samples. Our goal is to build on Crossley et al. (2014) by examining the ways in which highly rated texts differ in terms of their linguistic profiles in text-dependent and independent persuasive writing tasks.

To investigate different profiles of successful writing, we conduct a cluster analysis of linguistic features found within the texts using linguistic indices calculated by five NLP tools contained in SALAT. These tools calculate indices related to lexical diversity, cohesion, sentiment, lexical sophistication, and syntactic complexity. We interpret the clusters by examining indices most and least characteristic of each writing profile and validate our analysis with a close read of the essay nearest to each cluster centroid. Thus, our study differs from Crossley et al. (2014) in size, population, writing tasks, and linguistic features examined. The goal is to answer the following research questions:

1. What distinct writing profiles can be discerned from linguistic features explicit to the text?
2. What are the unique features of these distinct profiles in successful writing?
3. How does successful writing differ across text-dependent and independent writing tasks?

5. Methods

5.1 Corpus

Table 1. Demographic Information for PERSUADE Corpus – Total and Score > 4

	All Essays		Successful Writing (score > 4)	
Characteristic	n	%	n	%
Gender				
Female	13,142	50.55	2,369	56.81
Male	12,854	49.45	1,801	43.19
Grade				
Grade 6	1,372	5.28	10	0.24
Grade 8	9,629	37.04	1,095	26.26

Grade 9	2,114	8.13	212	5.08
Grade 10	8,471	32.59	844	20.24
Grade 11	3,461	13.31	7,871	44.87
Grade 12	949	3.65	128	3.31
English Language Learner				
No	22,451	86.36	3,834	91.94
Yes	2,244	8.63	74	1.77
Unknown	1,301	5.01	263	6.31
Race/Ethnicity				
White	11,571	44.51	2,084	49.98
Hispanic/Latino	6,560	25.24	687	16.47
Black/African American	4,959	19.08	582	13.96
Asian/Pacific Islander	1,743	6.71	619	14.84
Two or more races/Other	1,022	3.93	185	4.44
Amer. Indian/AK Native	141	0.54	13	0.31
Economic Disadvantage				
No	11,116	42.76	2,723	65.30
Yes	9,643	37.09	816	19.57
Unknown	5,237	20.15	631	15.13
Disability				
No	21,479	82.62	3,574	85.71
Yes	3,349	12.88	340	8.15
Unknown	1,168	4.49	256	6.14
Total	25,996	100	4170	100

Essays used in the study were sampled from the Persuasive Essays for Rating, Selecting, and Understanding Argumentative and Discourse Elements (PERSUADE) corpus of student persuasive writing (Crossley et al., 2022). The PERSUADE corpus comprises 25,996 essays based on fifteen writing prompts. The essays were selected from a much

larger corpus of around 500,000 essays typed by American students in grades 6-12 in several states across the United States. PERSUADE includes two subcorpora, one of which (n = 12,875) comprises text-dependent essays in which students give their opinion about a text which was provided to them, while the other (n = 12,121) comprises independent writing essays. The text-dependent essays were written by students in grades six through ten and required students to read a source text and integrate that source text into their essay. The independent essays were written by students in grades eight through twelve and required students to write essays on prompts that required no reference to other texts. Essays in the PERSUADE corpus have a minimum of 150 words, of which 75% are spelled according to the conventions of American English. In total, the corpus contains 10,783,494 words, with an average of 402.31 (SD = 188.38) words per essay. The essays were selected to include writers from diverse demographic backgrounds.

Every essay was reviewed by two expert raters from an educational consulting firm with two or more years of experience rating essays for quality. These raters undertook training beforehand to address possible bias. They assigned holistic essay scores of between 1 and 6 to each essay based on the standardized SAT essay rubric, with an inter-rater agreement of r=0.8. After the initial round of rating, a third rater assigned a final adjudicated score to all essays. This paper focuses on successful writing, operationalized as essays with adjudicated holistic scores of greater than four, meaning that at least one reviewer scored the essay a 5 out of 6. Using this threshold, 4,170 essays were retained for analysis. This high-scoring subset of the PERSUADE corpus comprised 2,806,228 words, with an average of 672.96 (SD = 204.45) words per essay. These essays were longer on average than the mean for the whole corpus (M = 402.31, SD = 188.38). Additionally, while the original corpus was balanced between independent and text dependent tasks, 76.5% (n=3,190) of the high-scoring essays were from independent writing tasks while only 23.5% (n=980) were from text-dependent writing tasks. Demographic information for both the entire corpus and the sub-corpus of successful writing are reported in Table 1.

5.2 Linguistic Features

We used five different automated natural language processing (NLP) tools to extract and quantify linguistic features from each text. These tools were the Sentiment Analysis and Cognition Engine (SEANCE; Crossley et al., 2017) which generates statistics on 250 indices related to sentiment analysis, emotion, and cognition, the Tool for the Automated Analysis of Cohesion (TAACO; Crossley et al., 2016) which calculates 169 indices based on type-token ratio, the presence of grammatical participants that have already been mentioned previously in the text, occurrences of semantic and lexical overlap, and the frequency of connectives, the Tool for the Automated Analysis of Lexical Diversity (TAALED; Kyle & Eguchi, 2021) which calculates type/token ratio as well as more sophisticated indices of lexical diversity, the Tool for the Automated Analysis of Lexical Sophistication (TAALES; Kyle & Crossley, 2015) which includes 135 indices of lexical

sophistication, including measures of bigram and trigram frequency, word frequency, the frequency of words that are on academic language word lists, and other psycholinguistic lexical features, and the Tool for the Automated Analysis of Syntactic Structure and Complexity (TAASSC; Kyle, 2016) which generates four groups of indices calculating aspects of syntactic complexity. All tools are open-source and available for free (www.linguisticanalysistools.org). Each of the five tools are discussed in more detail in Appendix A, along with the types of indices that they generate.

5.3 Statistical Analysis

5.3.1 Index Selection

To investigate distinctive writing styles of successful writing in the PERSUADE corpus, the 1,806 indices calculated by these five tools were first pruned to control for statistical assumptions. Many of the features (n=638) reported values of zero for more than 20% of essays, too low to be generalizable to the broader population and thus were removed. Additionally, although k-means clustering is considered fairly robust to non-normal data, it is known to be sensitive to outliers (Gan & Ng, 2017), so 148 indices that reported absolute-value Fisher coefficients of skew higher than two or kurtosis greater than three were also removed. Lastly, 683 indices were found to be only weakly correlated to essay quality measured by holistic essay score ($r < 0.1$; Cohen, 1988, 1992) and were removed. The remaining indices were checked for collinearity, and in cases where the Pearson's product-moment correlation between two or more of the indices was $r > 0.7$, only the index most closely correlated to essay quality was retained. Crossley et. al (2014) used a similar method, with the exception that they set a higher maximum threshold for collinearity ($r > 0.9$) which resulted in a greater proportion of removed features. After pruning, a total of fifty-one indices were available for analysis. These indices along with short descriptions are provided in Appendix B.

5.3.2 Statistical Modelling

After index selection, the remaining indices were z-score normalized for the entire corpus and used as features to conduct a k-means cluster analysis to assess the potential for highly rated essays to have distinct linguistic profiles. A k-means cluster analysis is an algorithm that sorts instances into groups by situating them in a high-dimensional space according to their features (Macqueen, 1967). In the standard algorithm (Hartigan & Wong, 1979), the model is constructed by first manually selecting k, the desired number of clusters. Next, k points are chosen at random to be the cluster centroids and each other point is assigned to a cluster based on its nearest centroid. Then the sum of squared Euclidian distances from each point to its cluster centroid is recorded as the sum of squares, and the centroid is moved to the new center of the cluster. These steps are repeated until the sum of squared distances stabilizes. This study used the k-means algorithm contained in the base R package (v3.6.3, R Core Team, 2021).

The first step in the k-means clustering process is to determine the optimum number of clusters. To determine this, we followed the 'elbow method' outlined by Kodinariya et al. (2013). The elbow method calculates the sum of the squared Euclidian distances from each point to its cluster centroid. This is done for a one-cluster solution then repeated for a two-cluster solution, a three-cluster solution and so on. When graphed, there may be a clear inflection point where increasing the number of clusters no longer has a strong effect on reducing the within-cluster sum of squares. Other methods used for determining the best number of clusters include the information criteria method (Kodinariya et al., 2013) in which Akaike Information Criteria (AIC), a statistic of prediction error which rewards parsimony, is calculated for varying numbers of clusters and an inflection point is detected where information loss begins to level out. Finally, we used cluster plots to flatten the 51 dimensions into a two-dimensional graph, then visually inspected the graph to determine the best number of clusters.

Once we selected an optimal number of clusters, we performed a post-hoc linear discriminant analysis (LDA) to validate the cluster selection. Linear discriminant analysis is a statistical technique commonly used in dimensionality reduction for classification tasks which finds a linear combination of features that best separates the classes while minimizing the variation within each class in a lower-dimensional space (Xanthopoulos et al., 2013). The LDA reduces dimensionality of high dimensional data by calculating new axes that best separate the data points, maximizing the distance between the means of the groups while minimizing the variation within each category. The datapoints are then projected onto the new axes in a way that maximizes the separation of the categories. We generated scatterplots by plotting each essay along each LDA axis and used the graph to verify the separability of the clusters. This validation step was not taken in the original study by Crossley et. al. (2014) but it has been used in similar studies to provide evidence for dimension reduction methods (Omuya et al., 2023). Finally, we used a multi-variant analysis through the manova function in base R to determine whether the differences between the clusters were significant, similarly to the procedure described by Crossley et. al. (2014).

