THESIS

OBJECTIVE MEASURES OF WRITING QUALITY

Submitted by

Kristopher Kyle

Department of English

In partial fulfillment of the requirements

For the degree of Master of Arts

Colorado State University

Fort Collins, Colorado

Summer 2011

Master’s Committee:

Advisor: Douglas Flahive

Nancy Berry
Marilyn Thayer

ABSTRACT

OBJECTIVE MEASURES OF WRITING QUALITY

This study explores the use of objective measures to assess writing quality. Coh-Metrix 2.0, an online text analysis tool, was used to measure 54 linguistic properties of argumentative essays written by English as a Second Language (ESL) and English as a Foreign Language (EFL) students. Using discriminant function analysis, this study found that two models of three objective-measure predictors each were able to significantly discriminate between holistically evaluated high and low quality essays written by ESL students. Two different models of three objective-measure predictors were also able to significantly discriminate between high and low quality EFL essays. These models could not, however, significantly discriminate between high and low quality essays across educational settings. The predictors used in each successful model support the idea that essays containing complex language are generally perceived to be of high quality. The results of this study question, however, the idea that coherent texts have more cohesive textual features, at least in the ESL setting. Furthermore, this study highlights the differences between ESL and EFL writing, though these differences may prove to be related to factors other than educational setting due to some limitations of this study.

TABLE OF CONTENTS

CHAPTER 1

Writing quality, and more specifically measuring writing quality, has been an object of study for over fifty years, first in first language studies (eg. Hunt, 1965) and then in second language studies as well. The measurement of writing quality can be separated into three general categories, namely holistic (assigning a single subjective score), analytic (using a number of subjective categories to create a score), and objective (analyzing occurrences of linguistic features). Although holistic scoring, or assigning a single, subjective global score to indicate the quality of an essay, is a widespread phenomenon for a variety of reasons, many have questioned both the reliability and validity of holistic measurements of writing quality (e.g. Charney,1984). The purpose of the current study, however, is not to dispute holistic scoring but to further the body of research that has investigated objective measures of writing quality.

Some researchers have investigated more objective ways to evaluate the quality of writing, which has generally been described as ‘analytic’ scoring. One example of this type of research is Freedman (1979), which sought to ascertain which factors affected holistic scores using a number of indices such as content, organization, sentence structure, and mechanics. Another example is Hamp-Lyons and Henning (1991), which attempted to measure writing quality using seven indices in order to create a multi-trait profile of a piece of writing, which they felt would give an in-depth understanding of a student’s writing competencies. While these studies and others like them attempted to isolate factors that affect writing quality, their more specific criteria were still essentially qualitative, where a score for linguistic accuracy, for example, is given on the basis of a general assessment, not a ratio of the number of correct versus incorrect sentences, clauses, or T-units (an independent clause and all its dependent clauses, as described in Hunt (1965)).

Other early studies have looked at writing quality from a quantitative perspective by calculating various indices of syntactic complexity, or sophistication of the grammar used in a text (e.g. Flahive and Snow (1980), Homburg (1984), Larsen-Freeman (1978), among others- refer to chapter 2 for an in-depth discussion), text length (e.g. Homburg, 1984), various indices of lexical complexity (sophistication of vocabulary used (e.g. Engber, 1995)). Most early objective-measure studies were necessarily laborious and time-consuming due to the need for a researcher to hand-count each instance of each measure, which was compounded by the potential for counting errors (see Polio (1997) as an example). More recent objective measure research has been made less onerous and arguably more accurate with the help of advancing technology. Common computer programs such as RANGE (Nation and Heatly, 2002), Wordsmith (Scott, 1996) and Compleat Lexical Tutor (Cobb, 2010) analyze some lexical features of texts with a click of a button. Other programs such as Grammatik (a word-processor grammar-checker) have, with limited success, been used to analyze errors in student-produced texts (see Li, 2000).

While computer programs such as the ones listed above can be useful for a variety of applications, the analysis of writing quality seems to be more complex than just the vocabulary used or the complexity or correctness of grammar in a text (see Flahive and Snow (1980), Li (2000) among many others-see Chapter 2 for a full discussion). To address the complexity in writing analysis, computer programs have recently been created based on a compilation of programs that can analyze over 50 different objective measures at the same time. Programs such as Coh-Metrix (Graesser, 2004) and e-rater (descriptions of various versions in Chodorow and Burstein (2004) and Attali and Burstein (2005), among others, see chapter 2) for example, can simultaneously analyze a number of different simple and complex lexical, syntactic, and cohesive aspects of a text. This allows researchers to fine tune their analyses and ultimately to create writing models (see Crossley and McNamara (2009), McNamara et al. (2010), Attali and Burstein (2005), among others, see chapter 2).

