

John Benjamins Publishing Company



This is a contribution from *Multi-Dimensional Analysis, 25 years on. A tribute to Douglas Biber*.

Edited by Tony Berber Sardinha and Marcia Veirano Pinto.

© 2012. John Benjamins Publishing Company

This electronic file may not be altered in any way.

The author(s) of this article is/are permitted to use this PDF file to generate printed copies to be used by way of offprints, for their personal use only.

Permission is granted by the publishers to post this file on a closed server which is accessible to members (students and staff) only of the author's/s' institute, it is not permitted to post this PDF on the open internet.

For any other use of this material prior written permission should be obtained from the publishers or through the Copyright Clearance Center (for USA: www.copyright.com).

Please contact rights@benjamins.nl or consult our website: www.benjamins.com

Tables of Contents, abstracts and guidelines are available at www.benjamins.com

A Multi-Dimensional analysis of essay writing

What linguistic features tell us about situational parameters and the effects of language functions on judgments of quality

Scott A. Crossley¹, Laura K. Allen² & Danielle S. McNamara²

¹Georgia State University / ²Arizona State University

This study applied the Multi-Dimensional analysis used by Biber (1988) to examine the functional parameters of essays. Co-occurrence patterns were identified within an essay corpus (n = 1529) using linguistic indices provided by Coh-Metrix. These patterns were used to identify essay groups that shared features based upon situational parameters. Results revealed that the linguistic features reliably co-occurred according to the parameters. Namely, four dimensions were interpreted and associated with essay quality, prompt, and grade level. Confirmatory analyses revealed that the dimensions reliably distinguished among the parameters. Results provide insight into the situational parameters that affect writing, and the extent to which essays vary among and between themselves. The results have important implications for composition pedagogy, writing assessment, and writing theory.

1. Introduction

Twenty-five years ago Biber (1988) investigated differences between speech and writing in his landmark book *Variation across speech and writing*. From this and previous investigations (Biber 1985,1986), an approach for assessing linguistic variation within texts called Multi-Dimensional analysis was born. The hallmarks of this approach included the use of large-scale text corpora, the sampling of texts from a variety of conditions, the automatic computation of a number of linguistic features, and the use of statistical analyses to identify relations among the conditions. The book spurred a revolution in quantitative corpus linguistic studies and led to hundreds (if not thousands) of Multi-Dimensional analyses

based on Biber's approach. These analyses were applied to a number of areas well beyond differences in speech and writing, such as language disabilities (Gregg, Coleman, Stennett & Davis 2002), bilingual creativity (Baker & Eggington 1999), grant proposals (Connor & Upton 2004), and world Englishes (Xiao 2009), to name but a few.

In the current study, like Biber (1988), we use a Multi-Dimensional approach; however, our emphasis is not on distinguishing between registers or genres. Rather, we are interested in determining co-occurrence patterns in linguistic features as they relate to first language (L1) argumentative essay writing. The context for this essay writing is with English as a first language writers writing within the American education system. Knowing that texts are related across specific situational and functional parameters, this study seeks to identify parameters in argumentative essays that will afford a better understanding of how writing situations, such as the prompt and the author's grade level, influence the co-occurrence of linguistic features. We are also interested in understanding how functional parameters, such as cohesion, propositional meaning, temporality, syntactic complexity, and word choice, affect human judgments of essay quality. Thus, unlike Biber (1988), our goal is not to distinguish differences between speech and writing, but to determine how linguistic elements can be used to reliably identify consistent co-occurrence patterns within the specific genre of argumentative essays.

The identification of such co-occurrence patterns would be beneficial for a variety of reasons. First, it would help first language (L1) and second language (L2) writers make more informed decisions throughout the writing process by providing them with detailed descriptions of quality written products. Second, it would give teachers a better understanding of the situational factors related to writing and, in turn, help them address such factors in classroom assignments and assessments. Finally, such knowledge would help teachers make more accurate and specific evaluations of writing quality, which would enable them to provide more precise and targeted feedback to struggling writers.

1.1 Essay writing

As Biber (1988) notes, once a community develops a writing system, there is no reason to assume that the system should be treated secondarily to speech. This is particularly true for modern societies in which writing is an essential element of academic and professional success. From an academic perspective, writing is a major component of high-stakes tests that require higher-order writing skills (Jenkins, Johnson & Hileman 2004). Underachievement in such high-stakes tests can have important consequences for the test taker, including failure to advance in grade level or to graduate high school (in the case of American state and federal

testing), failure to enter undergraduate programs (in the case of the Scholastic Aptitude Test, a test commonly used in the United States for college entrance), and failure to enter graduate school (in the case of the Graduate Record Examination). In addition to academic consequences, the failure to develop sufficient writing skills has other real-life consequences, including difficulty in articulating ideas, arguing opinions, and synthesizing multiple perspectives. These skills are essential for accurate and persuasive communication with peers, teachers, co-workers, and the community at large (Connor 1987; Crowhurst 1990; National Commission on Writing 2004).

While we know that writing is a critical skill necessary for academic and professional success (Kellogg & Raulerson 2007), large-scale assessments often show that writing proficiently is difficult for many students to accomplish (National Commission on Writing, NCW 2003). Thus, studies that provide a clearer understanding of writing proficiency, writing quality, and the effects of varying writing conditions have the potential to increase writing performance among students. By pinpointing how writing develops, how expert raters assess writing quality, and what effects writing tasks have on the writing product, researchers and educators can generate opportunities for extended practice in specific problem areas and, in turn, guide individualized feedback for students that serves to remediate these targeted areas.

1.2 Writing development

Knowledge of writing development is a key element for understanding and assessing writing proficiency. A common approach for investigating writing development is to compare writers of different grade levels (e.g. Berninger, Cartwright, Yates, Swanson & Abbott 1994; Crossley, Weston, Sullivan, & McNamara 2011; O'Donnell, Griffin & Norris 1967; Perfetti & McCutchen 1987). Such analyses can provide crucial information about how writing skills change as neural, cognitive, and linguistic functions develop (Berninger, Mizokawa & Bragg 1991). Studies that have focused on the differences among grade levels have shown that writing skills first develop with the mastery of producing legible letters and basic spelling (Abbott, Berninger & Fayol 2010). They then move toward mastery of basic grammar and sentence structure, followed by a mastery of the lexicon (Freedman & Pringle 1980; Haswell 2000). Writing continues to develop over time with writers next focusing on text cohesion (McCutchen 1986; Witte & Faigley 1981), then syntactic structures (Hunt 1965; McCutchen & Perfetti 1982), and lastly cognitive strategies such as planning and revising (Abbott et al. 2010; Berninger et al. 1991).

A major change in writing development seems to occur around high school with the trade-off between the use of explicit cohesive devices and the use of more

complex syntactic structures. Studies have demonstrated that the use of cohesive devices slows or decreases at the high school and college level. For instance, Freedman and Pringle (1980) found no differences between essays written by high school students and third-year college students in terms of textual unity, organization, development, or coherence. Similarly, Crossley et al. (2011) found that the incidence of cohesive devices declined as grade level advanced from 9th to 11th grade and from 11th grade to college. Finally, research reveals that, in the workforce, expert writers rely less on explicit markers of text cohesion, producing fewer referential links (pronouns, demonstratives, and comparatives), lexical overlap, and logical connectors than college undergraduate writers (Haswell 1986).

Conversely, as the use of cohesive devices declines, the use of more complex syntactic structures increases (Berninger et al. 2010; McCutchen & Perfetti 1982; Stewart 1978). For instance, Haswell (2000) found that college juniors tend to write longer sentences with longer clauses than college freshmen and Crossley et al. (2011) reported that the production of syntactically complex structures increased as a function of grade level from 9th grade until the freshman year of college.

1.3 Essay quality

Another approach used to investigate the characteristics of writing proficiency is to examine human judgments of writing quality. Such judgments are generally made by professional readers (e.g. teachers) who are trained to score essays based upon pre-determined rubrics. These judgments have important consequences for writers, because they generally determine passing or failing grades in high school and college. Most commonly, these judgments are investigated through an examination of the linguistic structures that best predict human ratings (e.g. McNamara, Crossley, & McCarthy 2010; Witte & Faigley 1981). For instance, McNamara et al. (2010) used a variety of linguistic features to examine differences between low- and high-rated untimed essays written by college freshman. Their results demonstrated that the three most predictive indices of essay quality were increased syntactic complexity, increased lexical diversity, and decreased word frequency.

In a similar fashion, Crossley, Roscoe, and McNamara (2011) used a number of linguistic indices to predict human ratings of essay quality on a corpus of timed essays written by college freshman composition students. The results demonstrated that 43% of the variance in human scores of essay quality was predicted by essay length (more word types), greater redundancy (higher given/new), fewer personal pronouns, more infrequent words, more expressions associated with higher quality conclusions (conclusion n-grams), and more complex syntax (fewer base

verb forms). Comparable findings have been reported in research concentrating on second language (L2) writing. Crossley and McNamara (2012), for instance, reported that the linguistic indices that distinguished high-quality L2 essays from low-quality L2 essays were mostly related to linguistic sophistication, with higher-quality L2 essays containing more lexical diversity, more infrequent words, less meaningful words, and less familiar words.

2. Method

In this study, we use a Multi-Dimensional analysis to examine the situational parameters of L1 argumentative essays. The essays in this analysis vary as a function of the following parameters: essay score, prompt, grade level of the writer, geographic area, production (handwritten or typed), and timing (i.e. length of time given to write the essay). Our grouping features are linguistic indices taken from the computational tool Coh-Metrix (Graesser, McNamara, Louwerse & Cai 2004; McNamara & Graesser 2012). Our goal is to first investigate if the linguistic features provided by Coh-Metrix report salient co-occurrence patterns in our corpus of essays.¹ If so, we then aim to determine whether these patterns can be used to identify groupings of essays that are similar in their Multi-Dimensional profiles according to particular situational and functional parameters. If dimensions can be identified, we will then assess whether the features that inform these dimensions can reliably distinguish among the parameters that define the dimensions. Such an approach will allow us to identify how linguistic patterns in argumentative essays co-occur and what the relation amongst these co-occurrences is with regard to the parameters that underlie the essays.