After generating and validating the cluster analysis, we examined the mean z-score for each index in each cluster. We extracted the indices in each cluster which were higher and lower than in any other cluster. These indices provided information what linguistic features are most and least representative of the essays in each profile, and we used the indices to draw conclusions about the profile of successful writing indicated by each cluster. Additionally, we examined the essay closest to each cluster centroid to illustrate and validate our interpretation based on the linguistic features. We further examined the distribution of task type, independent or text-dependent, within each cluster to determine whether each profile is characteristic of a specific task type.

6. Results

The elbow plot in which the mean sum of the squared distance from each data point to its cluster centroid is plotted against the number of clusters can be seen in Panel A of Figure 1. Although the inflection point is not obvious, the graph appears to level out at four or five clusters. Graphs of gap statistics and information criteria also provided support for a four- or five-cluster solution. We also calculated the AIC for between one and ten clusters (see Panel B of Figure 1). A clearer inflection point was reported with four clusters, with additional clusters providing diminishing returns of information. As a result of these analyses, either a four- or five-cluster solution appeared appropriate for the current study.

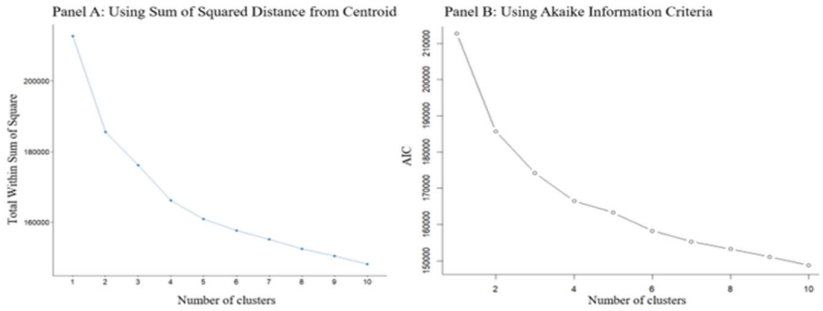

Figure 1: Selecting Best Number of Clusters using Sum of Squared Distance and AIC

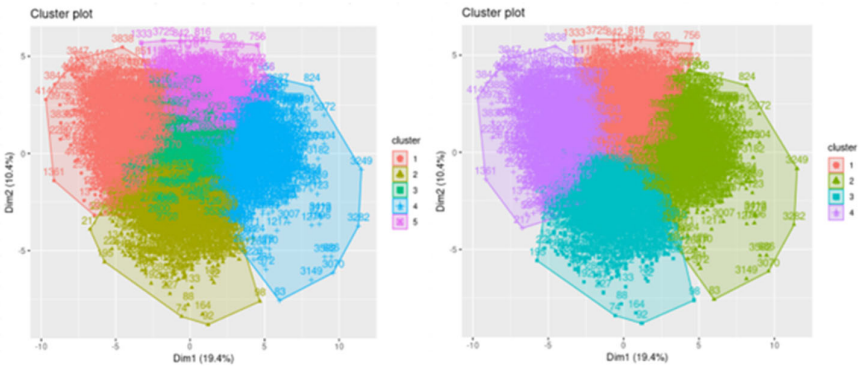

Figure 2: Cluster Plots for Five and Four-Cluster Solutions

Figure 2 displays cluster graphs in which all 51 dimensions are projected into two dimensions. The four-cluster solution cleanly divided the data into four distinct clusters. In the five-cluster solution, the fifth cluster was situated in the center, encompassing the datapoints that did not fit cleanly into any of the four clusters. As a result of this visual inspection of the cluster graph, we selected the four-cluster solution.

Table 2: Four-cluser-solution

Cluster	*n*	*%*	*Independent*	*Text Dependent*
1 – Structural	*1,349*	*32.4*	*1,279 (94.8%)*	*70 (5.2%)*
2 – Academic	*916*	*22*	*798 (87.1%)*	*118 (12.9%)*
3 – Reportive	*1,043*	*25*	*374 (35.9%)*	*669 (64.1%)*
4 – Conversational	*862*	*20.6*	*739 (85.7%)*	*123 (14.3%)*
Total	*4,170*	*100*	*3,190 (76.5%)*	*980 (23.5%)*

Table 2 reports the distribution of the essays into the four clusters, as well as the distribution of the of the clusters by task type. There was a significant difference $\chi^2(3) = 1308.2$, $p < 0.001$ in the distributions of clusters between task types, with Cluster 3 being most representative of text-dependent writing. When examining the differences in linguistic features scores among the clusters, a MANOVA reported a significant difference, Pillai's' Trace = 0.59, $F(51, 4118) = 117.06$, $p < 0.001$. As seen in Table 3, the MANOVA revealed significant differences between the clusters for forty-eight of the fifty-one indices, indicating evidence for four distinct clusters. The MANOVA was followed by a linear discriminant analysis which showed three discriminant functions, making up 60.7%, 33.2%, and 6.1% of the between-class variance respectively for the four clusters. The individual clusters and their interpretation are discussed below.

Table 3: Linguistic Differences Between Clusters – MANOVA Results

Index	F	p	□	Index	F	p	□
Abstract words (GI)	1053.000	<0.001	0.202	Faith (COCA Fic)	81.215	<0.001	0.019
VAC faithfulness SD	92.365	<0.001	0.022	Lemma Freq (COCA fic)	85.624	<0.001	0.020
Academic words (GI)	2860.600	<0.001	0.407	Constr. TTR (COCA fic)	45.588	<0.001	0.011
Paragraph overlap (adv)	47.417	<0.001	0.011	Lemma constr. (COCA fic)	85.767	<0.001	0.020
Sentence overlap (FW)	271.640	<0.001	0.061	Ortho. neighborhood	39.561	<0.001	0.009
Sentence overlap (N)	1241.400	<0.001	0.229	Hostile (GI)	405.940	<0.001	0.089
Paragraph overlap (adv)	135.860	<0.001	0.032	Paragraph overlap (LSA)	1213.200	<0.001	0.225
Affiliation (GI)	258.240	<0.001	0.058	Unigram familiarity	281.170	<0.001	0.063
Lemma Constr. Freq	32.595	<0.001	0.008	Subjects/clause	336.430	<0.001	0.075
Adj/Object of Prep	8.630	0.003	0.002	Objects component	572.720	<0.001	0.121
LDA Age of Onset	771.250	<0.001	0.156	Orthographic neighbors	307.150	<0.001	0.069
Arousal	720.200	<0.001	0.147	Polysemy (adj)	212.810	<0.001	0.049
Dep/Object of Prep	2.603	0.107	0.001	Positivity (GI)	301.830	<0.001	0.068
AWL Sublist 1	257.320	<0.001	0.058	Positivity (EmoLex)	383.520	<0.001	0.084
Num content tokens	188.760	<0.001	0.043	Prep/Nominal Group	0.732	0.392	0.000

BNC spoken bigrams	413.840	<0.001	0.090	Prep/Clause	21.777	<0.001	0.005
Complements/ clause	233.720	<0.001	0.053	Prep/Obj of Prep	0.542	0.462	0.000
CN/T	83.987	<0.001	0.020	Ethics (Lasswell)	155.170	<0.001	0.036
COCA Acad. trigrams	853.220	<0.001	0.170	Semantic variability	1287.100	<0.001	0.236
Coca Fiction trigrams	281.190	<0.001	0.063	SUBTLEXus (all words)	956.530	<0.001	0.187
CP/T	114.130	<0.001	0.027	SUBTLEXus (fun. words)	21.937	<0.001	0.005
Action verbs (GI)	552.700	<0.001	0.117	Synonym overlap (n)	780.440	<0.001	0.158
Doctrine (GI)	49.446	<0.001	0.012	Gain (Lasswell)	307.930	<0.001	0.069
Free association	1234.600	<0.001	0.229	Trust (EmoLex)	169.370	<0.001	0.039
Word naming react time	47.346	<0.001	0.011	Understanding (GI)	479.000	<0.001	0.103
Sentence overlap (w2v)	1024.200	<0.001	0.197				

LSA = Latent semantic analysis, COCA = Corpus of Contemporary American English, TTR = Type Token Ratio, GI = General Inquiry

6.1 Analysis of the Linguistic Profile of each Cluster

After grouping each essay into one of the four clusters based on their indices, we analyzed the characteristic linguistic profile of each cluster by identifying which cluster had the highest and lowest z-score for each index. We grouped the indices under the cluster which reported the highest and lowest mean z-score for that index. These z-scores represent how many standard deviations each cluster mean is from the global mean.

6.1.1 Cluster 1: Structural Writing

The features that comprise this cluster are associated with indices representing coherence and structure. Specifically, essays in Cluster 1 report above average-scores for semantic overlap across paragraphs and sentences based on latent semantic analysis and word2vec. These essays included high degrees of lexical cohesion which served to link their paragraphs together, similar to 'Accessible' from Crossley et. al. (2014). For instance, essays in this cluster make extensive use of cohesive devices such as repeated mentions and adverbial phrases like "similarly" and "for instance". They also utilize anaphoric pronouns in subsequent mentions of a concept that was previously introduced. In the extreme case, Structural texts may be overly repetitive. However, effective writers in this cluster seem to present a tightly connected argument that moves smoothly from point to point.

Table 4. Cluster One: Maximum and Minimum Scores

Cluster One - Structural			
Indices with Maximum Score		Indices with Minimum Score	
Index	Mean-z-score	Index	Mean-z-score
Paragraph overlap (LSA)	0.562	Construction TTR (COCA fic)	-0.230
Sentence overlap (word2vec)	0.536	Objects SÉANCE component	-0.422
Orthographic neighborhood	0.490	Arousal (GI)	-0.445
COCA Academic trigrams	0.479		
Paragraph overlap (adverbs)	0.278		
Paragraph overlap (adverbs)	0.227		

In addition, the essays included more nominalizations to represent abstract ideas, which may also relate to their higher-than-average use of words in the Corpus of Contemporary American English (COCA Davies, 2008) academic corpus. In contrast, the essays were

less likely to contain constructions that are common in fiction and words that indicate emotional arousal. They are less likely to discuss tangible objects. Essays in this cluster may be characterized as being formal with high cohesion and strong organization. Table 4 displays the indices most and least characteristic of Cluster 1. The Structural Writing profile is strongly associated with independent writing tasks, as 40.1% of independent essays were in this cluster compared to only 7.1% of all text-dependent essays. As a result of the high levels of structural cohesion, we designated this cluster as Structural.