A number of recent computer-assisted studies have explored objective measures of writing quality. Li (2000) investigated the relationship between computerized objective evaluations and human scored analytic and holistic evaluations of L2 writing samples (n = 132). The author ran a number of correlations between the computer generated scores and the human generated scores for each linguistic variable, but also tried to correlate each computer generated linguistic variable with the holistic rating variable. Though two significant correlations were found, a much more informative statistical analysis could have been conducted in order to determine whether the computer scores for the linguistic features could accurately distinguish the essays into each holistic rating group (eg. multiple regression, logistic regression, or discriminant analysis function for an in-depth discussion, see chapter 2).

Chodorow and Burstein (2004) investigated the use of two versions of a proprietary computerized essay rating program developed by Educational Testing Services (ETS), e-rater99 and e-rater01. Both e-rater99 and e-rater01 were capable of measuring approximately 50 objective measures of writing. During a training process, a number of essays were scored by human raters and then analyzed by the computer program for all measures. Finally, a stepwise linear regression statistic was conducted for each training set in order to select 8-10 measures that most accurately predicted human rated scores. The four basic measures that were used measured various syntactic, discoursal, topical and lexical aspects of the texts (see chapter 2 for more explanation). Although erater-99 predicted human scores within 1 point (on a six-point scale) with 96% accuracy and e-rater01 lessened the effects of essay length on the models, a new model based on human rated essays had to be created for each writing prompt.

Attali and Burstein (2005) described a study that tested a new version of e-rater, e-rater v. 2.0, which was designed to evaluate essays across a variety of writing prompts. One of the main additions that was made to the program was the addition of Criterion (another program developed by ETS) functions, which analyzed errors in grammar, usage, mechanics and style. Reliability of the program in evaluating 6th-12th grade essays, GMAT essays and TOEFL essays was .60, which was higher than single human rater (.50) and the average of two human raters (.58). Erater v. 2.0 essays scores correlated highly with averaged human rater's scores (.93). While the erater programs developed by ETS appear to be fairly adept at predicting human rater scores, these tools are unavailable to the public, as are the exact measures used, negating replication.

Crossley and McNamara (2009) explored the use of objective measures of writing to determine whether linguistic features could distinguish between L1 and L2 writing. They used Coh-Metrix (Graesser et al., 2004), to analyze a corpus consisting of 195 argumentative essays written in English by Spanish L1 students taken from a section of the International Coprus of Learner's English (ICLE) and 208 argumentative essays written in English by native speakers. Through the use of discriminant function analysis, the researchers identified seven indices (hypernymy (the use of word that are semantically related to each other in a hierarchy, such as chair and furniture-furniture is a hypernym of chair**)**, argument overlap, motion verbs, word frequency, polysemy (use of words with multiple meanings), latent semantic analysis givenness (a complex computerized assessment of cohesion in a text), and age-of-acquisition scores-see chapter 2 for a full discussion of these measures) that could distinguish between L1 and L2 writing with an accuracy of 79%. This study demonstrated the flexibility of Coh-Metrix, the usefulness of discriminant analysis in corpus linguistics, and identified a number of indices that can be used to distinguish between texts.

McNamara et al. (2010) used Coh-Metrix to determine if measures of cohesion, syntactic complexity, diversity of words, and characteristics of words could predict the holistic quality of essays (n = 120) written by students in a freshman composition course at Mississippi State University. The essays were first expertly graded using a holistic rubric that ranged from 1-6. After the data had been analyzed in Coh-Metrix, the researchers used discriminant analysis to determine whether the particular indices could predict either low or high proficiency membership. After using a training set of essays (n = 80) to select the most effective indices, they were able to correctly place 67% of the essays into the correct group on the test set (n = 40) using discriminant analysis. Using multiple regression analysis, they determined that syntactic complexity (operationalized as number of words before the main verb), lexical diversity (operationalized as the Measure of Textual and Lexical Diversity (McCarthy, 2005)), and word frequency (operationalized as measured by CELEX (Baayen, Piepenbrock, & Gulikers, 1995, which is based on frequency data found in the COBUILD corpus) were the strongest predictors of group membership. While this study indicates that objective measures can be used with limited success to differentiate between high and low proficiency L1 texts, it is unclear whether these indices would also predict high or low proficiency of texts written in an L2 and if so, whether the same indices could be used for both L1 and L2 writing across educational settings.

Based on the extant body of research, the current study investigated the following research questions:

1.) Can the various text analysis functions of Coh-Metrix be used to create a model that can distinguish between high and low quality essays in the L2 context?

2.) If a model can be created, which objective measures are the strongest predictors of writing quality?

3.) Are objective predictors of writing quality constant across educational settings?

Following McNamara et al. (2010), the current study used Coh-Metrix to analyze two corpora, but in the L2 context. The first corpus, the Intensive English Program Learner English Corpus (IEPLEC), was comprised of problem/solution essays that were written in an intermediate class at an intensive English program (n=60). The second corpus was taken from subcorpora of the International Corpus of Learner English (ICLE), a collection of argumentative essays written by advanced L2 writers from a variety of L1 backgrounds (n=60). Both sets of essays were analyzed holistically by human raters and then analyzed by Coh-Metrix. A model was created based on the first set of essays and then retested on the second set to check the versatility of the model. For information about this methods used in this study, including the two corpora used, the holistic assessment, the objective assessment, and the analysis of the data, see chapter 3.