2.1 Corpus

Like Biber (1988), our analysis depends on a large-scale text corpus. The corpus used in this analysis comprises 1529 argumentative essays collected under a variety of conditions discussed below. All essays were written by native speakers of English. Descriptive statistics for the corpus of essays are located in Table 1.

1. In addition to part of speech tags common in traditional MDA analyses, Coh-Metrix also reports on textual functions such as cohesion, contextual functions such as temporality and spatiality, and ideational functions such as propositional density.

Table 1. Descriptive statistics for essay corpus

Index	Mean	Standard deviation
Total number of words	334.016	186.655
Total number of sentences	18.835	10.102
Total number of paragraphs	3.358	1.668

2.2 Parameters

The argumentative essays were collected under a variety of different conditions. However, all essays were independent essays meaning that test-takers were prompted to produce an extended written argument built exclusively on their prior knowledge and experience. The identified situational and functional parameters are discussed below.

2.2.1 *Essay scoring*

A number of expert raters scored the essays that comprise the corpus. The majority of the raters had at least two years of experience teaching freshman composition courses at a large university. In all cases, at least two raters evaluated each essay based on a commonly used, standardized SAT rubric. In some cases, three raters evaluated each essay using the rubric. The rubric generated a holistic quality rating with a minimum score of 1 and a maximum score of 6. Raters were informed that the distance between each score was equal. In all cases, the raters were first trained to use the rubric with a subset of similar essays. Once the correlations within the raters reached a threshold of $r = .70$ ($p < .001$), the raters were considered trained. After training, raters scored each essay independently. In all cases, the final interrater reliability for the essays in the corpus was $r > .70$. We used the mean score between the raters as the final value for the quality of each essay unless the differences between the 2 raters was ≥ 2 , in which case a third expert rater adjudicated the score. Three scoring parameters were developed from the human ratings: interval scores from 1–6, low (1–2 scores), medium (3–4 scores), and high (5–6 scores) scores, and low (1–3 scores) and high (4–6 scores) scores.

2.2.2 *Prompt*

The essays were written on 15 different prompts. Each prompts was an SAT prompt. The shortened prompts and the number of essays written for each prompt are presented in Table 2. The full assignments and prompts are presented in Appendix 1.

Table 2. Essay count by prompt title

Prompt title	Prompt number	Number of essays
Choices	1	70
Competition	2	126
Dreaming	3	59
Equality	4	65
Fame	5	133
Fitting in	6	35
Heroes	7	158
Images and impressions	8	481
Memories	9	45
Optimism	10	56
Religion and television	11	60
Truth telling	12	51
Uniqueness	13	155
Winning	14	35

2.2.3 *Grade level*

The essays were collected from a variety of different grade levels including 9th grade, 10th grade, 11th grade, and college freshmen (see Table 3 for overview of this parameter). Two parameters were developed from these grade level conditions: grade level (9th grade, 10th grade, 11th grade, and college freshmen) and low (9th-10th grade levels) and high (11th-and college freshman) grade levels.

Table 3. Essay count by grade level

Grade level	Number of essays
Ninth grade	91
Tenth grade	550
Eleventh grade	266
College freshmen	622

2.2.4 *Geographic area*

The writers of the essays came from a number of different geographical areas including the District of Columbia, New York, Tennessee, Mississippi, and Louisiana. Descriptive statistics for each geographical area are presented in Table 4.

Table 4. Essay count by geographic area

Geographic area	Number of essays
District of Columbia	455
Louisiana	51
Tennessee	70
Mississippi	497
New York	456

2.2.5 *Production type*

Students wrote in two different production types: hand writing or typing (see Table 5 for descriptive statistics for this parameter).

Table 5. Essay count by production type

Production type	Number of essays
Hand written	526
Typed	1003

2.2.6 *Timing*

The essays were written under a variety of timing conditions. Some essays were untimed (i.e. take-home homework assignments) and some were timed. The timed essays included essays written in 10-, 15-, and 25-minute increments (see Table 6 for descriptive statistics for this parameter).

Table 6. Frequency count by timing

Timing	Number of essays
Untimed	184
10-minutes	51
15-minutes	355
25-minutes	939

2.3 Coh-Metrix indices

A number of different Coh-Metrix indices were used as potential grouping features. Most of these features also relate to functional parameters (i.e. textual functions such as cohesion, contextual functions such as temporality and spatiality, and

ideational functions such as propositional density). These indices are discussed briefly below. We refer the reader to Graesser et al. (2004) and McNamara & Graesser (2012) for a fuller description of the indices. In total, we selected 400 different linguistic indices from Coh-Metrix.

2.3.1 *Basic text properties*

Coh-Metrix provides a variety of indices that describe the basic properties and structure of a text, such as the number of words, the number of sentences, the number of paragraphs, the average length of words, and the average length of sentences.

2.3.2 *Basic lexical types*

Coh-Metrix reports the number of lexical types in the text (i.e. *total types*) and the number of content words contained in the text.

2.3.3 *Lexical overlap*

Lexical overlap refers to the extent to which words and phrases overlap across sentences and text, thus resulting in greater text cohesion and facilitating text comprehension (Kintsch & Van Dijk 1978). Coh-Metrix considers four forms of lexical overlap between sentences: noun overlap, argument overlap, stem overlap, and content word overlap.

2.3.4 *Semantic overlap*

Semantic overlap refers to the extent to which words, phrases, and sentences overlap semantically across text. Coh-Metrix measures semantic overlap using Latent Semantic Analysis (LSA), a mathematical and statistical technique for representing deeper world knowledge based on large corpora of texts (Landauer, McNamara, Dennis & Kintsch 2007).

2.3.5 *Givenness*

Given information is information that is recoverable from the preceding discourse (Halliday 1967) and, thus, does not require activation (Chafe 1975) and eases cognitive load. Coh-Metrix calculates text givenness using perpendicular and parallel LSA vectors (Hempelmann et al. 2005). This is referred to as LSA given/new.

2.3.6 *Causality*

Causal cohesion depends on causal relations between events and actions, which helps to create relationships between clauses (Pearson 1974–1975). Causal cohesion is measured in Coh-Metrix by calculating the ratio of causal verbs (e.g. *kill*,

break) to causal particles (e.g. *because, by, due to*). The causal verb count is based on the number of main causal verbs identified through WordNet (Fellbaum 1998).

2.3.7 *Connectives*

Connective phrases, such as *moreover* or *on the other hand*, make the relationships among clauses and sentences more explicit, and play an important role in the creation of cohesive links between ideas (Longo 1994). Coh-Metrix assesses the incidence of connectives on two dimensions. The first dimension contrasts positive vs. negative connectives, whereas the second dimension is associated with particular classes of cohesion identified by Halliday and Hasan (1976) and Louwse (2001). These connectives are associated with positive additive (*also, moreover*), negative additive (*however, but*), positive temporal (*after, before*), negative temporal (*until*), and causal (*because, so*) measures.

2.3.8 *Logical operators*

Logical operators make the logical flow and relations between ideas explicit, and include terms such as *or, and, not, and if-then*. Such terms have been shown to relate directly to the density and abstractness of a text (Costermans & Fayol 1997). Coh-Metrix assesses the incidence of these terms, combinations of these terms, and their common variants.

2.3.9 *Anaphoric reference*

Anaphoric reference refers to the presence of pronouns, which can be resolved by inferring the noun to which they refer from a previous sentence. Anaphoric reference is an important indicator of text cohesion (Halliday & Hasan 1976). Coh-Metrix measures anaphoric links between sentences by comparing pronouns in one sentence to pronouns and noun references in previous sentences.

2.3.10 *Spatial cohesion*

Spatial cohesion helps the reader construct a well-structured situational model (Zwaan, Langston & Graesser 1995) by developing a spatial representation of a text. Coh-Metrix measures spatiality by computing the number of motion verbs (*move, go, run*) and the number of location nouns (*place, region*) reported by WordNet (Fellbaum 1998).

2.3.11 *Temporal cohesion*

Temporal cohesion refers to the use of consistent temporal references, such as maintaining the same temporal tense (e.g. past, present, or future) throughout a section of text. Temporal cohesion is also an important element of situational knowledge. Temporal cohesion is measured in Coh-Metrix in three ways: aspect

repetition (e.g. progressive and perfect verb forms), tense repetition (e.g. present and past tense), and the combination of aspect and tense repetition.

2.3.12 Contextual cohesion indices

Coh-Metrix measures contextual cohesion by computing the degree of overlap between the prompt and the essay (using LSA and key word indices). Coh-Metrix also computes the number of key words and key types used for each essays using a reference corpus that is specific for individual prompts. Such a measure assesses how well the writer is producing words that are contextually relevant to the prompt.

2.3.13 Rhetorical indices

Coh-Metrix utilizes categories developed by Quirk, Greenbaum, Leech, and Svartvik (1985) to calculate the use of rhetorical strategies by measuring semantic categories that are related to, or proxies for, various rhetorical strategies. These include indirect pronouns (*all, none, some*), amplifiers and emphatics (*extremely, definitely*), downtoners (*slightly, somewhat, almost*), and exemplification (*for instance, namely*).

Coh-Metrix also calculates n-gram indices for words and phrases common in high-quality introduction, body, and conclusion paragraphs. These key n-grams are categorized based on rhetorical features. For instance, introductory paragraphs contain n-grams related to reported speech (i.e. *said*), contrast (i.e. *but some*), strength of argument (i.e. *we see*), and outside reference (i.e. *a person*). Body paragraphs contain n-grams related to providing examples (i.e. *addition to*), pronouns, conditionals (i.e. *if an*), and contrast (i.e. *while the*). Concluding paragraphs contain concluding statements (i.e. *in conclusion*), statements of fact (i.e. *it is*), negation, conditionals, modals, opinion (i.e. *I think*), and causality (i.e. *because*).