A closer look at the first paragraph of the essay nearest to the cluster centroid reveals a strong focus on semantic overlap across sentences (see example below). The flow of the text follows students, teachers, and summer projects with a high degree of lexical overlap from one sentence to the next. This profile of highly structured writing helps to build cohesion and allows the reader to follow the main ideas across the essay. In this excerpt, bolded text is used to show how the text follows the participant **students** through the paragraph

> Projects assigned over summer break are made to help **students** learn more about the subject and prepare **them** for the next year's curriculum. The projects assigned to **students** during summer break should be designed by teachers. This helps to ensure that what **the students** are learning over break will help **them** in the next year, **the students** are learning content that is relevant to school, and to make sure that the teachers know what is being put into those projects.

6.1.2 Cluster 2: Academic Writing

Essays in the second cluster are typified by high phrasal complexity and lexical sophistication. On average, the inverse age of exposure for essays in this cluster was nearly a full standard deviation from the mean, indicating that these essays employ a rich vocabulary. In terms of lexis, the essays are much more likely to contain words on the General Inquirer (GI) word lists for academic subjects and doctrine, meaning organized systems of knowledge, that are commonly discussed in academic settings. As a result, we designated this cluster as Academic style writing. In addition, the essays are more likely to contain words with positive connotations. The essays are much less likely to include action verbs or words with a hostile connotation, instead relying on more objective, academic terminologies. Lastly, the essays also contain bigrams common in spoken modes much less frequently than the mean. This cluster consists of essays that can be best described as academic and lexically dense. Cluster 2 from our data shares features with two clusters from Crossley et. al. (2014). In terms of syntactic complexity, demonstrated by high proportions of complex nominals and phrases per t-unit, it is most like 'Academic'. However, it also shares many of the lexical features with 'Lexical', specifically low scores in polysemy and high scores in lexical features. Table 5 shows indices related to this cluster. Similarly to Cluster 1, independent tasks were over-represented in Academic style writing, as this cluster comprised 25% of independent essays but only 12% of text-dependent essays.

Table 5. Cluster Two: Maximum and Minimum Scores

Cluster Two - Academic			
Indices with Maximum Score		Indices with Minimum Score	
Index	Mean-z-score	Index	Mean-z-score
LDA AOE	0.992	Unigram familiarity	-0.325
Coca Fiction trigrams	0.870	Polysemy (adj)	-0.479
Positivity (EmoLex)	0.868	SUBTLEXus (function words)	-0.499
Free association	0.841	Faith (COCA Fic)	-0.542
Academic words (GI)	0.819	Ethics (Lasswell)	-0.545
Doctrine (GI)	0.741	Lemma Freq (COCA Fic)	-0.548
Gain (Lasswell)	0.735	Complements/clause	-0.577
AWL Sublist 1	0.729	Subjects/clause	-0.577
Complex Nominals/T-unit	0.697	Understanding (GI)	-0.596
Semantic variability	0.694	Lemma construct. (COCA Fic)	-0.645
Positivity (GI)	0.664	SUBTLEXus (all words)	-0.662
Word naming react time	0.661	Hostile (GI)	-0.670
Abstract words (GI)	0.628	Action verbs (GI)	-0.760
Prepositions/Clause	0.607	BNC spoken bigrams	-0.948
Adjectives/Object of Prep	0.556		
Trust (EmoLex)	0.538		
Complex Phrases/T-unit	0.533		
Prep/Obj of Prep	0.531		
Sentence overlap (N)	0.529		
Synonym overlap (n)	0.515		
VAC faithfulness SD	0.499		
Dep/Object of Prep	0.448		
Num content tokens	0.438		
Orthographic neighbors	0.304		
Sentence overlap (FW)	0.276		

The first paragraph of the essay closest to the centroid of Cluster 2 reveals a strong reliance on academic vocabulary to convey the student's ideas. This essay uses a rich lexis at a higher rate than other writing profiles, including abstract words with higher age of exposure and higher word naming reaction time such as innovations and incorporate. The essay also includes multiple complex nominals per sentence on average such as nouns with adjectives (e.g., technological innovations of human history) and nominal clauses (e.g., that students would be able to benefit […]), which aligns with the syntactic complexity typically found in academic writing. In this paragraph, bolded text is added to show the use of academic lexis.

> If we have access to the greatest technological **innovations** of human history, shouldn't we be using them to allow students to choose the way they want to learn? Schools have started to **incorporate** modern technology into the classroom, where schools are surrounded with screens instead of chalkboards and laptops instead of binders. Thus, as technology becomes easier to **integrate** with learning, many schools have started to give the option of learning through online software rather than sitting in a classroom. I believe that students would be able to benefit from learning through online or **video conferencing**, because they would be able to learn in a method that is more convenient, less stressful for students with social issues, and more helpful through the use of learning with online **media**.

6.1.3 Cluster 3 – Reportive Writing

Of the four profiles of successful writing, this one is most strongly associated with text-dependent writing tasks. This cluster comprised 68.2% of all text-dependent essays but only 11.7% of independent essays. Essays in Cluster 3 contain indices related to cohesion that are much lower than the mean, indicating that they are less formally structured than essays in the previous two clusters. Instead of reporting high incidences of cohesive features, these essays contain a high number of prepositional phrases per nominal unit. Lexically, the essays include the least academically oriented vocabulary and avoid words with positive connotations. Instead, the essays contain words that generate emotional arousal and the vocabulary in the essays is more like the lexicon found in fiction reference subcorpora. Because these essays are more likely to include language reporting on language from outside the text, we designated this cluster as 'Reportive' writing. Based on the mean indices of essays in this cluster, it appears to be largely the opposite of Cluster 1. Cluster 3 shares many features with 'Action and Depiction' from Crossley et. al. (2014), specifically its low indices of sentence and paragraph level overlap. Also, essays in our Cluster 3 are more likely to discuss tangible objects, similarly to Crossley et. al. (2014)'s 'Action and Depiction'. Table 6 reports indices distinctive of this cluster.

The first paragraph of the essay most centrally located within the Reportive cluster reveals a high proportion of writing in which the student reports from the text on which the essay is written. It includes a high number of direct quotations, rare in other clusters. The practice of reporting information from another text may lead to some of the lower

indices of cohesion, as the text may be more of a pastiche of facts drawn from the source than a coherent narrative (e.g., According to Source 1…, Heidrun Walter, a media trainer and mother of two says…). The essay also relies significantly less on the use of academic words and abstract words, with the focus of the narrative being on tangible objects and locations (e.g., car, Germany, Europe, the U.S., etc.). In this paragraph, bolded text is added to show the use of reportive language

> In the United States of America, and all over the world, cars are used every day. People use them to get to work, to go see family, and to get simply, from A to B, but a new idea is sprouting up in Europe, the U.S., and elsewhere where people are doing something unheard of.... giving up their cars. **According to Source 1**, "In German Suburb, Life Goes On Without Cars", Vauban, Germany is a city that is almost completely car free. **Heidrun Walter, a media trainer and mother of two says,** "When I had a car I was always tense. I'm much happier this way," This shws that living without cars is not only possible, but could have some great consequences.

Table 6. Cluster Three: Maximum and Minimum Scores

Cluster Three - Reportive			
Indices with Maximum Score		Indices with Minimum Score	
Index	Mean-z-score	Index	Mean-z-score
Objects (SÉANCE) comp.	0.790	Paragraph overlap (adverbs)	-0.320
Arousal	0.683	Trust (EmoLex)	-0.332
Prepositions/Nominal Group	0.621	Num content tokens	-0.352
Lemma Construction Freq	0.530	Paragraph overlap (adverbs)	-0.376
Construction TTR (COCA Fic)	0.349	Sentence overlap (FW)	-0.406
Lemma Freq (COCA Fic)	0.327	Affiliation (GI)	-0.508
		Orthographic neighborhood	-0.546
		Synonym overlap (n)	-0.549
		Abstract words (GI)	-0.587
		Paragraph overlap (LSA)	-0.796
		Sentence overlap (word2vec)	-0.905
		Positivity (GI)	-0.912
		Academic words (GI)	-0.944

LSA = Latent semantic analysis, COCA = Corpus of Contemporary American English, TTR = Type Token Ratio, GI = General Inquiry

6.1.4 Cluster 4: Conversational Writing

Essays in this cluster have high clausal complexity and low phrasal complexity, containing complex clauses with multiple subjects and complement clauses. The essays also tend to include high-frequency words with lower lexical sophistication. Specifically, the essays tend to contain words that are often found in spoken corpora such as SUBTLEXus and the BNC Spoken Corpus instead of academic corpora, and the words found in the essays have the lowest average age of exposure, or self-reported age at which the participant first heard the word. Because of their reliance on lexis commonly found in spoken corpora, we designated this cluster as Conversational. The essays also tend to contain speech acts common in spoken language such as hedging, and they contain a high number of action verbs as compared to sophisticated nominalization. These essays can be characterized as being engaging and conversational. This cluster does not match neatly with any of the clusters presented by Crossley et. al. (2014). Table 7 shows indices most and least characteristic of this cluster. This cluster was like clusters 1 and 2 in that it was more common amongst independent essays, comprising 23.2% of independent essays but only 12.6% of text-dependent essays.

The first paragraph of the essay nearest to the centroid for the Conversational writing cluster demonstrates a story-telling profile commonly employed by essays in this cluster. The essay eschews academic language and structure, instead relying on a conversational tone, often telling personal stories to communicate the theme of the essay. The essay strikes a conversational tone, makes frequent use of personal pronouns, and relies on common words, prioritizing clarity of expression over precision and brevity. In this paragraphbold text is added to show personal pronouns typical of this type of text.