See chapter 4 for the results of this study, including descriptive data concerning the corpora and the predictive power of a number of sets of objective measure in discriminating between high and low quality essays.

See chapter 5 for a discussion of the findings, including pedagogical implications, how this study supports/questions previous research, and the limitations of this study.

CHAPTER 2

The measurement of writing quality has been the object of research for many years. Many studies have investigated the reliability and validity of holistic measures, often comparing them to human-evaluated analytical rubrics in which essays are assigned scores based on the summation of scores in sub-areas. Another vein of research in writing quality has taken the analytical side of previous studies a step further with human-evaluated objective measure studies. With the advance of technology, not only have traditional objective measures such as syntactic and lexical complexity been made less time consuming to study, but new objective measures, such as latent semantic analysis, have been developed, opening new possibilities for text analysis. In addition, new computer programs such as Coh-Metrix (Graesser et al., 2004) and e-rater v. 2.0 (Attali and Burstein, 2005) can analyze a number of objective measures at once, enabling researchers to test potentially robust models of writing quality.

Analytic Measures

A large body of research exists concerning analytic, or multi-trait scoring. In general, analytic scoring developed as a way to provide learners more information about their writing strengths as weaknesses than a single holistic score could (Hamp-Lyons, 1995). I have reviewed a few of these studies in this paper to highlight the similarities and differences between these types of measures and objective measures.

An early example of analytic evaluation of writing quality is described in Brown and Bailey (1984). Brown and Bailey hoped to focus 'raters' attention on a number of specifically defined criteria' (p. 28) in order to produce a more precise evaluation of writing quality. To these ends five equal categories were used, Organization, Logical Development of Ideas, Grammar, Mechanics, and Style. They found this scale to be fairly reliable (r = .72) using two raters after manipulating the data slightly. As no comparison was made to another scale, the sole purpose for the study was to determine the reliability and usability of the scale.

Another example of an analytic scoring system is described in Hamp-Lyons and Henning (1991). Hamp-Lyons and Henning evaluated L2 English texts from two sources, the Test of Written English (TWE) and the Michigan Writing Assessment (MWA), using a variation of the New Profile Scale, which they dubbed the Experimental Communicative Profile Scale (ECPS), to determine whether their scale was transferable between different tasks. The ECPS included seven categories, Communicative Quality, Interestingness, Referencing, Organization, Argumentation, Linguistic Accuracy, and Linguistic Appropriacy. Using three raters, they found that some of their categories were reliable in some contexts, while others were not. Communicative Quality and Linguistic accuracy, for example, obtained the highest interrater reliability scores based on adjusted Spearman-Brown adjusted coefficients (.896 and .905 respectively) with regards to the TWE papers. Estimates of reliability were generally lower with the MWA samples, and the most reliable categories were Referencing (.807) and Argumentation (.716). Overall, global scores (the summation of scores for each category) were highly reliable for both TWE (α = .94) and MWA (α = .90).

Although analytic measures were designed in part to provide in-depth feedback about writing performance to students (Hamp-Lyons, 1995), there is evidence that this goal is not always achieved. Lee, Gentile, and Kantor (2010) found that all six of the analytic measures used in their analysis (development, organization, vocabulary, sentence variety, grammar/usage, and mechanics) of 930 essays written for the TOEFL Computer Based Test (CBT) correlated highly and significantly with holistic scores, indicating that these analytic scores may not have provided a deeper evaluation of each writer's relative strengths and weaknesses.

Analytic measures seek to compartmentalize distinct features of writing in order to help students improve their writing in specific areas. Although this may occur (Hamp-Lyons, 1995), analytic measures are still subjective and subject to human evaluation. Objective measures, on the other hand, are not subjective. If valid objective measures of writing quality could be identified, then they could also be used to provide specific information about student writing.

Human-Analyzed Objective Measures

In an attempt to mitigate the subjective nature of writing evaluation, researchers have attempted to identify objective measures of writing quality that correlate with perceived quality as measured by holistic or analytic scales. In general, these studies focused on syntactic complexity measures concerned with the T-unit (Hunt, 1965). While many of these studies were necessarily time consuming and impractical for large corpus analysis, they form the building blocks of current computerized objective measure studies.

Larsen-Freeman (1978) described a subset of a larger study designed to create an index of ESL development in which a number of objective measures were used to evaluate 212 written placement exams written by prospective EAP students. The essays were divided into five groups based on each student's performance on the entire placement exam (which was not fully described but included components additional to the written exam). No information was given regarding the evaluation process or the inter-rater reliability on each measure. Using ANOVA statistics (F ratios), Larsen-Freeman found that the differences between groups regarding average words per composition were significant at the .001 level, average words per T-unit and percentage of error-free T-units were significantly different between proficiency groups (no p values were given), and the average number of words per error free T-units between proficiency levels were ‘highly significant’ (p. 446). Although this study is extremely vague regarding reliability and the methods used, it is a good example of early objective measure studies.