2.3.14 Syntactic complexity

Sentences that contain a higher number of words before the main verb, high-level constituents (sentences and embedded sentence constituents) per word in the sentence, and modifiers per noun phrase are more syntactically complex and, thus, more difficult to process and comprehend (Perfetti, Landi & Oakhill 2005). Coh-Metrix calculates the average number of these constructions across sentences in the text.

2.3.15 Syntactic similarity

Syntactic similarity refers to the uniformity and consistency of syntactic constructions in the text at the clause, phrase, and word level. More uniform syntactic constructions result in less complex syntax that is easier for the reader to process

(Crossley, Greenfield & McNamara 2008). Coh-Metrix calculates the mean level of syntax consistency at different levels of the text.

2.3.16 *Lexical categories*

Coh-Metrix uses the Charniak parser to calculate incidence scores for all of the part-of-speech tags reported by the Penn Tree Bank Tag Set (Marcus, Santorini & Marcinkiewicz 1993).

2.3.17 *Syntactic categories*

Similar to lexical categories for words, many clauses and phrases can also be assigned to particular syntactic categories. For example, phrasal components can include the incidence of noun, verb, and prepositional phrases. Clausal components can include declarative sentences and the number of embedded sentences (s-bars). Coh-Metrix uses the Charniak parser to calculate incidence scores for a variety of syntactic categories at the phrase and clause level.

2.3.18 *Readability formulas*

Coh-Metrix reports on a variety of readability formulas including Flesch Reading Ease (Flesch 1948), Flesch-Kincaid Grade Level (Kincaid, Fishburne, Rogers & Chissom 1975), and the Coh-Metrix Second Language Reading Index (Crossley, Greenfield & McNamara 2008).

2.3.19 *Easability scores*

Recent research on text readability has led to the development of text easability principal component scores that reflect the ease of processing a text. The eight components are narrativity, syntactic simplicity, word concreteness, referential cohesion, deep (situation model) cohesion, verb cohesion, connectivity, and temporality. These components are described in greater detail in Graesser, McNamara & Kulikowich (2011).

2.3.20 *Minimal edit distance*

Coh-Metrix computes the Minimal Edit Distance (MED) for a text sample by measuring differences in the sentential positioning of content words. A high MED value indicates that content words are located in different places within sentences across the text, suggesting lower structural cohesion.

2.3.21 *Word frequency*

Word frequency refers to how often particular words occur in the English language and is an important indicator of lexical knowledge. The presence of more uncommon words in a text suggests that the writer possesses a larger vocabulary. The

indices reported by Coh-Metrix are obtained from CELEX (Baayen, Piepenbrock & Gulikers 1995), a 17.9 million-word corpus. Coh-Metrix reports a mean frequency score across words.

2.3.22 *N-gram indices*

Coh-Metrix computes a variety of n-gram indices related to accuracy, frequency, and proportion. These are discussed below.

2.3.23 *N-gram accuracy*

Coh-Metrix assesses the n-gram accuracy of written texts by comparing the normalized frequency of n-grams shared in both a reference corpus taken from the British National Corpus (BNC) and the language sample of interest. The indices report correlations that represent the similarity between the frequency of occurrences in a representative corpus and a sample text. Higher rated essays contain n-grams that occur at similar frequencies as the representative corpus (Crossley, Cai & McNamara 2012).

2.3.24 *N-gram frequency*

Coh-Metrix assesses the frequency of n-grams found in a sample text. Higher proficiency writers use less frequent n-grams (Crossley et al. 2012).

2.3.25 *N-gram proportion*

Coh-Metrix reports n-gram values based on proportion scores. More proficient writers produce essays that contain proportionally fewer n-grams (Crossley et al. 2012).

2.3.26 *Lexical diversity*

Lexical diversity (LD) refers to the variety of words used in a text. LD indices generally measure the number of types (i.e. unique words occurring in the text) by tokens (i.e. all instances of words). Traditional indices of lexical diversity are highly correlated with text length, so Coh-Metrix also reports more sophisticated LD indices, including *MTLD* (McCarthy & Jarvis 2010) and *D* (Malvern, Richards, Chipere & Durán 2004). Lexical diversity measures relate to the number of words a writer knows.

2.3.27 *Word property indices*

Coh-Metrix reports word properties for concreteness, familiarity, imageability, and meaningfulness using human ratings provided by the Medical Research Council Psycholinguistic Database (MRC; Wilson 1988). Concrete words are more tangible than abstract words; familiar words are more recognizable and frequent.

Imageability indicates the ease at which a word can evoke a mental image and meaningfulness relates to the number of associations a word has with other words (Toglia & Battig 1978).

2.3.28 *Hypernymy*

Hypernymy describes the specificity or abstractness of a word. To assess hypernymy, Coh-Metrix uses the WordNet database (Fellbaum 1998; Miller, Beckwith, Fellbaum, Gross & Miller 1990), a computational, lexical database containing over 170,000 English nouns, verbs, adjectives, and adverbs. Each word in WordNet is located on a hierarchical scale allowing for the measurement of the number of subordinate words below and superordinate words above the target word. Coh-Metrix calculates a mean hypernymy rating across words in the text where a lower score reflects an overall use of less specific words and a higher score reflects an overall use of more specific words.

2.3.29 *Polysemy*

Polysemy refers to the number of senses or core meanings of a word, and is indicative of text ambiguity. For example, the word *bat* has at least two senses, one referring to an object used to play baseball and the other referring to a flying mammal. Texts that include more polysemous words are less precise because the words may be understood in different ways. Coh-Metrix measures word polysemy via WordNet and calculates an average polysemy value for content words in a text. A higher value indicates greater polysemy.

2.4 Statistical analysis

This study partially replicates the methods used in Biber's (1988) Multi-Dimensional analysis. Following this approach, we first entered the Coh-Metrix indices into a factor analysis (in this case, a principle component analysis; PCA) using a Promax rotation.² The PCA clustered the indices into groups that co-occurred frequently within the texts allowing for a large number of variables to be reduced into a smaller set of derived variables (i.e. the factors or dimensions). The dimensions reported by the PCA were interpreted based on writing parameters through a qualitative analysis of the linguistic indices that clustered on each dimension.

For inclusion into a factor, we set a conservative cut-off for the eigenvalues of $\lambda > .35$. This ensured that only salient linguistics indices would be included in the analysis. As with Biber (1988), we seek to find underlying functional interpreta-

2. Unlike Biber's (1988) analysis, we used a PCA because the underlying structures were undefined. In such cases, a PCA is appropriate because it reduces the variables to meaningful units.

tions that explain the co-occurrence factors among the indices that load onto each dimension. These functional interpretations represent a common function of the texts, which help explain the co-occurrence patterns. We hypothesize that the underlying functional interpretations will be related to the parameters discussed above (e.g. essay quality, grade level, prompt). To help confirm our hypotheses, we computed factor scores for each factor. These factor scores allow group comparisons for each factor based on the parameters of interest and can be used to visually demonstrate which parameters load high and low on the dimensions. For this interpretation, we computed factor scores for the indices on each dimension that loaded highest on that dimension only. Thus, if an index loaded higher in factor 1 than factor 2, the index was only included in the factor score for factor 1. Following Biber, the factor score was calculated by subtracting the mean of all scores for an index from the score for that index on a specific essay. This value was then divided by the standard deviation of the index across all essays. So, for instance, if *number of words* loaded into the first factor, we would calculate the mean (M) score and the standard deviation (SD) score for the *number of words* in all the essays. Let us assume that this was $M = 334.016$, $SD = 186.655$. If an essay had 230 words, it would receive a score of -0.702 (230 minus 334.016 divided by 186.665). For each essay, the average for all the indices in the factor was then calculated providing a factor score for the essay. The final factor scores for each essay were then averaged based on the essay conditions discussed above affording the opportunity to interpret the factors in consideration of situational writing parameters.

To complement the factor scores, we also conducted confirmatory statistical analyses by conducting Analyses of Variance (ANOVAs) followed by stepwise discriminant function analyses (DFAs). The ANOVAs examined if statistical differences were reported for the Coh-Metrix indices based on the dimension interpretation. The DFAs were used to provide evidence that the linguistic indices that load onto each dimension could be used to discriminate the essays based on the parameters selected. Such an approach is uncommon in most MDAs, but can provide important information about the fidelity of the factors derived from the factor analysis. Thus, if a dimension was interpreted as representing essay quality, we would conduct a follow-up DFA analysis that treated the essay score as the dependent variable and the Coh-Metrix indices from the factor loadings as the independent variables. We used the DFA to predict the degree of accuracy with which the independent variables could classify the dependent variables. Such an analysis allowed us to confirm that the Coh-Metrix variables that load into each factor could reliably distinguish the essay parameters. The stepwise DFA also allowed us to assess which variables were the strongest predictors of a dimension, providing us with evidence for the importance of specific indices in assessing essay parameters.

3. Results

3.1 Factor analysis

The eigenvalues for the first ten factors are reported in Table 7. In total, the first ten factors accounted for about 65% of the shared variance, with factor 1 explaining 21% of the variance alone. The scree plot corresponding to these factors is shown in Figure 1. The scree plot can be used to find a characteristic break that indicates at which point additional factors explain little additional variance in the analysis (Biber 1988). The clearest break in the scree plot appears to occur between the fourth and fifth factors, indicating that a four-factor solution is the best interpretation.

Table 7. First 10 eigenvalues from the principal component analysis

Factor number	Eigenvalue	Percent of variance	Cumulative variance
1	23.004	20.539	20.539
2	12.876	11.496	32.035
3	8.446	7.541	39.576
4	5.895	5.263	44.839
5	5.332	4.761	49.600
6	4.881	4.358	53.958
7	3.427	3.060	57.018
8	2.754	2.459	59.477
9	2.324	2.075	61.552
10	2.171	1.938	63.490

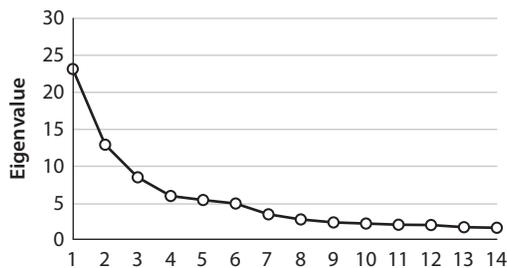


Figure 1. Scree plot

The final factor pattern for the four-factor solution included 95 of the original 400 indices selected from Coh-Metrix. The factor loadings for each of the linguistic features in each of the four factors are presented in Tables 8 through 11.