> **I** was stuck in between two decisions, live with **my** mom or live with **my** dad. **I** could ask for help, but would it make a difference if **I** didn't feel happy about it? Asking friends could lead to fights, and if **I** asked parents and step parents they would probably make it some emotional lesson. **I** decided to ask my uncle, Generic_Name, who could relate to me on many occasions. He said that it was up to **me** but he thought I would be happier at **my** dads house. It wasn't enough only being one person, so **I** went to **my** aunt, Generic_Name, who said **my** mom needed me more. I needed tie breaker. **My** last resort was **my** brother, Generic_Name. Although he was not close to **me** at all, he was honest with me and said that **I** should stay with **my** dad. He said that if **I** wanted to be mentally stable and not have **my** clothes carry the stench of smoke form cigarettes, dad was the right option.

Table 7. Cluster Four: Maximum and Minimum Scores

Cluster Four - Conversational			
Indices with Maximum Score		Indices with Minimum Score	
Index	Mean-z-score	Index	Mean-z-score
SUBTLEXus (all words)	0.972	Lemma Construction Freq	-0.248
Affiliation (GI)	0.957	Complex phrases/T-unit	-0.340
Polysemy (adj)	0.871	Orthographic neighbors	-0.348
Lemma construct. (COCA fic)	0.771	Adjectives/Object of Prep	-0.351
BNC spoken bigrams	0.724	Doctrine (GI)	-0.421
Subjects/clause	0.693	Prepositions/Clause	-0.424
Complements/clause	0.679	Gain (Lasswell)	-0.483
Faith (COCA Fic)	0.667	Dep/Object of Prep	-0.486
Understanding (GI)	0.645	COCA Academic trigrams	-0.492
Action verbs (GI)	0.616	Preposition/Obj of Prep	-0.517
Unigram familiarity	0.604	Positivity (EmoLex)	-0.533
Ethics (Lasswell)	0.579	VAC faithfulness SD	-0.533
SUBTLEXus (function words)	0.538	Complex Nominals/T-unit	-0.633
Hostile (GI)	0.463	Sentence overlap (N)	-0.668
		Preposition/Nominal Group	-0.745
		Academic Word List	-0.750
		Word naming react time	-0.789
		Coca Fiction trigrams	-0.851
		Free association	-0.957
		LDA Age of Exposure	-0.981
		Semantic variability	-1.355

LSA = Latent semantic analysis, COCA = Corpus of Contemporary American English, TTR = Type Token Ratio, GI = General Inquiry

7. Discussion

In this study we extracted indices of linguistic features from 4,170 persuasive essays by secondary school students which were highly scored by expert raters. These essays came from two writing tasks (independent and text-dependent writing). We then clustered the essays according to the extracted linguistic indices to examine whether there are multiple profiles of high-quality essays. We used the results of the cluster analysis to extrapolate the profiles of successful writing by identifying the indices most and least characteristic of each profile. Finally, we examined the text of the essays most typical of each cluster, defined as the essays closest to each cluster centroid.

Our first research question asked whether distinct writing profiles could be extracted through a k-means cluster analysis on linguistic features of highly rated persuasive essays. The results support the hypothesis that high quality persuasive essays comprise multiple linguistic profiles. Students use a variety of linguistic resources to write high quality persuasive essays, and these profiles can be inferred through observation and analysis of the observable linguistic features present in their writing. We extracted four clusters, representing different linguistic profiles, based on indices of these linguistic features. We validated these clusters based on AIC, and a post-hoc MANOVA test indicated significant differences among the indices representative of each cluster. The clustering approach was further validated through linear discriminant analysis.

Our second research question asked about the distinctive linguistic features for each of the writing profiles. The cluster analysis results indicated four distinct writing styles for which labels were extrapolated. Essays that employ the Structured style tended to be highly organized and coherent with ideas presented using logical and systematic approaches. In contrast, Reportive style essays were more related to text-dependent writing, using stream-of-consciousness writing styles that often incorporate material from external texts. Academic style essays used a rich lexis and a variety of technical terms to communicate complex ideas, while Conversational style essays used a more informal vocabulary with more high frequency words. However, all were identified by expert raters as high quality. Each of these four writing profiles has their own characteristic linguistic choices, and they can be identified quantitatively through machine learning models.

This study provides support for the findings reported by Crossley et. al. (2014). Specifically, three of the clusters in the current study align closely with those reported in Crossley et al. The cluster identified as Structural in this study is similar to Crossley et. al.'s cluster identified as Accessible. Both clusters emphasized coherence and used lexical cohesion to guide the reader through the essay. Essays in this cluster were the most numerous, comprising 32.4% of the highly rated essays in the corpus. Additionally, the cluster identified as Reportive in this study is closely related to Crossley et. al.'s cluster identified as Action-Depiction. In this dataset, 25% of the essays followed the Reportive writing profile. The cluster in this study that we identified as Academic appears to combine two clusters from Crossley et. al. – Academic and Lexical. The Academic profile

reported here shared the high proportion of complex nominals per t-unit of the Academic profile from Crossley et. al., but it also had the low polysemy scores and high lexical diversity of the Lexical profile from Crossley et al. Academic essays comprised 22% of the essays in the PERSUADE corpus. The current work also identifies a fourth cluster not included in Crossley et al. which we labeled Conversational. This cluster can be characterized by its reliance on words and n-grams that are commonly found in spoken corpora such as SUBTLEXus or the spoken subcorpus of the BNC. Essays of this type were less common, comprising 20.6% of all essays, although this type comprised most of the text-based persuasive essays.

The indices provided by TAALES were most essential for discerning this and other clusters, since they include indices related to the relative occurrence of words in various corpora and sub-corpora, including academic and spoken corpora. Features related to spoken and academic corpora were not available in Coh-Metrix, which Crossley et al. used in their analysis. In addition, the corpus used by Crossley et al. did not include any text-dependent essays, which may help explain why the Reportive style was not represented in their analysis. Finally, the majority (76%) of the persuasive essays in the Crossley et al. corpus were written by college freshmen whereas all essays in the PERSUADE corpus were written by students in middle and high school. This may further explain some of the differences in the results between the two studies.

In answer to the third research question, this study supports previous research (Guo et al., 2013; Tywoniw & Crossley, 2019) indicating that successful writers use different writing styles when executing different writing tasks. Specifically, the Reportive writing profile (Cluster 3) was most prevalent when students were engaged in text-dependent writing, making up 64.1% of essays in this cluster despite comprising only 23.5% of the total corpus of highly rated essays. The high proportion of essays in this cluster may be the result of writers describing the actions of other writers. For example, text-dependent essays are more likely to use phraseology such as, "The author of this paper believes that …" thus partially explaining the prevalence of clausal objects in this cluster. This hypothesis is supported by a reading of essays in the Reportive cluster. Furthermore, these essays may be more likely to quote their source material, which might increase the proportion of lemmas from COCA's fiction subcorpus, especially if the source itself was in a narrative format.

8. Implications

The social purpose of persuasive writing is to convince an audience of some position or activity and the genre of persuasive writing often employs a network of linguistic resources to accomplish this task (Devitt, 2015; Swales, 1990). However, our study indicates substantial variation in the types of linguistic resources that may be effectively brought to bear to accomplish this task. Academic texts may attempt to convince readers through intellectual argumentation, Reportive texts may refer to other texts in an appeal to authority, Structural texts may rely on clear and structured logic, and Conversational

texts might attempt to create solidarity and pathos in the reader. These within-genre individual differences that might be idiosyncratic to the text or to the writer could have important implications in the field of genre analysis.

These findings may also have important implications for the modeling and prediction of writing quality using statistical methods. As statistical algorithms grow in prominence and popularity due to their convenience and affordability, it is important that they be able to capture the nuance of essay quality as a complex latent variable. When utilizing statistical methods, developers should be sure to provide datasets of sufficient depth and diversity so that the models are able to train on corpora of natural language that are rich enough to capture the various profiles of writing that might be deployed by the population of interest. Likewise, models of writing quality should account for the complex interactions between linguistic features in their relationship to writing quality.

In addition to language assessment, our results also have implications in the field of pedagogy and writing instruction in secondary education. Our findings indicate that human raters may perceive very different writing profiles as equally effective. Rather than focusing exclusively on formulaic writing curricula that may presume a standardized construct of writing quality, high school teachers may also choose to encourage students to develop their own voice and experiment with different writing styles that can lead to successful writing (Vengadasalam, 2020; Zhao & Wu, 2022). Indeed, literature on critical pedagogy has highlighted the ways in which formulaic writing standards may serve to entrench existing power structures (Au et al., 2016) and alternative curricula which encourage students to develop their own voice have already been proposed (De Los Rhos, 2020). These results lend support to those theories, showing that a single standard of linguistic competence may be insufficient to describe how humans judge the quality of a text.

9. Limitations and Recommendations

Our study has several limitations which may provide directions for future research. The first is that, while Crossley et al. (2014) looked primarily at college freshmen, our dataset consisted of persuasive essays by students in secondary schools. Because these students are still in the process of learning to write well, this corpus provides valuable insights into the development of writing skills and the linguistic features used by emerging writers. However, our findings may not generalize outside of that age range. Future studies on linguistic features used by more mature writers may uncover different patterns of successful writing by examining writing by mature writers only, including published works. Our study is further limited by its focus on texts that were judged by at least one human to be of at least a five on a six-point scale. A different distinct list of linguistic profiles may emerge when a different set of criteria is used.