Flahive and Snow (1980) investigated whether four syntactic complexity measures could discriminate between essays written by three proficiency groups (N = 300) under examination conditions and what if any relationship existed between these complexity measures and writing quality . The four measures included were the mean length of T-units, mean ratio of clauses per T-unit, the mean number of errors per T-unit, and the Index of Complexity. The Index of Complexity, which was influenced by Endicott (1973), was calculated by assigning a complexity score to each T-unit and dividing that score by the number of words in the T-unit. Complexity scores were based on the incidence of certain grammatical structures selected on the basis of their frequency in advanced ESL writing. Derivational morphemes and adjectives received a score of ‘1’, relative clauses, passive sentences, embedded questions, possessives, and comparatives received a score of ‘2’, and noun clauses received a score of ‘3’. Analyses indicated that only two of the measures, mean length of T-unit and the ratio of clauses to T-units, were able to discriminate between the proficiency levels. Neither errors per T-unit nor the Index of Complexity were able to discriminate between the proficiency groups. To investigate whether there were any significant correlations between the measures of syntactic complexity and holistic evaluations of essays, the essays were scored on a 5-point scale by experienced ESL teachers. Inter-rater reliability was quite high ($\alpha > .90$). They found that the relationship between mean ratio of clauses per T-units increased as quality increased. They acknowledged that while writing is too complex to be evaluated on the basis of T-unit measures alone, these measures can be useful indicators of proficiency level and writing ability.

Perkins (1980) attempted to determine which of ten objective measures would discriminate between essays that were holistically evaluated to be passing, low passing and failing compositions (n = 29) written by ESL students studying EAP in the same level. The objective measures used were categorized as word and sentence counts, syntactic complexity with no attention to error, and syntactic complexity with attention to error. In addition, he tested a standardized multiple-choice writing test used on the SAT. He found that objective measures that took error into account were able to discriminate between the three essay quality levels. These measures were error free T-units per composition, number of words in error-free T-units per composition, errors/T-unit, and total errors. Interestingly, the standardized test approached but did not reach significance with regard to discriminating between the essays.

Homburg (1984) attempted to identify objective measures that affected reader's choices in assigning holistic scores. From a collection of texts from the writing section of the Michigan Test of English Language Proficiency (MTELP) which had been assigned holistic scores ranging from one to ten, 30 were selected from the five, six and seven levels each, for a total of 90 texts. Only texts that achieved an inter-rater reliability of 1.0 were included in the data set. The essays were then evaluated the essays using objective measures of length, subordination and relativization, sentence connectors and number and types of errors. A stepwise discriminant analysis was conducted and showed that five measures differentiated between the three levels accounting for 84% of the variance. Second-degree errors per T-unit which as defined by Nas (1975) included spelling, word choice, and grammatical errors that make comprehension difficult (but not impossible) accounted for the largest variance, followed by dependent clauses per composition, words per sentence, coordinating conjunctions per composition, and error-free t-units.

Sparks (1998) investigated the validity of objective syntactic complexity measures and subjective indices of the frequency and seriousness of errors when compared to a holistic scale using practice TWE essays written by 30 college-level ESL students. Two 'experienced ESL teachers' rated the essays using a six-point holistic scale developed by ETS for the Test of Written English (TWE). No inter-rater reliability scores were given. The essays were then evaluated for syntactic complexity based on three measures devised by Arena (1982), which included the number of clauses per sentence, the number of clauses per main clause, and the average value of embedded clauses. These measures were deemed to be 'reliable' (p.43) but no reliability figures were given. Finally, the essays were evaluated for errors based on a method adapted from Brodkey and Young (1982), where errors are given a seriousness score on a scale of 3. An error rated '3' seriously distorted meaning, an error rated '2' moderately distorted meaning, and an error rated '1' did not significantly distort meaning. Interestingly, only the first 120 words in each essay were evaluated using this scale. The correctness measure was mentioned to be 'not highly dependable' (p.43) based on intra- and inter-rater reliability, but no reliability figures were given. Only one measure, the correctness score, correlated significantly with the holistic rating (spearman's rho = .644, p = .01).

Polio (1997), in a departure from other studies that attempted to use objective measures to discriminate between the proficiency level of ELLs or writing quality, investigated the reliability of a number of measures of writing quality that had been previously employed. In this study, she investigated the intra- and inter-rater reliabilities of a holistic scale, error-free T-units, and an error classification system. As data, she used 38 one-hour essays written by graduate and undergraduate ESL students. For each measure she rated each essay twice and a graduate student rated each essay once. The holistic scale, which was based on linguistic accuracy, earned an intra-rater reliability of .77 and inter-rater reliabilities of .44 and .55, respectively. The ratio of error-free T-units earned an intra-rater reliability of .91, while inter-rater reliabilities were both .80. The ratio of error-free t-units to total clauses earned an intra-rater reliability of .93, while inter-rater reliabilities were .80 and .85. The ratio of error-free T-units to total words earned an intra-rater reliability of .93 while inter-rater reliabilities were .76 and .78. With regards to error-counts, intra-rater reliability was .89 while inter-rater reliabilities were .94 and .89. This study is pertinent to the present study as it shows that although human-rated objective measures can be much more reliable than holistic scoring, reliability is still an issue when human analysis is involved.