Table 8. Factor 1 loadings

Indices	Loadings
Total type count	.914
Total number of words	.878
Total number of prepositional phrases	.877
Total number of attributive adjectives	.830
Total number of sentences	.822
Total number of adverbs	.741
Total number of main 'Be' verbs	.678
Incidence of existential 'there'	.678
Incidence of 'that' verb complements	.678
Total number of nominalizations	.677
Total number of present tense verbs	.650
Total number of paragraphs	.648
Incidence of 'and'	.638
Total keyword type count	.637
Incidence of agentless passive verbs	.597
Incidence of predicative adjectives	.586
Incidence of time adverbials	.578
Incidence of stranded prepositions	.569
Incidence of gerunds	.565
Mean hypernymy value	.511
Total number of modifiers per noun phrase	.479
Lexical diversity (VOCD)	.478
Incidence of conjunctions	.475
Incidence of 'not'	.466
Lexical density	.462
Incidence of emphatics	.451
Total number of past tense verbs	.448
Total number of 'that' relative clauses	.448
Total number of private verbs	.442
LSA paragraph to paragraph	.439
Incidence of body paragraph n-grams	.434
Incidence of present participial clauses	.427
Incidence of nouns per thousand words	.411
Incidence of deleted past participle relatives	.410
Incidence of downtoners	.404

(Continued)

Table 8. Factor 1 loadings (Continued)

Indices	Loadings
Incidence of prepositional phrases	.399
Incidence of determiners	.378
Incidence of amplifiers	.375
Incidence of predictive modals	.372
Incidence of other adverbial subordinators	.368
Correlation of spoken bi-grams	.357
Incidence score of religion words	.356
Incidence of second person pronouns	-.375
Incidence of verbs per thousand words	-.401
Familiarity of content words	-.459
Incidence score of non-3rd person singular verbs	-.469
Indices of verb phrases	-.492
Incidence score of simple declarative sentences	-.546
Frequency of written bi-grams	-.581
Frequency of written tri-grams	-.591
Frequency of spoken tri-grams	-.615
Frequency of spoken bi-grams	-.753
Frequency of written tri-grams logarithm	-.798
Frequency of spoken tri-grams logarithm	-.799
Frequency of spoken bi-grams logarithm	-.810
Frequency of written bi-grams logarithm	-.811

Table 9. Factor 2 loadings

Indices	Loadings
Coh-Metrix L2 reading index	.675
Word frequency content words	.675
Proportion of spoken tri-grams	.670
Word frequency all words	.592
Narrativity component score	.574
Content word overlap	.572
Incidence of subordinating clauses	.516
Incidence of impersonal pronouns	.434
Incidence of conclusion paragraph n-grams	.408
Incidence of subordinating conditionals	.382

(Continued)

Table 9. (Continued)

Indices	Loadings
Average syllables per word	-.455
Average concreteness of content words	-.493
Average imageability of content words	-.495
Lexical diversity (MTLD)	-.563
Incidence of nouns	-.589
Type-token ratio of all content words	-.638

Table 10. Factor 3 loadings

Indices	Loadings
Stem overlap for all sentences	.803
Noun overlap for all sentences	.768
Average LSA cosine for all sentences	.768
Average LSA cosine for adjacent sentences	.735
Referential cohesion component score	.645
Argument overlap for all sentences	.637
Proportion of keywords	.581
LSA given/new	.520
Flesch reading ease score	-.483
Minimal edit distance all tags	-.524
Minimal edit distance all lemmas	-.545
Minimal edit distance all words	-.549

Table 11. Factor 4 loadings

Indices	Loadings
Syntactic simplicity component score	.662
Normed number of sentences (per text length)	.642
Syntactic similarity across paragraphs	.593
Syntactic similarity adjacent sentences	.572
Incidence of causal verbs	.530
Incidence of verbs	.444
Proportion of key types	.396
Incidence of words related to work	.369
Flesch-Kincaid grade level	-.538
Average words per sentence	-.655

3.2 Dimension analysis

Each factor was loaded onto a dimension using factor scores and then interpreted. We then conducted confirmatory DFA analyses to assess the strength of these interpretations. These four dimensions are discussed below.

3.2.1 Dimension 1 (*Essay Quality: Interval Scale*)

The first factor comprised 57 Coh-Metrix indices and explained 21% of the total variance. When the factor scores were computed for this dimension, the most appropriate label for this dimension was ‘Essay Quality: Interval Scale’ because the dimension separated high- and low-quality essays based on interval scales (i.e. from 1–6 scores). The dimension is presented in Figure 2.

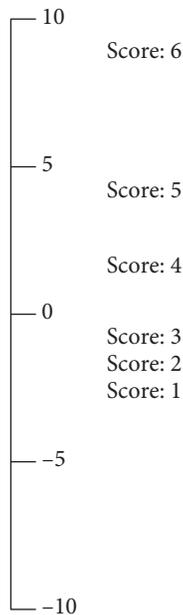


Figure 2. Dimension 1 (*Essay Quality: Interval Scale*)

The linguistic indices that separated essays based on scores can be grouped into functional categories that include fluency (number of words, type count), modifications (prepositions, adjectives, and adverbs), syntactic complexity (clause and phrase complexity), lexical sophistication (nominalizations, word hypernymy, n-grams, lexical diversity, word familiarity), structure (number of sentences and paragraphs), connectives (conjuncts and incidence of *and*), and rhetorical features (emphatics, downtoners, amplifiers).

An ANOVA and a confirmatory DFA were conducted on the factor scores reported for Dimension 1 (see Appendix 2). In this analysis, the Coh-Metrix indices were used as independent variables to classify the essays according to essay score (1 through 6). The ANOVA results (see Table 12) demonstrated that each index demonstrated significant differences based on essay score. The stepwise DFA retained 12 variables as significant predictors of proficiency level (*frequency spoken bigrams, nominalizations, number of paragraphs, attributive adjectives, incidence of verbs, word familiarity, predicative adjectives, time adverbials, proportion of key types, lexical diversity, LSA paragraph to paragraph scores, frequency written bigrams*) and removed the remaining 45 variables as non-significant predictors.

Table 12. Analysis of variance (ANOVA) results for factor 1

Indices	<i>f</i> value	<i>p</i> value	h_p^2
Frequency of spoken bi-grams logarithm	134.614	<.001	.306
Total type count	129.816	<.001	.299
Frequency of written bi-grams logarithm	126.692	<.001	.294
Total number of paragraphs	122.107	<.001	.286
Frequency of spoken tri-grams logarithm	118.725	<.001	.280
Total number of attributive adjectives	117.129	<.001	.278
Total number of prepositional phrases	114.867	<.001	.274
Total number of nominalizations	111.385	<.001	.268
Frequency of written tri-grams logarithm	105.825	<.001	.258
Total number of words	104.523	<.001	.255
Frequency of spoken bi-grams	104.461	<.001	.255
Total number of sentences	99.216	<.001	.246
Total keyword type count	85.774	<.001	.220
LSA paragraph to paragraph	69.874	<.001	.187
Total number of adverbs	63.634	<.001	.173
Incidence of 'and'	56.898	<.001	.157
Incidence of agentless passive verbs	55.921	<.001	.155
Frequency of spoken tri-grams	54.993	<.001	.153
Total number of main 'Be' verbs	55.116	<.001	.153
Incidence of existential 'there'	55.116	<.001	.153
Incidence of 'that' verb complements	55.116	<.001	.153
Frequency of written bi-grams	54.042	<.001	.151

(Continued)

Table 12. Factor 1 loadings (Continued)

Indices	<i>f</i> value	<i>p</i> value	h_p^2
Frequency of written tri-grams	50.558	<.001	.142
Incidence score of simple declarative sentences	49.611	<.001	.140
Incidence of predicative adjectives	44.034	<.001	.126
Mean hypernymy value	42.866	<.001	.123
Lexical diversity (VOCD)	39.088	<.001	.114
Familiarity of content words	39.110	<.001	.114
Mean hypernymy value	39.021	<.001	.114
Incidence of verbs per thousand words	37.988	<.001	.111
Incidence of conjunctions	37.182	<.001	.109
Total number of present tense verbs	35.447	<.001	.104
Total number of modifiers per noun phrase	33.829	<.001	.100
Incidence of stranded prepositions	32.705	<.001	.097
Lexical density	31.684	<.001	.094
Incidence of gerunds	29.263	<.001	.088
Incidence of body paragraph n-grams	27.867	<.001	.084
Incidence of deleted past participle relatives	27.814	<.001	.084
Incidence score of non-3rd person singular verbs	24.162	<.001	.073
Incidence of prepositional phrases	23.712	<.001	.072
Incidence of second person pronouns	23.373	<.001	.071
Incidence of 'not'	22.611	<.001	.069
Incidence score of downtoners	21.833	<.001	.067
Total number of 'that' relative clauses	21.669	<.001	.066
Total number of past tense verbs	20.688	<.001	.064
Incidence of other adverbial subordinators	20.737	<.001	.064
Incidence of present participial clauses	19.689	<.001	.061
Incidence of time adverbials	18.134	<.001	.056
Correlation of spoken bi-grams	17.397	<.001	.054
Incidence of amplifiers	17.340	<.001	.054
Incidence of emphatics	17.379	<.001	.054
Incidence of nouns per thousand words	16.497	<.001	.051
Incidence of determiners	13.239	<.001	.042
Total number of private verbs	10.621	<.001	.034
Incidence of predictive modals	10.107	<.001	.032
Incidence score of religion words	5.290	<.001	.017

The results demonstrate that the DFA using the 12 significant Coh-Metrix indices correctly allocated 702 of the 1529 essay based on essay score in the total set, χ^2 ($df = 25$, $n = 1529$) = 1089.639, $p < .001$, for an accuracy of 45.9% (the chance level for this analysis is 17%). We also conducted a leave-one-out cross-validation (LOOCV) analysis to assess the generalizability of the model reported by the DFA. The LOOCV discriminant analysis correctly allocated 672 of the 1529 essays based on essay score for an accuracy of 44.0% (see the confusion matrix reported in Table 13 for results). In the confusion matrix, all correct classifications begin at the top-left and run diagonal to the bottom-right. Thus, for the LOOCV set, 34 essays assigned a 1 by the human raters were correctly classified as a 1. Twenty-five were classified as a 2, six were classified as a 3, and one was classified as a 4. The measure of agreement between the actual text type and the text type assigned by the model produced a weighted Cohen's Kappa of 0.467, demonstrating a moderate agreement.