Additionally, our study focuses on linguistic profiles that emerge from text-level linguistic features, intentionally refraining from making inferences about person-level writing strategies. Multi-level studies that examine multiple texts nested in writers may

help to discern whether individual writers prefer specific linguistic profiles or whether they adjust their writing styles depending on the task or prompt. It would also be interesting to examine the connection between the cognitive and behavioral writing strategies and the linguistic profiles manifest in the text. While previous research using questionnaires and behavioral data has uncovered that writers employ various planning and revision strategies to address writing tasks (De Smedt et al., 2022; Torrance et al., 2000; Zhang et al., 2019), our study indicates that texts of differing linguistic profiles can be perceived by human raters as being of equal quality. One interesting avenue to pursue would be to determine whether person-level writing strategies predict text-level linguistic profiles.

Finally, the focus on persuasive writing provides a strong overview of writing profiles within this specific genre, but it is likely that other genres of writing have different profiles of successful writing, depending on the social purpose of the genre and the range of expected registers. One avenue of future research may be to focus in on the profiles of effective writing in specific writing tasks such as narrative writing, expository writing, and creative writing. Research in these directions should support the idea that there is more than one way to write well and that writing profiles can be classified using the language features found in texts.

10. Conclusion

In this study, we assessed whether different profiles of successful writing can be uncovered based on linguistic features from texts. We examined the features of each of these profiles of successful writing and analyzed how different writing profiles may be more common in different writing tasks. To do so, we extracted indices of linguistic features from 4,170 highly rated persuasive essays on text-dependent and independent tasks from students in middle and high school. We then used these indices to perform a cluster analysis to search for profiles of successful writing. We found four clusters based on the linguistic features in the texts, three of which (Structural, Academic, and Conversational) were associated with independent writing tasks and were like those reported in Crossley et al., (2014) while one (Reportive) was more associated with text-dependent writing tasks.

This paper supports the findings of previous studies (Crossley et al. 2014; Jarvis et al., 2003) by identifying the construct of writing quality as a complex latent variable dependent on many mutually interacting observable language features, a property of high-quality writing which has important implications for language assessment. If essay quality is a complex construct, models that depend on linear combinations of indices and parameters may be unable to sufficiently describe the quality of an essay. Furthermore, the intensive rater training that is often used to achieve high levels of inter-rater reliability may over-emphasize certain profiles of successful writing at the expense of others which untrained expert readers may also identify as high-quality. We hope that

this study will inform further research into writing styles as well as pedagogical interventions.

References

Aryadoust, V., Ng, L. Y., & Sayama, H. (2021). A comprehensive review of Rasch measurement in language assessment: Recommendations and guidelines for research. Language Testing, 38(1), 6–40. https://doi.org/10.1177/0265532220927487

Attali, Y., & Burstein, J. (2006). Automated essay scoring with e-rater® V. 2. The Journal of Technology, Learning and Assessment, 4(3). https://ejournals.bc.edu/index.php/jtla/article/view/1650

Au, W., Brown, A. L., Calderyn, D., & Dumas, M. (2016). Reclaiming the multicultural roots of U.S. curriculum: Communities of color and official knowledge and education. Teachers College Press. https://doi.org/10.1086/705262

Biber, D. (1991). Variation across speech and writing. Cambridge University Press. https://doi.org/10.1017/cbo9780511621024

Biber, D., Gray, B., & Poonpon, K. (2011). Should We Use Characteristics of Conversation to Measure Grammatical Complexity in L2 Writing Development? TESOL Quarterly, 45(1), 5–35. https://doi.org/10.5054/tq.2011.244483

Cambria, E., & Hussain, A. (2015). SenticNet. In E. Cambria & A. Hussain, Sentic Computing (pp. 23–71). Springer International Publishing. https://doi.org/10.1007/978-3-319-23654-4_2

Chen, D., & Manning, C. D. (2014). A fast and accurate dependency parser using neural networks. Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), 740–750. https://aclanthology.org/D14-1082.pdf

Cohen, J. (1988). Set Correlation and Contingency Tables. Applied Psychological Measurement, 12(4), 425–434. https://doi.org/10.1177/014662168801200410

Cohen, J. (1992). A power primer. Psychological Bulletin, 112(1), 155–159. https://doi.org/10.1037/0033-2909.112.1.155

Coxhead, A. (2000). A New Academic Word List. TESOL Quarterly, 34(2), 213. https://doi.org/10.2307/3587951

Crossley, S. A., Baffour, P., Tian, Y., Picou, A., Benner, M., & Boser, U. (2022). The persuasive essays for rating, selecting, and understanding argumentative and discourse elements (PERSUADE) corpus 1.0. Assessing Writing, 54, 100667. https://doi.org/10.1016/j.asw.2022.100667

Crossley, S. A., & Kim, M. (2022). Linguistic Features of Writing Quality and Development: A Longitudinal Approach. The Journal of Writing Analytics, 6(1), 59–93. https://doi.org/10.37514/JWA-J.2022.6.1.04

Crossley, S. A., Kyle, K., & Dascalu, M. (2019). The Tool for the Automatic Analysis of Cohesion 2.0: Integrating semantic similarity and text overlap. Behavior Research Methods, 51(1), 14–27. https://doi.org/10.3758/s13428-018-1142-4

Crossley, S. A., Kyle, K., & McNamara, D. S. (2016). The tool for the automatic analysis of text cohesion (TAACO): Automatic assessment of local, global, and text cohesion. Behavior Research Methods, 48(4), 1227–1237. https://doi.org/10.3758/s13428-015-0651-7

Crossley, S. A., Kyle, K., & McNamara, D. S. (2017). Sentiment Analysis and Social Cognition Engine (SEANCE): An automatic tool for sentiment, social cognition, and social-order analysis. Behavior Research Methods, 49(3), 803–821. https://doi.org/10.3758/s13428-016-0743-z

Crossley, S. A., & McNamara, D. S. (2010). Cohesion, Coherence, and Expert Evaluations of Writing Proficiency. Proceedings of the Annual Meeting of the Cognitive Science Society, 32(32). https://escholarship.org/uc/item/6n5908qx

Crossley, S. A., & McNamara, D. S. (2014). Does writing development equal writing quality? A computational investigation of syntactic complexity in L2 learners. Journal of Second Language Writing, 26, 66–79. https://doi.org/10.1016/j.jslw.2014.09.006

Crossley, S. A., Roscoe, R., & McNamara, D. S. (2011). Predicting Human Scores of Essay Quality Using Computational Indices of Linguistic and Textual Features. In G. Biswas, S. Bull, J. Kay, & A. Mitrovic (Eds.), Artificial Intelligence in Education (Vol. 6738, pp. 438–440). Springer Berlin Heidelberg. https://doi.org/10.1007/978-3-642-21869-9_62

Crossley, S. A., Roscoe, R., & McNamara, D. S. (2014). What Is Successful Writing? An Investigation Into the Multiple Ways Writers Can Write Successful Essays. Written Communication, 31(2), 184–214. https://doi.org/10.1177/0741088314526354

Crossley, S. A., Tian, Y., Baffour, P., Franklin, A., Kim, Y., Morris, W., Benner, M., Picou, A., & Boser, U. (2023). The English Language Learner Insight, Proficiency and Skills Evaluation (ELLIPSE) Corpus. International Journal of Learner Corpus Research, 9(2), 250–271. https://doi.org/10.1075/ijlcr.22026.cro

Crossley, S., & McNamara, D. (2016). Say more and be more coherent: How text elaboration and cohesion can increase writing quality. Journal of Writing Research, 7(3 (February 2016)), 351–370. https://doi.org/10.17239/jowr-2016.07.3.02

Dascalu, M., McNamara, D., Crossley, S., & Trausan-Matu, S. (2016). Age of Exposure: A Model of Word Learning. Proceedings of the AAAI Conference on Artificial Intelligence, 30(1). https://doi.org/10.1609/aaai.v30i1.10372

Davies, M. (2008). The Corpus of Contemporary American English (COCA). https://www.english-corpora.org/coca/

De Los Rhos, C. V. (2020). Writing Oneself Into the Curriculum: Photovoice Journaling in a Secondary Ethnic Studies Course. Written Communication, 37(4), 487–511. https://doi.org/10.1177/0741088320938794

De Smedt, F., Landrieu, Y., De Wever, B., & Van Keer, H. (2022). Do cognitive processes and motives for argumentative writing converge in writer profiles? The Journal of Educational Research, 115(4), 258–270. https://doi.org/10.1080/00220671.2022.2122020

Devitt, A. J. (2015). Genre performances: John Swales' Genre Analysis and rhetorical-linguistic genre studies. Journal of English for Academic Purposes, 19, 44–51. https://doi.org/10.1016/j.jeap.2015.05.008

Gaies, S. J. (1980). T-Unit Analysis in Second Language Research: Applications, Problems and Limitations. TESOL Quarterly, 14(1), 53. https://doi.org/10.2307/3586808

Gan, G., & Ng, M. K.-P. (2017). K -means clustering with outlier removal. Pattern Recognition Letters, 90, 8–14. https://doi.org/10.1016/j.patrec.2017.03.008

Garner, J., Crossley, S., & Kyle, K. (2019). N-gram measures and L2 writing proficiency. System, 80, 176–187. https://doi.org/10.1016/j.system.2018.12.001

Gonz6lez, M. C. (2017). The Contribution of Lexical Diversity to College□Level Writing. TESOL Journal, 8(4), 899–919. https://doi.org/10.1002/tesj.342

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, & Computers, 36(2), 193–202. https://doi.org/10.3758/BF03195564

Guo, L., Crossley, S. A., & McNamara, D. S. (2013). Predicting human judgments of essay quality in both integrated and independent second language writing samples: A comparison study. Assessing Writing, 18(3), 218–238. https://doi.org/10.1016/j.asw.2013.05.002

Halliday, M. A. K., & Hasan, R. (1976). Cohesion in English (0 ed.). Routledge. https://doi.org/10.4324/9781315836010