Computer-Analyzed Objective Measures

Although computer analyzed objective measures have departed from traditional T-unit analysis of syntactic complexity, new measures have continued to be developed and tested. Some of these new measures are related below. Although the measures reviewed are by no means comprehensive, I have focused on features and compilations of features that are either longstanding proponents of reading and writing research (such as lexical measures) or have been well researched considering how recently they have been developed, and that have either been shown to be quite successful (such as e-rater 2.0), or show potential for success in the area of writing (such as Latent Semantic Analysis and Coh-Metrix).

Lexical measures.

Although lexical measures were devised long before the use of computers, computer programs such as Compleat Lexical Tutor (Cobb, 2010) have made the analysis of various lexical features of texts, such as frequency counts very easy and efficient. One particularly promising lexical measure of writing quality is lexical diversity, which is calculated, broadly speaking, by dividing the number of different words in a text (types) by the number of total words in a text (tokens). Engber (1995), for example, demonstrated that lexical diversity was significantly correlated with holistic evaluations of essays. Yu (2010), also found lexical diversity was significantly correlated with human evaluations of writing quality. Lexical diversity is also one of the predictors in the highly accurate essay evaluation program e-rater 2.0 (Attali and Burstein, 2006), and is included in the computerized text analysis tool Coh-Metrix (Graesser et al., 2004). Although many more studies have addressed this well-researched topic, I have chosen to exclude an in-depth treatment in favor of discussing new measures, such as latent semantic analysis, and text analysis tools that include lexical diversity.

Latent semantic analysis.

One increasingly prevalent text analysis measure that has been created with the assistance of computers is Latent Semantic Analysis (LSA). Although LSA has been used both to model acquisition of knowledge and to represent the “contextual usage substitutability” (Landauer, Foltz, and Laham, 1998a, p. 260) of words, only the latter will be discussed in this paper as this is the way it has been used with regard to writing quality. Following is an explanation of the conceptual underpinnings of LSA and examples of ways it has been used in relation to measuring text features to discriminate between texts.

LSA, according to Landauer et al., (1998a) is "a fully automatic mathematical and statistical technique for extracting and inferring relations of expected contextual usage of words in passages of discourse." (p. 263). It uses no dictionaries or any other information outside of the analyzed texts to determine the semantic relatedness of words, but calculates semantic meaning based on raw data alone. LSA is also not merely a function of co-occurrence counts in that it analyzes "unitary expressions of meaning" (p. 261) – or the sum of the words used in sentences and paragraphs instead of word-word co-occurrence frequencies. Usually this is done using a large corpus or a number of corpora in order to account for a large number of contexts in which a word may be used.

LSA first creates a matrix where each word in a text (or groups of texts) represents a row and each meaning unit, such as a sentence or a paragraph, represent a column. Weight is then given both locally and globally to each word, the matrices are reduced using singular value decomposition, where each word is given three values, 'a term-concept vector matrix, a singular values matrix, and a concept-document vector matrix' (Jorge-Botana, Leon, Olmos, and Escudero, 2010, p. 2). To measure the similarity between two words or sections of texts, LSA calculates the cosine between the semantic vector of each section. The closer the cosine is to 1, the more related the two words or sections of text are (Jorge-Botana et al. 2010).

Although LSA is limited by the fact that it does not take into account syntax, logic, or morphology, researchers have demonstrated its ability to model semantic relationships with a degree of accuracy. Landauer, Foltz, and Laham (1998b) (unpublished manuscript described in Landauer et al., (1998a)), for example, reported the results a study in which LSA was trained on a ‘large corpus of representative English’ and then took a TOEFL vocabulary test (which was for the most part a synonym test) based on the information gained from the training corpus. On an 80-item multiple-choice test, LSA earned a score of 65% by making choices based on the cosines between the prompt word and each potential answer. Whichever potential answer had the highest cosine value with the prompt was chosen. Although a score of 65% may seem low, it was reported to be the same as the average score of students applying to universities in the United States. This study clearly showed that LSA is capable of determining the semantic relatedness between words with the same accuracy as English L2 students applying for universities.

Foltz, Kintsch and Landauer (1998) demonstrated the ability of LSA to automatically measure the coherence of texts in the L1 context. To do so, they used an essay from a study conducted by Britton and Gulgoz (1991) that was re-written three different ways and then read by readers who were subsequently assessed on comprehension of the passage. In the Britton and Gulgoz (1991) study, two of the rewritten versions (Principled and Heuristic) earned higher comprehension scores than the other two versions (Original and Readability). After conducting an LSA training session with the first 2,000 characters of 30,473 articles from *Grolier’s Academic American Encyclopedia*, LSA determined that the two versions of the essays that earned higher comprehension scores also the most coherent.