Table 13. Classification results for dimension 1 (Essay quality: Interval scale)

Predicted group membership			1	2	3	4	5	6
Original	Count	1	35	24	6	1	0	0
		2	59	151	69	24	10	2
		3	17	128	288	138	49	7
		4	0	15	112	157	76	13
		5	0	4	10	36	65	25
		6	0	0	0	0	2	6
Cross-validation	Count	1	34	25	6	1	0	0
		2	63	146	70	24	10	2
		3	18	129	282	142	49	7
		4	0	16	113	149	81	14
		5	0	4	10	43	57	26
		6	0	0	0	0	4	4

3.2.2 Dimension 2 (Essay Quality: Low, Middle, High)

The second factor comprised 16 Coh-Metrix indices and explained 12% of the total variance. When the factor scores were computed for this dimension, the most appropriate label for this dimension was 'Essay Quality (Low, Middle, and High)' because the dimension separated essays based on those scored low (1–2), middle (3–4) and high (5–6). The dimension is presented in Figure 3.

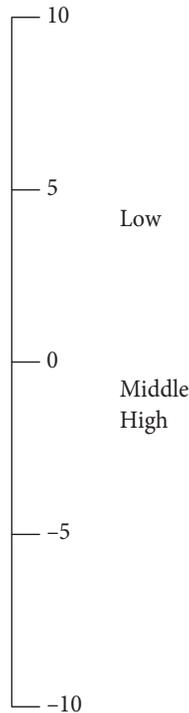


Figure 3. Dimension 2 (Essay Quality: Low, Middle, High)

The linguistic indices that separated high- from low-quality essays can be grouped into functional features that include lexical sophistication (lexical diversity, word frequency, word concreteness), narrativity (narrative component scores, impersonal pronouns), and syntactic complexity (subordinating conditionals and subordinating clauses).

An ANOVA and a confirmatory DFA were conducted on the factor scores reported for Dimension 2 (see Appendix 2). In this analysis, the Coh-Metrix indices were used as independent variables to classify the essays according to score grouping (low, middle, and high). The ANOVA results (see Table 14) demonstrated that all indices except *subordinating conditionals* demonstrated significant differences based on score grouping. The stepwise DFA retained nine variables as significant predictors of proficiency level (*average syllables per word*, *content word overlap*, *type-token ratio*, *lexical diversity D*, *word frequency content words*, *word frequency all words*, *word imageability*, *conclusion n-grams*, and *impersonal pronouns*) and removed the remaining seven variables as non-significant predictors.

The results of the DFA demonstrate that nine significant Coh-Metrix indices correctly allocated 863 of the 1529 essay based on essay score in the total set, χ^2

Table 14. Analysis of variance (ANOVA) results for factor 2

Indices	<i>f</i> value	<i>p</i> value	h_p^2
Narrativity component score	136.658	<.001	.152
Word frequency content words	113.027	<.001	.129
Average syllables per word	105.974	<.001	.122
Coh-Metrix L2 reading index	87.919	<.001	.103
Word frequency all words	87.573	<.001	.103
Incidence of nouns	85.862	<.001	.101
Content word overlap	60.626	<.001	.074
Type-token ratio of all content words	58.580	<.001	.071
Incidence of subordinating clauses	56.15	<.001	.069
Average imageability of content words	54.819	<.001	.067
Proportion of spoken tri-grams	42.834	<.001	.053
Average concreteness of content words	37.271	<.001	.047
Lexical diversity (MTLD)	30.172	<.001	.038
Incidence of conclusion paragraph <i>n</i> -grams	25.283	<.001	.032
Incidence of impersonal pronouns	22.430	<.001	.029
Incidence of subordinating conditionals	.876	<.050	.001

Table 15. Classification results for dimension 2 (Essay quality: Low, Middle, High)

Predicted group membership			Low	Middle	High
Original	Count	Low	260	86	35
		Middle	194	388	251
		High	13	87	215
Cross-validation	Count	Low	257	89	35
		Middle	197	384	252
		High	13	93	209

($df = 4, n = 1529$) = 501.046, $p < .001$, for an accuracy of 56.4% (the chance level for this analysis is 33%). The LOOCV discriminant analysis correctly allocated 850 of the 1529 essays based on essay score for an accuracy of 55.6% (see the confusion matrix reported in Table 15 for results). The measure of agreement between the actual text type and that assigned by the model produced a weighted Cohen's Kappa of 0.417, demonstrating a moderate agreement.

3.2.3 Dimension 3 (Prompt and Assignment)

The third factor comprised 12 Coh-Metrix indices and explained 8% of the total variance. When the factor scores were computed for this dimension, the most appropriate label for this dimension was ‘Prompt and Assignment’ because the dimension separated essays based on the linguistic features found in the prompt and assignment. The dimension is presented in Figure 4.

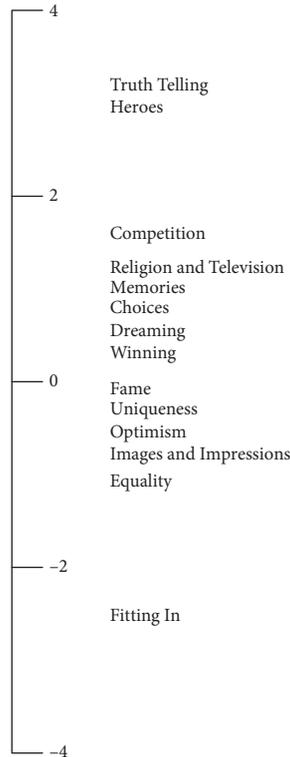


Figure 4. Dimension 3 (Prompt and Assignment)

The linguistic indices that separated the different prompts and assignments can be grouped into functional parameters that include lexical overlap (stem and noun overlap), semantic similarity (LSA indices), and structural cohesion (minimal edit distance indices). We interpreted these indices as related to prompts and assignments because the prompts and assignments that loaded positively on this dimension shared similar patterns in their linguistic features as the essays that were written in response to them. In addition, the essays written on these prompts came from a variety of grade levels and regions, indicating that demographics were not in play. Thus, prompts that loaded high on this dimension had higher

lexical overlap and semantic similarity scores and lower structural cohesion scores than the prompts that loaded low (see Table 16 for details).

Table 16. Prompt and assignment comparison to essays

Indices	Low loading prompts	High loading prompts	Same pattern as essays
Overlap	0.370	0.549	+
LSA	0.116	0.300	+
MED	0.859	0.810	+

An ANOVA and a confirmatory DFA were conducted on the factor scores reported for Dimension 3 (see Appendix 2). In this analysis, the Coh-Metrix indices were used as independent variables to classify the essays according to prompt (14 prompts). The ANOVA results (see Table 17) demonstrated that all indices demonstrated significant differences based on prompt and assignment differences. The stepwise DFA retained ten variables as significant predictors of proficiency level (removing *Flesch Reading Ease* and *Minimal Edit Distance All Words*).

Table 17. Analysis of variance (ANOVA) results for Factor 3

Indices	<i>f</i> value	<i>p</i> value	h_p^2
Proportion of keywords	46.212	<.001	.299
LSA given/new	39.957	<.001	.270
Flesch reading ease score	20.831	<.001	.162
Referential cohesion component score	14.102	<.001	.115
Average LSA cosine for all sentences	12.808	<.001	.106
Stem overlap for all sentences	12.176	<.001	.101
Noun overlap for all sentences	12.029	<.001	.100
Argument overlap for all sentences	11.930	<.001	.099
Average LSA cosine for adjacent sentences	11.706	<.001	.098
Minimal edit distance all lemmas	11.053	<.001	.093
Minimal edit distance all words	8.664	<.001	.074
Minimal edit distance all tags	5.496	<.001	.048

The results demonstrate that the DFA using the ten significant Coh-Metrix indices correctly allocated 624 of the 1529 essay based on prompt in the total set, χ^2 ($df = 169$, $n = 1529$) = 2582.0044, $p < .001$, for an accuracy of 40.8% (the chance level for this analysis is 7%). The LOOCV discriminant analysis correctly

allocated 587 of the 1529 essays based on prompt for an accuracy of 38.4% (see the confusion matrix reported in Table 18 for results). The measure of agreement between the actual text type and that assigned by the model produced a weighted Cohen's Kappa of 0.338, demonstrating a fair agreement.