Hammill, D. D., Mather, N., Allen, E. A., & Roberts, R. (2002). Using Semantics, Grammar, Phonology, and Rapid Naming Tasks to Predict Word Identification. Journal of Learning Disabilities, 35(2), 121–136. https://doi.org/10.1177/002221940203500204

Hartigan, J. A., & Wong, M. A. (1979). Algorithm AS 136: A K-Means Clustering Algorithm. Applied Statistics, 28(1), 100. https://doi.org/10.2307/2346830

Hartley, J., & Branthwaite, A. (1989). The psychologist as wordsmith: A questionnaire study of the writing strategies of productive British psychologists. Higher Education, 18(4), 423–452. https://doi.org/10.1007/BF00140748

Hunt, K. W. (1965). A Synopsis of Clause-to-Sentence Length Factors. The English Journal, 54(4), 300. https://doi.org/10.2307/811114

Jarvis, S. (2013). Capturing the Diversity in Lexical Diversity: Lexical Diversity. Language Learning, 63, 87–106. https://doi.org/10.1111/j.1467-9922.2012.00739.x

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. https://doi.org/10.1016/j.jslw.2003.09.001

Jnoub, N., Al Machot, F., & Klas, W. (2020). A Domain-Independent Classification Model for Sentiment Analysis Using Neural Models. Applied Sciences, 10(18), 6221. https://doi.org/10.3390/app10186221

Jung, Y. J., Crossley, S., & McNamara, D. (2019). Predicting Second Language Writing Proficiency in Learner Texts Using Computational Tools. The Journal of AsiaTEFL, 16(1), 37–52. https://doi.org/10.18823/asiatefl.2019.16.1.3.37

Kim, M., & Crossley, S. A. (2018). Modeling second language writing quality: A structural equation investigation of lexical, syntactic, and cohesive features in source-based and independent writing. Assessing Writing, 37, 39–56. https://doi.org/10.1016/j.asw.2018.03.002

Klein, D., & Manning, C. D. (2003). Accurate unlexicalized parsing. Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics, 423–430. https://doi.org/10.3115/1075096.1075150

Kodinariya, T. M., Makwana, P. R., & others. (2013). Review on determining number of Cluster in K-Means Clustering. International Journal, 1(6), 90–95.

Kyle, K. (2016). Measuring syntactic development in L2 writing: Fine grained indices of syntactic complexity and usage-based indices of syntactic sophistication [Doctoral Dissertation].

Kyle, K., & Crossley, S. (2016). The relationship between lexical sophistication and independent and source-based writing. Journal of Second Language Writing, 34, 12–24. https://doi.org/10.1016/j.jslw.2016.10.003

Kyle, K., & Crossley, S. A. (2015). Automatically Assessing Lexical Sophistication: Indices, Tools, Findings, and Application. TESOL Quarterly, 49(4), 757–786. https://doi.org/10.1002/tesq.194

Kyle, K., & Crossley, S. A. (2018). Measuring Syntactic Complexity in L2 Writing Using Fine-Grained Clausal and Phrasal Indices. The Modern Language Journal, 102(2), 333–349. https://doi.org/10.1111/modl.12468

Kyle, K., Crossley, S. A., & Jarvis, S. (2021). Assessing the Validity of Lexical Diversity Indices Using Direct Judgements. Language Assessment Quarterly, 18(2), 154–170. https://doi.org/10.1080/15434303.2020.1844205

Kyle, K., Crossley, S., & Berger, C. (2018). The tool for the automatic analysis of lexical sophistication (TAALES): Version 2.0. Behavior Research Methods, 50(3), 1030–1046. https://doi.org/10.3758/s13428-017-0924-4

Kyle, K., & Eguchi, M. (2021). 6 Automatically Assessing Lexical Sophistication Using Word, Bigram, and Dependency Indices. In S. Granger (Ed.), Perspectives on the L2 Phrasicon (pp. 126–151). Multilingual Matters. https://doi.org/10.21832/9781788924863-007

Lasswell, H. D., & Namenwirth, J. Z. (1969). The Lasswell value dictionary. New Haven.

Laufer, B., & Nation, P. (1995). Vocabulary Size and Use: Lexical Richness in L2 Written Production. Applied Linguistics, 16(3), 307–322. https://doi.org/10.1093/applin/16.3.307

Levy, R., & Andrew, G. (2006). Tregex and Tsurgeon: Tools for querying and manipulating tree data structures. LREC, 2231–2234.

Liu, B. (2022). Sentiment analysis and opinion mining. Springer Nature. https://www.cs.uic.edu/~liub/FBS/liub-SA-and-OM-book.pdf

Lu, X. (2010). Automatic analysis of syntactic complexity in second language writing. International Journal of Corpus Linguistics, 15(4), 474–496. https://doi.org/10.1075/ijcl.15.4.02lu

Lu, X., Casal, J. E., & Liu, Y. (2021). Towards the Synergy of Genre- and Corpus-Based Approaches to Academic Writing Research and Pedagogy: International Journal of Computer-Assisted Language Learning and Teaching, 11(1), 59–71. https://doi.org/10.4018/IJCALLT.2021010104

MacArthur, C. A., Jennings, A., & Philippakos, Z. A. (2019). Which linguistic features predict quality of argumentative writing for college basic writers, and how do those features change with instruction? Reading and Writing, 32(6), 1553–1574. https://doi.org/10.1007/s11145-018-9853-6

Macqueen, J. (1967). Some Methods for Classification and Analysis of Multivariate Observations. Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability, 1, 281–297.

McCarthy, P. M., & Jarvis, S. (2010). MTLD, vocd-D, and HD-D: A validation study of sophisticated approaches to lexical diversity assessment. Behavior Research Methods, 42(2), 381–392. https://doi.org/10.3758/BRM.42.2.381

McNamara, D. S., Allen, L. K., Crossley, S. A., Dascalu, M., & Perret, C. A. (2017). Natural Language Processing and Learning Analytics. In C. Lang, G. Siemens, A. Wise, New York University, USA, & D. Gasevic (Eds.), Handbook of Learning Analytics (First, pp. 93–104). Society for Learning Analytics Research (SoLAR). https://doi.org/10.18608/hla17.008

McNamara, D. S., Crossley, S. A., & McCarthy, P. M. (2010). Linguistic Features of Writing Quality. Written Communication, 27(1), 57–86. https://doi.org/10.1177/0741088309351547

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. https://doi.org/10.3758/s13428-012-0258-1

McNamara, D. S., Crossley, S. A., Roscoe, R. D., Allen, L. K., & Dai, J. (2015). A hierarchical classification approach to automated essay scoring. Assessing Writing, 23, 35–59. https://doi.org/10.1016/j.asw.2014.09.002

Medhat, W., Hassan, A., & Korashy, H. (2014). Sentiment analysis algorithms and applications: A survey. Ain Shams Engineering Journal, 5(4), 1093–1113. https://doi.org/10.1016/j.asej.2014.04.011

Mohammad, S. M. (2016). Sentiment Analysis. In Emotion Measurement (pp. 201–237). Elsevier. https://doi.org/10.1016/B978-0-08-100508-8.00009-6

Mohammad, S. M., & Turney, P. D. (2013). Crowdsourcing a Word-Emotion Association Lexicon. Computational Intelligence, 29(3), 436–465. https://doi.org/10.1111/j.1467-8640.2012.00460.x

Moskal, B. M. (2000). Scoring Rubrics: What, When and How? Practical Assessment, Research and Evaluation, 7(3). https://doi.org/10.7275/A5VQ-7Q66

Muangkammuen, P., & Fukumoto, F. (2020). Multi-task learning for automated essay scoring with sentiment analysis. Proceedings of the 1st Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 10th International Joint Conference on Natural Language Processing: Student Research Workshop, 116–123. https://aclanthology.org/2020.aacl-srw.17/

Myhill, D. (2008). Towards a Linguistic Model of Sentence Development in Writing. Language and Education, 22(5), 271–288. https://doi.org/10.1080/09500780802152655

Nakayama, M., Sears, C. R., & Lupker, S. J. (2008). Masked priming with orthographic neighbors: A test of the lexical competition assumption. Journal of Experimental Psychology: Human Perception and Performance, 34(5), 1236–1260. https://doi.org/10.1037/0096-1523.34.5.1236

National Center for Educational Statistics. (2012). The Nation ‘s Report Card: Writing 2011. https://nces.ed.gov/nationsreportcard/pdf/main2011/2012470.pdf

Olinghouse, N. G., & Wilson, J. (2013). The relationship between vocabulary and writing quality in three genres. Reading and Writing, 26(1), 45–65. https://doi.org/10.1007/s11145-012-9392-5

Omuya, E. O., Okeyo, G., & Kimwele, M. (2023). Sentiment analysis on social media tweets using dimensionality reduction and natural language processing. Engineering Reports, 5(3), e12579. https://doi.org/10.1002/eng2.12579

Ortega, L. (2003). Syntactic Complexity Measures and their Relationship to L2 Proficiency: A Research Synthesis of College-level L2 Writing. Applied Linguistics, 24(4), 492–518. https://doi.org/10.1093/applin/24.4.492

Perin, D., & Lauterbach, M. (2018). Assessing Text-Based Writing of Low-Skilled College Students. International Journal of Artificial Intelligence in Education, 28(1), 56–78. https://doi.org/10.1007/s40593-016-0122-z

R Core Team. (2021). R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing. https://www.R-project.org/

Richards, B. (1987). Type/Token Ratios: What do they really tell us? Journal of Child Language, 14(2), 201–209. https://doi.org/10.1017/S0305000900012885

Rudner, L. M., Garcia, V., & Welch, C. (2006). An evaluation of IntelliMetricTM essay scoring system. The Journal of Technology, Learning and Assessment, 4(4). https://ejournals.bc.edu/index.php/jtla/article/view/1651