In the same study, Foltz et al. (1998) used data from McNamara, Kintsch, Songer, and Kintsch (1996) was also used to test the ability of LSA to measure coherence. McNamara et al. (1996) investigated the role that previous knowledge played on text coherence. They manipulated a text with respect to local and macro coherence, creating four versions of the text- one with low local coherence and low macro coherence, one with high local coherence but low macro coherence, one with low local coherence but high macro coherence and one with high local coherence and high macro coherence. After conducting two training sessions to create two sets of LSA scores, one using a small corpus of encyclopedia articles and one using a large corpus of encyclopedia article, Foltz et al. (1998) evaluated each version of the text with each version of LSA. In addition, they tested the four versions of the test with regard to argument overlap. Although significant differences were not found between the LSA analyses of each test, a linear trend was observed in which LSA coherence values from both the small-corpus trained LSA and the large-corpus trained LSA rose as coherence did. Additionally, LSA scores achieved a linear relationship with coherence, while argument overlap did not. Foltz et al. (1998) concluded that the analysis of the first data set indicated the ability of LSA to accurately measure coherence, while the second data set demonstrated that LSA measured coherence much better than and therefore distinct from mere word overlap measures.

Leon, Olmos, Escudero, Canas, and Salmeron (2006) investigated the effectiveness of LSA in evaluating short (50-word) narrative and expository summaries. To train LSA for this study, a corpus of 2,059,234 documents taken from the internet, textbooks, encyopledias, newspapers, and literary books were used. They used 390 middle and high school student-produced summaries and six expertly written summaries about the narrative text and 192 summaries written by middle and high school students about the expository text for their study. They tested six different ways of evaluating the quality of summary writing including comparing the LSA cosine of the source text to the summary, comparing the cosine of each summary to that of all of the summaries written by students as a whole, comparing the cosine of each essay with that of an expertly-written summary, comparing the cosines of 100 holistically graded summaries with those of ungraded essays to obtain a score, comparing the cosine of each sentence of the source text with that of a summary, and finally comparing the cosines of sentences from the source document deemed to be important by experts to the cosine of a summary. All summaries were holistically graded by four PhD students on a four-point scale for content and then on a six-point scale for coherence. Inter-rater reliability was acceptable for the narrative summaries (ranging from .81-.86) with regards to content, but fairly low (.66-.75) for coherence. For the expository summaries, inter-rater reliability ranged from .53-.81 for content and from .58-.79 for coherence. All six LSA evaluation methods significantly correlated with the human ratings of narrative summaries at the .001 level. Although the correlations between most LSA methods and human raters were significant for the expository summaries, the overall strength of the correlations were weaker than those of the narrative summaries. For the expository summaries, the first four LSA evaluation methods correlated significantly at the .001 level with all human evaluators. The fifth measure significantly correlated with two of the human raters at the .001 level and with the other two raters at the .01 level. The final LSA evaluation method correlated significantly with two human raters at the .01 level, with one of the human raters at the .05 level, and did not significantly correlate with the final human rater. They concluded that LSA can be considered a reliable tool for writing evaluation, even when the texts to be evaluated are quite short.

Many of the studies demonstrating the usefulness of LSA focus on its use within Coh-Metrix (Graesser et al., 2004), which are reviewed later in this chapter.

Multiple Measure Studies

E-rater.

Among recent computerized objective measure studies, studies evaluating various automatic essay scoring (AES) systems have played a prominent role. Among these, studies reporting on various versions of e-rater, an automatic writing evaluation tool developed by Educational Testing Services (ETS), have played a large role. I have included a review of some of these studies to provide some background concerning current methods of objectively evaluating writing quality which studies have shown to be quite successful (eg. Attali and Burstein, 2005; Attali and Burstein, 2006; Attali (submitted for publication); Chodorow and Burstein, 2004). One caveat to these studies, however, is that the specific measures used are often not divulged for proprietary reasons, and ultimately are not available for public investigation.