Table 18. Classification results for dimension 3 (Prompt and Assignment)

Predicted group membership		1	2	3	4	5	6	7	8	9	10	11	12	13	14	
Original	Count	1	3	3	2	2	3	13	7	9	8	5	4	4	1	6
		2	5	48	11	3	7	3	5	7	5	2	10	5	7	8
		3	1	0	25	8	0	4	5	0	0	0	8	0	8	0
		4	0	1	10	34	0	5	0	2	0	3	3	0	3	3
		5	2	17	1	0	60	9	9	12	3	5	4	5	2	4
		6	2	0	0	5	0	16	0	1	3	0	2	0	3	3
		7	4	9	8	3	2	10	71	13	7	1	13	1	9	7
		8	14	33	0	18	12	21	13	217	6	38	33	28	31	17
		9	2	1	1	1	4	5	4	2	15	0	0	8	1	1
		10	2	0	0	9	0	5	0	9	0	19	1	0	6	5
		11	0	1	1	9	0	0	3	0	0	0	44	0	1	1
		12	2	2	0	1	7	2	0	2	10	0	1	23	1	0
		13	2	11	14	15	5	13	9	13	5	11	6	2	38	11
		14	1	2	0	4	1	3	2	4	1	3	0	2	1	11
Cross-validation	Count	1	3	3	2	2	3	13	7	9	8	5	4	4	1	6
		2	5	46	11	3	9	3	5	7	5	2	19	5	7	8
		3	1	0	25	8	0	4	5	0	0	0	8	0	8	0
		4	0	1	11	29	0	7	0	3	0	3	4	0	4	3
		5	2	18	1	0	58	9	9	12	4	5	4	5	2	4
		6	6	0	0	5	0	9	0	1	5	0	3	0	3	3
		7	4	9	11	3	2	11	66	13	7	1	14	1	9	7
		8	14	33	0	18	12	23	14	214	6	38	33	28	31	17
		9	2	1	1	1	4	5	4	2	14	0	0	9	1	1
		10	1	0	0	10	0	6	0	9	0	16	1	0	6	7
		11	0	1	1	10	0	0	3	1	0	0	42	0	1	1
		12	2	2	0	1	7	2	0	2	12	0	1	21	1	0
		13	2	11	15	15	5	13	10	13	5	11	6	2	36	11
		14	1	2	0	4	2	5	2	4	1	3	0	2	1	8

3.2.4 Dimension 4 (Low vs. High Grade Level)

The fourth factor comprised ten Coh-Metrix indices and explained 5% of the total variance. When the factor scores were computed for this dimension, the most appropriate label for this dimension was 'Low vs. High Grade Level' because the dimension separated essays based on whether they were written by students in 9th and 10th grade or students in 11th grade or the first year of college. The dimension is presented in Figure 5.

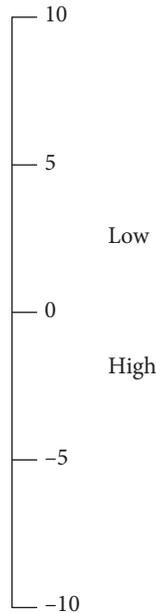


Figure 5. Dimension 4 (Grade Level)

The linguistic indices that separated grade levels can be grouped into functional parameters that include syntactic complexity (syntactic component scores, syntactic similarity) and verb density (incidence of verbs and verb causality). An ANOVA and a confirmatory DFA were conducted on the factor scores reported for Dimension 4 (see Appendix 2). In this analysis, the Coh-Metrix indices were used as independent variables to classify the essays according to grade level (two levels). The ANOVA results (see Table 19) demonstrated that all indices demonstrated significant differences between the grade level groupings. The stepwise DFA retained six variables as significant predictors of proficiency level (*Words Related to Work*, *Flesch-Kincaid Grade level*, *Syntactic Simplicity Component*, *Incidence of Causal Verbs*, *Syntactic Similarity across Paragraphs*, and *Syntactic Similarity Adjacent Sentences*).

Table 19. Analysis of variance (ANOVA) results for factor 4

Indices	<i>f</i> value	<i>p</i> value	h_p^2
Incidence of words related to work	238.031	<.001	.135
Proportion of key types	32.443	<.001	.021
Structural similarity across paragraphs	31.010	<.001	.020
Incidence of causal verbs	27.065	<.001	.017
Normed number of sentence (by text length)	25.422	<.001	.016
Average words per sentence	23.980	<.001	.015
Flesch-Kincaid grade level	15.447	<.001	.010
Structural similarity adjacent sentences	13.461	<.001	.009
Syntactic simplicity component score	5.854	<.050	.004
Incidence of verbs	.980	>.050	.001

The results demonstrate that the DFA using the six Coh-Metrix indices correctly allocated 1065 of the 1529 essay based on essay score in the total set, χ^2 ($df = 1, n = 1529$) = 214.695, $p < .001$, for an accuracy of 69.7% (the chance level for this analysis is 50%). The LOOCV discriminant analysis correctly allocated 1059 of the 1529 essays based on essay score for an accuracy of 69.3% (see the confusion matrix reported in Table 20 for results). The measure of agreement between the actual text type and that assigned by the model produced a weighted Cohen's Kappa of 0.375, demonstrating a fair agreement.

Table 20. Classification Results for Dimension 4 (Low vs. High Grade Level)

Predicted group membership			Low	High
Original	Count	Low	401	240
		High	224	664
Cross-validation	Count	Low	399	242
		High	228	660

4. Discussion

This study has shown that linguistic features related to lexical sophistication, syntactic complexity, cohesion, rhetorical strategies, contextual relevance, and text structure reliably co-occur in specific situations. Three main situational parameters were interpreted in this analysis to be related to essay quality, prompt, and grade level. Each of these situational parameters was realized using a number of

different functional parameters comprising the co-occurring linguistic features. Each of the situational parameters and their corresponding functional parameters are discussed below.

4.1 Essay quality

The first two factors reported by the principal component analysis explained 32% of the variance and were associated with essay quality. The first factor reliably discriminated between essay scores (i.e. 1 through 6 scores assigned by human raters). This factor provided evidence that high-quality essays are longer, more descriptive (i.e. more adjectives and adverbs), more lexically sophisticated and contain more clausal subordination, clausal embeddings, nominalizations, gerunds, longer phrases, connectives, and rhetorical elements. In addition, higher-quality essays contain fewer frequent n-grams, verbs and verb phrases, and second person pronouns. Many of these linguistic indices have long been associated with the quality of essays. For instance text length is a common indicator of essay quality (Crossley, Roscoe & McNamara 2011) as are syntactic complexity and lexical sophistication (McNamara et al. 2010; Crossley, Weston, Sullivan & McNamara 2011; Crossley et al. 2012). Phrasal and nominal sophistication have long been associated with academic writing (Biber, Gray & Poonpon 2011; Fang, Schleppegrell & Cox 2006; Halliday 1989; Halliday & Matthiessen 1999; Wells 1960), as have rhetorical features such as amplifiers and downtoners. However, unlike some previous studies (McNamara et al. 2010; Crossley, Roscoe & McNamara 2011), cohesion features were also positively indicative of essay quality, demonstrating that higher-scored essays contained more instances of conjuncts and coordinating conjunctions. Additionally, linguistic elements related to text description (e.g. adjectives, adverbs, and prepositional phrases) were also indicative of essay quality, denoting that higher-scored essays were also more descriptive and provided more verbal and nominal modifications.

Our second factor indicates that there is likely more than one profile of a high-quality essay. This factor did not strongly separate essays based on individual score, but rather groupings of scores into low, middle, and high-scored essays. This factor, like the first, included lexical and syntactic indices, but some of the patterns of occurrence were different. For instance, in the second factor, essays grouped together as high quality were less lexically sophisticated (for instance, containing more concrete and imageable nouns), less syntactically complex (for instance, containing less clausal subordination), and contained less cohesion (for instance, containing less word overlap). However, in a similar fashion to factor 1, essays grouped as high were also more lexically sophisticated, containing greater lexical diversity and more infrequent words. Essay quality in factor 2 was also evidenced by less narrativity and text readability.

Our first two factors thus demonstrate that there are multiple profiles for high-quality essays. In both cases, some level of lexical sophistication equates to essay quality, although the second factor included words that were more concrete and imageable. In addition, differences between the two factors are evident in the use of syntactic complexity and cohesion features. Other differences can be observed in phrasal components, rhetorical features, text descriptors (in the case of the first profile) and narrativity and readability (in the case of the second profile). Interpretation of these two profiles indicates that quality essays in the first factor are characterized as lexically complex, more cohesive, more descriptive, and containing more rhetorical features, while the quality essays in the second factor are characterized as lexically complex, lexically specific and written in a non-narrative style.

4.2 Prompt effects

Our third factor explained 8% of the variance in our analysis and indicates that the cohesion features found in the prompt strongly influence the cohesion features produced by the writer. Thus, if a prompt contains greater noun overlap, greater semantic similarity, and greater structural similarity, the essays written about the prompt will contain similar levels of these cohesion features (see Table 16). Therefore, it appears that a prompt-based priming effect may exist for some cohesion features. Notably, many of these cohesion features are similar to those that are predictive of essay quality (for instance, semantic similarity is positively correlated with essay score in factor 1, while content word overlap is negatively correlated with essay score in factor 2). Such a finding indicates that the prompt itself may lead writers to produce cohesion features that may lead to higher or lower judgments of essay quality. Additionally, this finding suggests that teachers and test administrators need to use caution when selecting prompts and attempt to control the level of cohesion found in prompts in order to ensure that writing samples are based on writer proficiency and not a result of prompt-based priming effects.

4.3 Grade level

Our final factor explained 6% of the variance in our analysis and indicates that linguistic features in essays reliably co-vary as a result of grade level. Specifically, as grade level increased, writers began to produce more complex syntactic structures, more syntactic variation, longer sentences, and fewer verbs. This finding supports earlier research by Crossley, Weston, Sullivan and McNamara (2011) in which writers produced more complex syntactic structures as a function of increasing grade level. A movement toward greater syntactic complexity may represent

attempts by writers to embed cohesive elements of a text in syntactic structures as compared with the use of explicit cohesive devices. Additionally, the use of fewer verbs may indicate that developing writers are moving more toward the production of noun phrases and a nominal style of writing. Such a style is suggestive of increased academic writing (Biber et al. 2011; Fang et al. 2006; Halliday 1989; Halliday & Matthiessen 1999; Wells 1960).

5. Conclusion

This study has demonstrated the utility of a Multi-Dimensional analysis approach for assessing co-occurrence patterns in argumentative essays. The study is unique in the writing literature because it allows for the simultaneous examination of a variety of different situational parameters. These situational parameters (quality, prompt, and grade level) are all important research areas in composition studies, and a Multi-Dimensional analysis provides concomitant information about how essays vary among and between themselves.