Saito, K., Webb, S., Trofimovich, P., & Isaacs, T. (2016). Lexical Profiles of Comprehensible Second Language Speech: The Role of Appropriateness, Fluency, Variation, Sophistication, Abstractness, and Sense Relations. Studies in Second Language Acquisition, 38(4), 677–701. https://doi.org/10.1017/S0272263115000297

Salsbury, T., Crossley, S. A., & McNamara, D. S. (2011). Psycholinguistic word information in second language oral discourse. Second Language Research, 27(3), 343–360. https://doi.org/10.1177/0267658310395851

Schwartz, M. (1983). Revision Profiles: Patterns and Implications. College English, 45(6), 549. https://doi.org/10.2307/377139

Seyoum, W. M., Yigzaw, A., & Bewuketu, H. K. (2022). Students' Attitudes and Problems on Question-Based Argumentative Essay Writing Instruction. Journal of English Language Teaching and Learning, 3(2), 58–63. https://doi.org/10.33365/jeltl.v3i2.2106

Shermis, M. D., Burstein, J., Higgins, D., & Zechner, K. (2010). Automated essay scoring: Writing assessment and instruction. International Encyclopedia of Education, 4(1), 20–26. https://doi.org/10.1016/b978-0-08-044894-7.00233-5

Sinclair, J. (1991). Corpus, concordance, collocation (4. impr). Oxford Univ. Pr. https://doi.org/10.2307/330144

Stone, P. J., Dunphy, D. C., & Smith, M. S. (1966). The general inquirer: A computer approach to content analysis. https://psycnet.apa.org/record/1967-04539-000

Struthers, L., Lapadat, J. C., & MacMillan, P. D. (2013). Assessing cohesion in children's writing: Development of a checklist. Assessing Writing, 18(3), 187–201. https://doi.org/10.1016/j.asw.2013.05.001

Swales, J. (1990). Genre analysis: English in academic and research settings (1. publ., 14. print). Cambridge Univ. Pr. https://doi.org/10.2307/416471

Torrance, M., Thomas, G., & Robinson, E. (1999). Individual differences in the writing behaviour of undergraduate students. British Journal of Educational Technology, 69, 189–199. https://doi.org/10.1348/000709999157662

Torrance, M., Thomas, G. V., & Robinson, E. J. (1994). The writing strategies of graduate research students in the social sciences. Higher Education, 27(3), 379–392. https://doi.org/10.1007/BF03179901

Torrance, M., Thomas, G. V., & Robinson, E. J. (2000). Individual differences in undergraduate essay-writing strategies: A longitudinal study. Higher Education, 39(2), 181–200. https://doi.org/10.1023/A:1003990432398

Torruella, J., & Capsada, R. (2013). Lexical Statistics and Tipological Structures: A Measure of Lexical Richness. Procedia - Social and Behavioral Sciences, 95, 447–454. https://doi.org/10.1016/j.sbspro.2013.10.668

Tywoniw, R., & Crossley, S. (2019). The effect of cohesive features in integrated and independent L2 writing quality and text classification. Language Education and Assessment, 2(3), 110–134. https://doi.org/10.29140/lea.v2n3.151

Uccelli, P., Dobbs, C. L., & Scott, J. (2013). Mastering Academic Language: Organization and Stance in the Persuasive Writing of High School Students. Written Communication, 30(1), 36–62. https://doi.org/10.1177/0741088312469013

Van Waes, L., & Schellens, P. J. (2003). Writing profiles: The effect of the writing mode on pausing and revision patterns of experienced writers. Journal of Pragmatics, 35(6), 829–853. https://doi.org/10.1016/S0378-2166(02)00121-2

Vengadasalam, S. S. (2020). Transformative Pedagogy and Student Voice: Using S.E.A. Principles in Teaching Academic Writing. Journal of Effective Teaching in Higher Education, 3(2), 12–27. https://doi.org/10.36021/jethe.v3i2.95

Verhavert, S., Bouwer, R., Donche, V., & De Maeyer, S. (2019). A meta-analysis on the reliability of comparative judgement. Assessment in Education: Principles, Policy & Practice, 26(5), 541–562. https://doi.org/10.1080/0969594X.2019.1602027

Warschauer, M., & Ware, P. (2006). Automated writing evaluation: Defining the classroom research agenda. Language Teaching Research, 10(2), 157–180. https://doi.org/10.1191/1362168806lr190oa

Weigle, S. C. (2007). Teaching writing teachers about assessment. Journal of Second Language Writing, 16(3), 194–209. https://doi.org/10.1016/j.jslw.2007.07.004

Wiseman, C. S. (2012). A comparison of the performance of analytic vs. Holistic scoring rubrics to assess L2 writing. International Journal of Language Testing, 2(1), 59–92.

Xanthopoulos, P., Pardalos, P. M., & Trafalis, T. B. (2013). Linear Discriminant Analysis. In P. Xanthopoulos, P. M. Pardalos, & T. B. Trafalis, Robust Data Mining (pp. 27–33). Springer New York. https://doi.org/10.1007/978-1-4419-9878-1_4

Zhang, M., Zhu, M., Deane, P., & Guo, H. (2019). Identifying and Comparing Writing Process Patterns Using Keystroke Logs. In M. Wiberg, S. Culpepper, R. Janssen, J. Gonzőlez, & D. Molenaar (Eds.), Quantitative Psychology (Vol. 265, pp. 367–381). Springer International Publishing. https://doi.org/10.1007/978-3-030-01310-3_32

Zhao, C. G., & Wu, J. (2022). Perceptions of authorial voice: Why discrepancies exist. Assessing Writing, 53, 100632. https://doi.org/10.1016/j.asw.2022.100632

Appendix A: Description of Linguistic Analysis Tools

SEANCE. Sentiment analysis, or opinion mining, is a technique often employed to predict consumer choices, but it also represents important elements of a writer's distinctive style, especially in persuasive writing (Liu, 2022). Sentiment analysis is often performed by a bag-of-words method, collecting vector representations of words and phrases in a text (Medhat et al., 2014). These vector representations can then be used in a classification task by a machine learning model trained on the target domain. Such classification models perform well within the target domain but may not generalize outside of it (Hussein, 2018). Alternatively, the model can compare the bag-of-words representation to a domain-independent sentiment dictionaries that consist of labelled vectors, such as General Inquirer (GI; Stone et al., 1966), EmoLex (Mohammad & Turney, 2013), and SenticNet (Cambria & Hussain, 2015). While less accurate within a specific domain, the domain-independent approach has been found to be robust for general use (Jnoub et al., 2020). Sentiment analysis has been used to investigate affect, valence, and opinions in student writing (Mohammad, 2016; Seyoum et al., 2022) and measures of sentiment have been found to help improve the accuracy of automatic essay scoring tools according to a holistic rubric (Muangkammuen & Fukumoto, 2020).

The Sentiment Analysis and Cognition Engine (SEANCE) is a linguistic analysis tool which generates statistics on 250 indices related to sentiment analysis, emotion, and cognition (Crossley et al., 2017). These analyses are primarily conducted by converting the words of a text into a numeric representation called an embedding or a vector. The vectors associated with each word are drawn from open-source databases such as SenticNet (Cambria & Hussain, 2015) and EmoLex (Mohammad & Turney, 2013). Additionally, older word lists such as General Inquirer (GI; Stone & Kirsch, 1966) and the Lasswell Value Dictionary (Lasswell & Namenwirth, 1969) are used, in which lists of words are organized into semantic categories such as positivity or ethics. Frequencies of words included in these lists are computed to derive indices of each of the categories.

TAACO. Cohesion refers to the linguistic resources that are used to connect linguistic elements within or between texts and is an important way to build a sense of coherence (Halliday & Hasan, 1976). Measures of cohesion are strong predictors of essay quality in essays written by young writers (Struthers et al., 2013). Specifically, Myhill (2008) found that the use of adverbials by students in year 8 was positively correlated to measures of essay quality, but by year 10 the correlation was no longer significant. Other studies on the relationship between cohesion features and essay quality on adult learners achieved mixed results. Studies using Coh-Metrix found significant positive relationships between referential cohesion and essay quality (MacArthur et al., 2019) as well as negative relationships between essay quality and argument overlap, a measure of referential cohesion (Perin & Lauterbach, 2018). Other studies have found no significant effect (Crossley & McNamara, 2010; McNamara et al., 2010). Despite these ambiguous results, there is strong theoretical evidence that cohesion is important to text quality and more

nuanced analyses have demonstrated that features of global cohesion are related to text quality (Crossley et al., 2011) and that modifying student essays to improve global cohesion leads to significantly increased measures of essay quality (Crossley & McNamara, 2016).

The Tool for Automatic Analysis of Cohesion (TAACO) was developed to collect indices specifically related to cohesion (Crossley et al., 2016). TAACO was later updated to collect pseudo-semantic indices based on state-of-the-art word embeddings such as word2vec and latent semantic analysis (Crossley et al., 2019). The current version of TAACO calculates 169 indices based on type-token ratio, the presence of grammatical participants that have already been mentioned previously in the text, occurrences of semantic and lexical overlap, and the frequency of connectives.

TAALED. Lexical diversity is a measure of the number of unique words relative to the number of total words in a text (Jarvis, 2013). While lexical diversity can be calculated most simply as type-token ratio (TTR), or the ratio of unique words to the total number of words (Richards, 1987), this approach has been found to overstate lexical diversity in shorter texts (Jarvis, 2013). As a result, several other measures have been developed, including mean segmental type-token ratio (MSTTR), or the average type-token ratio over subsamples of a given number of words in a text (Torruella & Capsada, 2013). Another approach to overcome the sample size problem, such as the HD-D measure, examines the probability encountering the same token twice in a sample of text (McCarthy & Jarvis, 2010). Finally, MTLD is calculated as the mean length of sequential tokens that fall above a given TTR value (McCarthy & Jarvis, 2010). These approaches come some ways in addressing the problem of calculating lexical diversity among texts of variable lengths. Measures of lexical diversity show medium to strong correlations with human judgments of lexical diversity (Kyle et al., 2021) and studies indicate that lexical diversity has a greater impact than word frequency on human evaluations of essay quality among English learners (Gonz6lez, 2017).