Chodorow and Burstein (2004), investigated the accuracy of two versions of e-rater, e-rater99 and e-rater01 when the effect of essay length was removed. Although these two versions of e-rater were similar, e-rater01 included measures not included in e-rater99. They used a large number of essays (1,855 for the training set and 9,597 for the cross-validation set) written for the computer-based version (CBT) of the TOEFL which were written on seven prompts. Participants were given 30 minutes to complete the writing task. TOEFL CBT essays were first evaluated holistically on a six point scale by two raters. To calculate the final score, the mean of the two raters' score was taken, provided that the scores did not vary by more than one point. If scores varied by more than a point, a third rater assigned a score and the mean of the adjacent or matching score is used. A comparison was made between holistic scores and essay length in the training set, which indicated that there was a significant relationship between the two. Holistic scores were then predicted for the cross validation set based on the comparisons made in the training set. They found that scores produced using length as the only predictor matched holistic scores half of the time and came within one point of holistic scores 95% of the time. They then evaluated the essays using e-rater99 and e-rater01. Through a number of analyses, they determined that while e-rater99 predicted holistic essays no better than essay length alone, while e-rater01 was significantly better than length at predicting holistic scores. Compared with two human raters, e-rater01 predicted matching scores slightly less often than human raters, but predicted adjacent scores just as often as human raters did. Overall, this study demonstrated the effectiveness of e-rater01, and the potential for automatic scoring systems based on objective measures to correspond with holistic ratings, at least in a very controlled setting. One integral weakness with e-rater99 and e-rater01 was that they relied on comparing linguistic traits of essays with those of training essays of the same prompt. In addition, 8-10 measures out of 50 were selected for the best fit within each prompt and potentially vary between prompts. This is a potential issue as these versions of e-rater must be trained for every writing prompt, which can be fairly inefficient, and scores can be hard to interpret as it is possible for two prompts to use a completely different set of predictive measures. In addition, although general information about the indices used in e-rater has been given, no specific information regarding each measure has been given, limiting subsequent research outside of the ETS realm to test these measures in contexts outside of standardized testing.

Attali and Burstein (2005) described a new version of e-rater, e-rater 2.0, which addressed a number of the issues in previous versions. In addition to including new features, such as grammatical accuracy measures, e-rater 2.0 included a general model of

8-12 measures that could be used across prompts, eliminating the need for a large number of training sets. In addition, the general model allowed for comparison across prompts. Using data from Criterion Essays written by students from different grade levels (6-12th grade), scores produced by e-rater 2.0 correlated highly (.93) with holistic scores. Attali and Burstein (2005) demonstrated that e-rater 2.0 models may be more generally useful than previous versions of e-rater. In addition, they demonstrated that automatic scoring programs can create scores that correlate highly with holistic scores.

Attali and Burstein (2006) elaborate on the measures used in e-rater 2.0. The eight features (some of which are comprised of a number of micro features) include Grammar, Usage, Mechanics, Style, Organization, Development, Lexical Complexity, and Prompt-Specific Vocabulary Usage. Enright and Quinlan (2010) elaborated upon some of the information provided in Attali and Burstein (2005) and Attali and Burstein (2006), and reported on Attali (submitted for publication), which showed that e-rater 2.0 scores correlated higher with a human rater's scores than two human scores correlated with each other. One potentially useful piece of information that can be gleaned from these studies were the weights given to each feature in the general e-rater model that has been shown to be quite successful. In this model, Organization accounted for 32% of the score, Development accounted for 29%, Mechanics for 10%, Usage for 8%, Grammar and Lexical Complexity (word length) accounted for 7% each, Lexical Complexity (use of less frequent words) accounted for 4%, and Style accounted for 3% of the score. Again, although these studies indicated that computerized automatic essay scoring can mirror that of human scoring, proprietary issues surrounding e-rater make it difficult to ascertain exactly how some of these features are measured. It is therefore unclear how useful these models are outside of a standardized testing setting, and no clear way to follow this line of investigation.

Independent Multiple Measure Studies

Li (2000) investigated the relationship between computerized objective evaluations and human scored quasi-objective and holistic evaluations of L2 writing samples (n = 132). The computerized measures included measures of syntactic complexity, lexical complexity, and grammatical accuracy. Syntactic complexity was operationalized as average sentence length and ratio of subordinated structures to the combination of subordinated and coordinated structures. Lexical diversity was operationalized as the number of different words divided by the number of total words, including both content and function words (lexical diversity) and the number of lexical items (no function words) divided by the total number of words (lexical density, as explained by Laufer and Nation (1995)). Grammatical accuracy was operationalized as the ratio of grammatical errors to total number of sentences and the ratio of different types of grammar errors to total number of sentences. Human raters evaluated the essays on the basis of sentence, vocabulary, and grammar, and also gave the essays a holistic score. The only statistically significant correlations that were observed between computer and human rating were between both computerized measures of grammatical accuracy and the human-evaluated measure of grammar (r = .30, p < .01 for both measures). The author postulated that the computerized syntactic measures failed to correlate significantly with the human rater category of sentence because the chosen measures did not fully account for the variables that comprise sentence dynamics. The study had some particularly important limitations that the authors did not identify. One of these is that the samples included four separate text types (narrative, informative, persuasive, and expressive), but differences in text types were not accounted for. No mention was made about the ratio of different text types or how/if text type affected the results. As shown by Cumming, et al. (2005), text types can be distinguished on the basis of linguistic features, necessitating the separation of text types in an analysis in order to have comparable samples. Another limitation was the level of acceptance for inter-rater reliability, which was set at 0.60. Results may have been different had the researchers set a more rigorous standard for inter-rater reliability (such as 0.80). The final limitation was the way in which the holistic rating variable was statistically analyzed. The author ran a number of correlations between the computer generated scores and the human generated scores for each linguistic variable, but also tried to correlate each computer generated linguistic variable with the holistic rating variable. Though two significant correlations were found, a much more informative statistic could have been conducted in order to determine whether the computer generated scores for the linguistic features could accurately distinguish the essays into each holistic rating group (i.e. multiple regression, logistic regression, or discriminant analysis function).