The study is also unique as a Multi-Dimensional analysis because it adds an additional layer of confirmatory statistical analyses: linear discriminate analyses. Unlike past Multi-Dimensional analyses, this study provides statistical confirmation that the variables that co-occur in the reported factors can, in fact, be used to discriminate between the situational parameters as interpreted on the dimensions. The use of linear discriminate analyses provide a means to assess the strength of factor interpretations and the strength of co-occurring variables to classify groupings as reported on the dimensions. Researchers should consider adding DFA analyses to future MDA studies.

In conclusion, this study not only provides strong indicators of how linguistic features co-occur in varying writing situations, but also methodological opportunities to investigate the strength of the features to interpret these writing situations. The former has important implications for writing theory, writing assessment, and writing pedagogy. The latter has important implications for further developing Multi-Dimensional analysis approaches.

Acknowledgments

This research was supported in part by the Institute for Education Sciences (IES R305A080589 and IES R305G20018-02). Ideas expressed in this material are those of the authors and do not necessarily reflect the views of the IES.

References

- Abbott, R. D., Berninger, V. W., & Fayol, M. (2010). Longitudinal relationships of levels of language in writing and between writing and reading in grades 1 to 7. *Journal of Educational Psychology, 102*, 281–298. DOI: 10.1037/a0019318
- Baayen, H., Piepenbrock, R., & Van Rijn, H. (Eds.). (1995). *The CELEX lexical database* (CD-ROM). Philadelphia, PA: Linguistic Data Consortium.
- Baker, W., & Egginton, W. (1999). Bilingual creativity, multi-dimensional analysis, and world Englishes. *World Englishes, 18*(3), 343–358. DOI: 10.1111/1467-971X.00148
- Berninger, V. W., Abbott, R. D., Swanson, H. L., Lovitt, D., Trivedi, P., Lin, S.-J., & Amtmann, D. (2010). Relationship of word and sentence-level working memory to reading and writing in second, fourth, and sixth grade. *Language, Speech, and Hearing Services in Schools, 41*, 179–193. DOI: 10.1044/0161-1461(2009/08-0002)
- Berninger, V. W., Cartwright, A. C., Yates, C. M., Swanson, L., & Abbott, R. D. (1994). Developmental skills related to writing and reading acquisition in the intermediate grades: Shared and unique functional systems. *Reading and Writing, 6*, 161–196. DOI: 10.1007/BF01026911
- Berninger, V. W., Mizokawa, D., & Bragg, R. (1991). Theory-based diagnosis and remediation of writing disabilities. *Journal of School Psychology, 29*, 57–79. DOI: 10.1016/0022-4405(91)90016-K
- Biber, D. (1985). Investigating macroscopic textual variation through multi-feature/multi-dimensional analyses. *Linguistics, 23*, 337–360. DOI: 10.1515/ling.1985.23.2.337
- Biber, D. (1986). Spoken and written textual dimensions in English: Resolving the contradictory findings. *Language, 62*, 384–414. DOI: 10.2307/414678
- Biber, D. (1988). *Variation across speech and writing*. Cambridge: CUP. DOI: 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. DOI: 10.5054/tq.2011.244483
- Chafe, W. (1975). Givenness, contrastiveness, definiteness, subjects, topics, and point of view in subject and topic. In C.N. Li (Ed.), *Subject and topic* (pp. 25–55). New York, NY: Academic Press.
- Connor, U. (1987). Research frontiers in writing analysis. *TESOL Quarterly, 21*, 677–696. DOI: 10.2307/3586989
- Connor, U., & Upton, T. (2004). The genre of grant proposals: A corpus linguistic analysis. In U. Connor, & T. Upton (Eds.), *Discourse in the professions: Perspectives from corpus linguistics* (pp. 235–256). Amsterdam: John Benjamins. DOI: 10.1075/scl.16
- Costermans, J., & Fayol, M. (1997). *Processing interclausal relationships: Studies in production and comprehension of text*. Hillsdale, NJ: Lawrence Erlbaum Associates.
- Crossley, S., Greenfield, J., & McNamara, D. (2008). Assessing text readability using psycholinguistic indices. *TESOL Quarterly, 42*, 475–493.
- Crossley, S., Roscoe, R., & McNamara, D. (2011). Predicting human scores of essay quality using computational indices of linguistic and textual features. In S. Bull, J. Kay, & A. Mitrovic (Eds.), *Proceedings of the 15th International Conference on Artificial Intelligence in Education* (pp. 438–440). Gautam Biswas: AIED.
- Crossley, S., Weston, J., Sullivan, S., & McNamara, D. (2011). The development of writing proficiency as a function of grade level: A linguistic analysis. *Written Communication, 28*, 282–311. DOI: 10.1177/0741088311410188

- Crossley, S., Cai, Z., & McNamara, D. (2012). Syntagmatic, paradigmatic, and automatic n-gram approaches to assessing essay quality. In P. McCarthy & M. Youngblood (Eds.), *Proceedings of the 25th International Florida Artificial Intelligence Research Society (FLAIRS) Conference* (pp. 214–219). Menlo Park, CA: The AAAI Press.
- Crossley, S., & McNamara, D. (2012). Predicting second language writing proficiency: The roles of cohesion and linguistic sophistication. *Journal of Research in Reading*, 35, 115–136. DOI: 10.1111/j.1467-9817.2010.01449.x
- Crowhurst, M. (1990). Reading/writing relationships: An intervention study. *Canadian Journal of Education*, 15, 155–172. DOI: 10.2307/1495373
- Fang, Z., Schleppegrell, M., & Cox, B. (2006). Understanding the language demands of schooling: Nouns in academic registers. *Journal of Literacy Research*, 38, 247–273. DOI: 10.1207/s15548430jlr3803_1
- Fellbaum, C. (1998). *WordNet: An electronic lexical database*. Cambridge, MA: The MIT Press.
- Flesch, R. (1948). A new readability yardstick. *Journal of Applied Psychology*, 32, 221–233. DOI: 10.1037/h0057532
- Freedman, A., & Pringle, I. (1980). Writing in the college years: Some indices of growth. *College Composition and Communication*, 31, 311–322. DOI: 10.2307/356491
- Graesser, A., McNamara, D., & Kulikowich, J. (2011). Coh-Metrix: Providing multilevel analyses of text characteristics. *Educational Researcher*, 40, 223–234. DOI: 10.3102/0013189X11413260
- Graesser, A., McNamara, D., Louwerse, M., & Cai, Z. (2004). Coh-Metrix: Analysis of text on cohesion and language. *Behavior Research Methods, Instruments, & Computers*, 36, 193–202. DOI: 10.3758/BF03195564
- Gregg, N., Coleman, C., Stennett, R., & Davis, M. (2002). Discourse complexity of college writers with and without disabilities: A Multi-Dimensional analysis. *Journal of Learning Disabilities*, 35(1), 23–38. DOI: 10.1177/002221940203500103
- Halliday, M. A. K. (1967). Notes on transitivity and theme in English. *Journal of Linguistics*, 3, 199–244. DOI: 10.1017/S0022226700016613
- Halliday, M. A. K., & Hasan, R. (1976). *Cohesion in English*. London: Longman.
- Halliday, M. A. K. (1989). *Spoken and written language*. Oxford: OUP.
- Halliday, M. A. K., & Matthiessen, C. (1999). *Construing experience through meaning: A language-based approach to cognition*. London: Cassell.
- Haswell, R. (1986). *Change in undergraduate and post-graduate writing performance: Quantified findings* (Technical report; ERIC database ED269780).
- Haswell, R. (2000). Documenting improvement in college writing: A longitudinal approach. *Written Communication*, 17, 307–352. DOI: 10.1177/0741088300017003001
- Hempelmann, C., Dufty, D., McCarthy, P., Graesser, A., Cai, Z., & McNamara, D. (2005). Using LSA to automatically identify givenness and newness of noun phrases in written discourse. In B. Bara, L. Barsalou, & M. Bucciarelli (Eds.), *Proceedings of the 27th Annual Conference of the Cognitive Science Society* (pp. 941–946). Mahwah, NJ: Lawrence Erlbaum Associates.
- Hunt, K. (1965). *Grammatical structures written at three grade levels* (NCTE Research report No. 3). Urbana, IL: National Council of Teachers of English.
- Jenkins, J., Johnson, E., & Hileman, J. (2004). When is reading also writing: Sources of individual differences on the new reading performance assessments. *Scientific Studies in Reading*, 8, 125–151. DOI: 10.1207/s1532799xssr0802_2
- Kellogg, R., & Raulerson, B. (2007). Improving the writing skills of college students. *Psychonomic Bulletin and Review*, 14, 237–242. DOI: 10.3758/BF03194058