The Tool for Automatic Analysis of LExical Diversity (TAALED) calculates type/token ratio as well as more sophisticated indices of lexical diversity such as Moving Average TTR (MATTR), and the Measure of Textual Lexical Diversity (MTLD) (Kyle & Eguchi, 2021). These additional indices are valuable because of the problems with using type/token ratio as a measure of lexical diversity in a corpus consisting of texts that are of variable length (Jarvis, 2013). Each of these indices are calculated for all words, and they are separately calculated for the diversity of function words and content words in the text. In total, TAALED provides 38 indices of lexical diversity.

TAALES. Lexical sophistication has long been thought to be a predictor of writing quality. Sophisticated words have historically been defined as lower frequency words, with a writer's Lexical Frequency Profile representing the percentage of words used by the writer at different frequency levels (Laufer & Nation, 1995). More recently, the construct of

lexical sophistication has been expanded to include words commonly found in academic texts (Coxhead, 2000) and words that are more abstract (Saito et al., 2016; Salsbury et al., 2011). Lexical sophistication may also be expanded to include the use of low-frequency or typically academic phrases consisting of multiple words known as n-grams (Sinclair, 1991), with research indicating that the proportion of academic n-grams predicts human ratings of writing quality (Garner et al., 2019). Psycholinguistic properties of words can also serve as a measure of lexical sophistication. For example, words that elicit a greater response time before being recognized as words have been shown to be more sophisticated (Kim & Crossley, 2018). Previous studies have shown that greater lexical sophistication as measured by proportion of academic words is positively correlated with writing quality (Kyle & Crossley, 2015). Psycholinguistic features of lexical items including age of acquisition, imageability, and familiarity have also been used to model holistic essay scores (Kyle & Crossley, 2016). Studies have indicated that lexical sophistication may correlate more strongly with essay quality in text-dependent tasks than independent writing tasks (Kyle & Crossley, 2016).

The Tool for Automatic Analysis of Lexical Sophistication (TAALES) includes 135 indices of lexical sophistication, including measures of bigram and trigram frequency, word frequency, the frequency of words that are on academic language word lists, and other psycholinguistic lexical features (Kyle & Crossley, 2015). These indices compare the relative frequencies of words, bigrams, and trigrams in a text to a variety of references lists and corpora such as the Academic Word List, the academic and fictional sub-corpora of the Corpus of Contemporary American English (COCA; Davies, 2008), the spoken sub-corpus of the British National Corpus, and the SUBTLEXus corpus of television and movie subtitles. These indices can give an idea not only of the diversity of words, but also of the registers the vocabulary most resembles. In addition to comparing the text to reference corpora, TAALES includes indices based on psycholinguistic studies such as rapid naming tasks in which the speed at which subjects were able to read the word aloud was timed in order to develop an idea of the orthographic complexity of the word (Hammill et al., 2002) and number of orthographic neighbors, words that can be produced by changing just one letter (Nakayama et al., 2008). It also calculates an inverse age of exposure metric for each word, which expresses the average self-reported age at which a learner has been exposed to enough context to be able to understand the word (Dascalu et al., 2016).

TAASSC. Syntactic complexity, or the syntactic nestedness of language in a text, has a somewhat complicated relationship with text quality (Crossley & McNamara, 2014). T-units, the shortest grammatical units that can be punctuated at the level of the sentence, have long been used to measure the syntactic complexity of a text ever since their introduction by Hunt (1965), with longer T-units often being associated with greater syntactic complexity (Gaies, 1980). The use of mean words per T-unit as a measure of syntactic complexity has been criticized as overly simplistic. It is difficult to interpret

(Ortega, 2003), and it fails to distinguish between phrasal complexity and clausal complexity (Kyle & Crossley, 2018), the first being more common in academic English and the second more common in spoken English (Biber et al., 2011). In response, more fine-grained NLP tools have been introduced in order to evaluate the frequency of specific components of syntactic complexity, such as the number of dependent clauses per T-unit as a specific measure of clausal complexity and complex nominals per T-unit as a specific measure of phrasal complexity (Lu, 2010). Many such indices are calculated by the Tool for the Automated Assessment of Syntactic Sophistication and Complexity (TASSC; Kyle & Crossley, 2018). Measures of syntactic complexity have been shown to predict essay quality with increased phrasal complexity, for example mean length of noun phrases (Jung et al., 2019) and diverse syntactic structures (Ortega, 2003), predicting higher ratings. Conversely, measures indicating reduced syntactic complexity such as incidents of declarative sentences have shown negative correlations with essay scores (McNamara et al., 2013).

The Tool for Automatic Assessment of Syntactic Sophistication and Complexity (TAASSC), developed by Kyle (2016) generates four groups of indices which calculate aspects of syntactic complexity. The first group are the SCA indices developed by Lu (2010), which count structures such as complex nominals per t-unit and complex phrases per t-unit using the Stanford Parser (Klein & Manning, 2003) and Tregex (Levy & Andrew, 2006). The other three groups of indices use the Stanford Neural Network Dependency Parser (Chen & Manning, 2014) to count features such as adjectival modifiers within a prepositional object or dependents of a prepositional object. In addition to providing counts and frequencies of syntactic structures, it also compares these measures to various subcorpora representing different registers of English in reference corpora such as COCA (Davies, 2008).

Appendix B: List and Descriptions of Indices

Index	Tool	Description
Abstract words (GI)	SEANCE	Frequency of 276 abstract words
Academic words (GI)	SEANCE	Frequency of 153 words associated with academics
Affiliation (GI)	SEANCE	Frequency of 557 words indicative of affiliation
Arousal	SEANCE	Frequency of words denoting arousal
Action verbs (GI)	SEANCE	Frequency of 540 descriptive action verbs
Doctrine (GI)	SEANCE	Frequency of 217 words describing organized systems of belief
Hostile (GI)	SEANCE	Frequency of 833 words indicative of hostility
Objects (SEANCE) component	SEANCE	Principal component score indicating objects
Positivity (GI)	SEANCE	Frequency of 1915 words associated with positivity
Positivity (EmoLex)	SEANCE	Positive emotion words
Ethics (Lasswell)	SEANCE	Frequency of 151 words associated with ethics
Gain (Lasswell)	SEANCE	Frequency of 129 words associated with accomplishment
Trust (EmoLex)	SEANCE	Trust emotion words
Understanding (GI)	SEANCE	Frequency of 309 words expressing caution
Paragraph overlap (adverbs)	TAACO	Adverb overlap across two paragraphs
Paragraph overlap (FW)	TAACO	Function word overlap across two paragraphs
Sentence overlap (N)	TAACO	Noun overlap across two sentences
Paragraph overlap (adverbs)	TAACO	Adverb overlap across two paragraphs (binary)
Sentence overlap (word2vec)	TAACO	Semantic overlap across sentences based on word2vec
Paragraph overlap (LSA)	TAACO	Semantic overlap across paragraphs based on LSA
Synonym overlap (n)	TAACO	Noun synonym overlap across paragraphs
Num content tokens	TAALED	Number of content word tokens

LDA AOE	TAALES	Latent Dirichlet Allocation Age of Exposure (inverse average)
AWL Sublist 1	TAALES	Academic Word List Sublist 1
BNC spoken bigrams	TAALES	BNC Spoken Bigram Frequency Logarithm
COCA Academic trigrams	TAALES	Prop. of trigrams among 30,000 most frequent in COCA Academic
Free association	TAALES	Num of word types arising in response to the word in free association
Word naming react time	TAALES	Standard deviation of mean word naming reaction time
Orthographic neighborhood	TAALES	Mean frequency of orthographic neighborhood
Unigram familiarity	TAALES	Mean word familiarity score
Orthographic neighbors	TAALES	Mean number of orthographic neighbors
Polysemy (adj)	TAALES	Measure of adjective polysemy
Semantic variability	TAALES	Semantic variability of contexts in which word occurs
SUBTLEXus (all words)	TAALES	Frequency of words in SUBTLEXus corpus
SUBTLEXus (function words)	TAALES	Score for average range of function words in SUBTLEXus (log)
VAC faithfulness SD	TAASSC	Standard dev. of faith score for verb construction in academic texts
Lemma Construction Freq	TAASSC	Average lemma construction frequency all (types)
Adjectives/Object of Prep	TAASSC	Adjectival modifiers per object of preposition
Dependents/Object of Prep	TAASSC	Dependents per object of preposition not including pronouns
Complements/clause	TAASSC	Complement clauses per clause
Coca Fiction trigrams	TAASSC	Trigram Bigram to Unigram Association Strength (delta P)
Faith (COCA fic)	TAASSC	Average faith verb construction score (COCA fiction)
Lemma Freq (COCA fic)	TAASSC	Average lemma frequency (COCA fiction)
Construction TTR (COCA fic)	TAASSC	Type-token ratio for constructions (COCA fiction)
Lemma constructions (COCA fic)	TAASSC	Percent of lemma constructions in the reference corpus (COCA fiction)
Subjects/clause	TAASSC	Nominal subjects per clause

Prep/Nominal Group	TAASSC	Prepositions per nominal group
Prep/Clause	TAASSC	Prepositions per clause
Prep/Obj of Prep	TAASSC	Prepositions per object of preposition
CN/T	TAASSC (L2SCA)	Complex nominals per t-unit
CP/T	TAASSC (L2SCA)	Complex phrases per t-unit

FW = Function Words, GI = General Inquiry, SD = Standard Deviation, LSA = Latent semantic analysis, COCA = Corpus of Contemporary American English, TTR = Type Token Ratio, GI = General Inquiry