- Kincaid, J. P., Fishburne, R. P., Rogers, R., & Chissom, B. (1975). *Derivation of new readability formulas* (Automated readability index, Fog count and Flesch reading ease formula) for Navy enlisted personnel (Research branch report 8-75). Millington, TN: Naval Technical Training, U. S. Naval Air Station.
- Kintsch, W., & van Dijk, T. (1978). Toward a model of text comprehension and production. *Psychological Review*, 85, 363–394. DOI: 10.1037/0033-295X.85.5.363
- Landauer, T., McNamara, D., Dennis, S., & Kintsch, W. (2007). *Handbook of latent semantic analysis*. Mahwah, NJ: Lawrence Erlbaum Associates.
- Longo, B. (1994). Current research in technical communication: The role of metadiscourse in persuasion. *Technical Communication*, 41, 348–352.
- Louwerse, M. (2001). An analytic and cognitive parameterization of coherence relations. *Cognitive Linguistics*, 12, 291–315.
- Malvern, D., Richards, B., Chipere, N., & Durán, P. (2004). *Lexical diversity and language development: Quantification and assessment*. Houndmills: Palgrave Macmillan. DOI: 10.1057/9780230511804
- Marcus, M., Santorini, B., & Marcinkiewicz, M. A. (1993). Building a large annotated corpus of English: The Penn treebank. *Computational Linguistics*, 19, 313–330.
- McCarthy, P., & Jarvis, S. (2010). MTL, D, and HD-D: A validation study of sophisticated approaches to lexical diversity assessment. *Behavior Research Methods*, 42, 381–392. DOI: 10.3758/BRM.42.2.381
- McCutchen, D. (1986). Domain knowledge and linguistic knowledge in the development of writing ability. *Journal of Memory and Language*, 25, 431–444. DOI: 10.1016/0749-596X(86)90036-7
- McCutchen, D., & Perfetti, C. (1982). Coherence and connectedness in the development of discourse production. *Text*, 2, 113–139. DOI: 10.1515/text.1.1982.2.1-3.113
- McNamara, D., Crossley, S., & McCarthy, P. (2010). Linguistic features of writing quality. *Written Communication*, 27, 57–86. DOI: 10.1177/0741088309351547
- McNamara, D., & Graesser, A. (2012). Coh-Metrix: An automated tool for theoretical and applied natural language processing. In P. McCarthy, & C. Boonthum-Denecke (Eds.), *Applied natural language processing and content analysis: Identification, investigation, and resolution* (pp. 188–205). Hershey, PA: IGI Global.
- Miller, G., Beckwith, R., Fellbaum, C., Gross, D., & Miller, K. (1990). *Five papers on WordNet*. Princeton, NJ: Princeton University, Cognitive Science Laboratory.
- National Commission on Writing. (2003). *The neglected R: The need for a writing revolution*. New York, NY: College Board.
- National Commission on Writing. (2004). *Writing: A ticket to work... or a ticket out*. New York, NY: College Board.
- O'Donnell, R., Griffin, W., & Norris, R. (1967). *Syntax of kindergarten and elementary school children: A transformational analysis* (NCTE Research report No. 8). Champaign, IL: National Council of Teachers of English.
- Pearson, P. D. (1974–1975). The effects of grammatical complexity on children's comprehension, recall, and conception of certain semantic relationships. *Reading Research Quarterly*, 10, 155–192. DOI: 10.2307/747180
- Perfetti, C., Landi, N., & Oakhill, J. (2005). The acquisition of reading comprehension skill. In M. J. Snowling, & C. Hulme (Eds.), *The science of reading: A handbook* (pp. 227–247). Oxford: Blackwell. DOI: 10.1002/9780470757642.ch13

- Perfetti, C., & McCutchen, D. (1987). Schooled language competence: Linguistic abilities in reading and writing. In S. Rosenberg (Ed.), *Advances in applied psycholinguistics: Reading, writing, and language learning* (Vol. 2; pp. 105–141). Cambridge: CUP.
- Quirk, R., Greenbaum, S., Leech, G., & Svartvik, J. (1985). *A Comprehensive grammar of the English language*. London: Longman.
- Stewart, M. (1978). Syntactic maturity from high school to university: A first look. *Research in the Teaching of English*, 12(1), 37–46.
- Toglia, M., & Battig, W. (1978). *Handbook of semantic word norms*. New York, NY: Lawrence Erlbaum Associates.
- Wells, R. (1960). Nominal and verbal style. In T. Sebeok (Ed.), *Style in language* (pp. 213–220). Cambridge, MA: The MIT Press.
- Wilson, M. (1988). The MRC psycholinguistic database: Machine-readable dictionary, version 2. *Behavioral Research Methods, Instruments and Computers*, 20(1), 6–11. DOI: 10.3758/BF03202594
- Witte, S., & Faigley, L. (1981). Coherence, cohesion, and writing quality. *College Composition and Communication*, 32, 189–204. DOI: 10.2307/356693
- Xiao, R. (2009). Multi-Dimensional analysis and the study of world Englishes. *World Englishes*, 28(4), 421–450. DOI: 10.1111/j.1467-971X.2009.01606.x
- Zwaan, R., Langston, M., & Graesser, A. (1995). The construction of situation models in narrative comprehension: An event-indexing model. *Psychological Science*, 6(5), 292–297. DOI: 10.1111/j.1467-9280.1995.tb00513.x

Appendix 1

Full Prompt Assignment for each Prompt Title

Prompt Title	Full Essay Prompt
Choices	Do you believe that the choices we make, rather than our abilities and talents, show who we truly are?
Competition	<p>Think carefully about the issue presented in the following excerpt and the assignment below.</p> <p>While some people promote competition as the only way to achieve success, others emphasize the power of cooperation. Intense rivalry at work or play or engaging in competition involving ideas or skills may indeed drive people either to avoid failure or to achieve important victories. In a complex world, however, cooperation is much more likely to produce significant, lasting accomplishments.</p> <p>Do people achieve more success by cooperation or by competition?</p>
Dreaming	Some people say that in our modern world, dominated by science, technology, and industrialization, there is no longer a place for dreaming and imagination. What is your opinion?
Equality	In his novel “Animal Farm,” George Orwell wrote “All men are equal: but some are more equal than others.” How true is this today?
Fame	<p>A sense of happiness and fulfillment, not personal gain, is the best motivation and reward for one’s achievements. Expecting a reward of wealth or recognition for achieving a goal can lead to disappointment and frustration. If we want to be happy in what we do in life, we should not seek achievement for the sake of winning wealth and fame. The personal satisfaction of a job well done is its own reward.</p> <p>Are people motivated to achieve by personal satisfaction rather than by money or fame? Plan and write an essay in which you develop your point of view on this issue. Support your position with reasoning and examples taken from your reading, studies, experience, or observations.</p>
Fitting In	<p>From the time people are very young, they are urged to get along with others, to try to “fit in.” Indeed, people are often rewarded for being agreeable and obedient. But this approach is misguided because it promotes uniformity instead of encouraging people to be unique and different. Differences among people give each of us greater perspective and allow us to make better judgments.</p> <p>Is it more valuable for people to fit in than to be unique and different? Plan and write an essay in which you develop your point of view on this issue. Support your position with reasoning and examples taken from your reading, studies, experience, or observations.</p>

(Continued)

Prompt Title	Full Essay Prompt
Heroes	<p>Having many admirers is one way to become a celebrity, but it is not the way to become a hero. Heroes are self-made. Yet in our daily lives we see no difference between “celebrities” and “heroes.” For this reason, we deprive ourselves of real role models. We should admire heroes – people who are famous because they are great – but not celebrities – people who simply seem great because they are famous.</p> <p>Should we admire heroes but not celebrities? Plan and write an essay in which you develop your point of view on this issue. Support your position with reasoning and examples taken from your reading, studies, experience, or observations.</p>
Images	<p>All around us appearances are mistaken for reality. Clever advertisements create favorable impressions but say little or nothing about the products they promote. In stores, colorful packages are often better than their contents. In the media, how certain entertainers, politicians, and other public figures appear is more important than their abilities. All too often, what we think we see becomes far more important than what really is.</p> <p>Do images and impressions have too much of an effect on people?</p>
Memories	<p>Many persons believe that to move up the ladder of success and achievement, they must forget the past, repress it, and relinquish it. But others have just the opposite view. They see old memories as a chance to reckon with the past and integrate past and present.</p> <p>Do personal memories hinder or help people in their effort to learn from their past and succeed in the present?</p>
Optimism	<p>In many circumstances, optimism – the expectation that one’s ideas and plans will always turn out for the best – is unwarranted. In these situations what is needed is not an upbeat view but a realistic one. There are times when people need to take a tough-minded view of the possibilities of success, give up, and invest their energies elsewhere rather than find reasons to continue to pursue the original project or idea.</p> <p>Is it better for people to be realistic or optimistic?</p>
Religion and Television	<p>Marx once said that religion was the opium of the people. If he was alive at the end of the 20th century, would he replace religion with television?</p>
Truth Telling	<p>It is often the case that revealing the complete truth may bring trouble – discomfort, embarrassment, sadness, or even harm – to oneself or to another person. In these circumstances, it is better not to express our real thoughts and feelings. Whether or not we should tell the truth, therefore, depends on the circumstances.</p> <p>Is it better to always tell the truth?</p>

(Continued)

Appendix 1 (Continued)

Prompt Title	Full Essay Prompt
Originality	<p>We value uniqueness and originality, but it seems that everywhere we turn, we are surrounded by ideas and things that are copies or even copies of copies. Writers, artists, and musicians seek new ideas for paintings, books, songs, and movies, but many sadly realize, "It's been done." The same is true for scientists, scholars, and businesspeople. Everyone wants to create something new, but at best we can hope only to repeat or imitate what has already been done.</p> <p>Can people ever be truly original? Plan and write an essay in which you develop your point of view on this issue. Support your position with reasoning and examples taken from your reading, studies, experience, or observations.</p>
Winning	<p>From talent contests to the Olympics to the Nobel and Pulitzer prizes, we constantly seek to reward those who are "number one." This emphasis on recognizing the winner creates the impression that other competitors, despite working hard and well, have lost. In many cases, however, the difference between the winner and the losers is slight. The wrong person may even be selected as the winner. Awards and prizes merely distract us from valuable qualities possessed by others besides the winners.</p> <p>Do people place too much emphasis on winning? Plan and write an essay in which you develop your point of view on this issue. Support your position with reasoning and examples taken from your reading, studies, experience, or observations.</p>

Appendix 2

Factor 1 (Essay quality: Interval scale)

Essay quality (interval scale)	Factor score
Score: 1	-.154
Score: 2	-.140
Score: 3	-.056
Score: 4	.106
Score: 5	.308
Score: 6	.879

Factor 2 (Essay quality: Low, Middle, High)

Essay quality (low, middle, high)	Factor score
Low	.107
Middle	-.016
High	-.085

Factor 3 (Prompt and Assignment)

Prompt and Assignment	Factor score
Truth telling	.272
Heroes	.272
Competition	.186
Religion and television	.145
Memories	.140
Choices	.067
Dreaming	.059
Winning	.005
Fame	-.011
Uniqueness	-.109
Optimism	-.109
Images and impressions	-.127
Equality	-.160
Fitting in	-.228

Factor 4 (Grade Level)

Grade Level	Factor score
Low	.102
High	-.074