Coh-metrix.

Graesser et al. (2004) provided rationale for and an overview of the text analysis tool Coh-Metrix. Essentially, Coh-Metrix is compilation of a number of computational linguistic measures. A majority of the included measures focus on a variety of ways to measure cohesion, although it also includes a number of other measures including but not limited to lexical counts and readability measures. Although the first version of Coh-Metrix was designed for in-house use only and included 200 measures (Graesser et al., 2004), the current version, Coh-Metrix 2.0, is available for public use over the Internet, includes 54 measures (See appendix A for a complete listing and short description of the measures included in Coh-Metrix 2.0. See chapter 3 for a more in-depth description of select measures). As Graesser et al. (2004) indicated, Coh-Metrix is very easy to use as it automatically processes a text for all measures and produces an output file that can be read by Excel or Word.

Coh-Metrix has been used to discriminate between a variety of texts and for a variety of purposes. Hall, McCarthy, Lewis, Lee, and McCarthy (2007), for example, investigated the use of Coh-Metrix to discriminate between the language used in American and English/Welsh legal cases. In order to mitigate the effect of genre, a corpus of court cases were selected that dealt with commercial competition. 200 American cases were selected as well as 208 English/Welsh cases. To control for text length, only continuous sections of 1000 words or more were included in the corpus. The corpus was divided into two roughly equal parts in order to establish training set and a test set. Based on previous research, five Coh-Metrix categories were chosen for analysis, including Co-referential Cohesion, Causal Cohesion, Local-grammatical Cohesion, Latent Semantic Analysis, and Lexical Diversity. As Coh-Metrix includes a number of measures for each category, an ANOVA was conducted on the training set between the measures in each category. The measures with the largest effect size in each category (which were not listed) were then used as predictors in a discriminate function analysis of the test set. Overall, Coh-Metrix was able to predict which community (American or English/Welsh) produced a text concerning a commercial competition legal case with 85% accuracy.

Crossley, Louwerse, McCarthy, and McNamara (2007) used Coh-Metrix to analyze linguistic differences between simplified and authentic texts. They used a corpus of 36,747 words taken from 105 texts from ESL textbooks for beginners in three skill areas: grammar, reading, and writing. In addition, texts were selected from basic readers. The simplified texts totaled 21,117 words, while the authentic texts totaled 15,640 words. The texts were analyzed using seven categories of measures including causal cohesion, connectives and logical operators, coreference, density of major parts of speech, polysemy and hypernymy, syntactic complexity, word information, and frequency. After running t-tests between simplified and authentic texts on each measure, they concluded that simplified texts ‘provide ESL learners with more coreferential cohesion and more common connectives and rely more on frequent and familiar words than do authentic texts’ (p. 27). Furthermore, simplified texts used less diverse parts of speech, had les causality, and relied less on logical operators, while demonstrating a higher level of syntactic complexity than authentic texts. They also concluded that there was no significant difference between simplified and authentic texts with regard to abstractness and ambiguity. Although this study analyzed a fairly small corpus of texts, it demonstrated the ability of Coh-Metrix to provide an in-depth analysis of textual features.

Crossley and McNamara (2008) described a principled replication of Crossley et al. (2007). Instead of analyzing a small corpus (36,747 words) of authentic and simplified texts written for beginner learners of English irrespective of age as in Crossley et al. (2007), Crossley and McNamara (2008) analyzed the differences between a larger corpus (128,294 words) of authentic and simplified texts written for intermediate adult learners of English. These texts were analyzed by Coh-Metrix using the same categories as the Crossley et al. (2007) study, which included causal cohesion, connectives and logical operators, coreference, density of major parts of speech, polysemy and hypernymy, syntactic complexity, word information, and frequency. After the texts were analyzed by Coh-Metrix, a series of t-tests were conducted in order to determine if any significant differences were observed between the authentic and simplified texts. Overall, the differences observed in Crossley et al. (2007) were also observed in Crossley and McNamara (2008), although there were a few differences. For example, in the Crossley et al. study, higher values were found for authentic texts than simplified texts for Causality and infrequent parts of speech, while in the Crossley and McNamara study, no significant difference was observed. Furthermore, in the Crossley et al. (2007) study, higher values were found for simplified texts than authentic texts for syntactic complexity, while in the McNamara and Crossley (2008) study, higher values in this area were found for authentic texts. More discrepancies between the studies existed where one study found significant results while the other didn't, but these are not necessarily pertinent to the current study. When differences occurred, Crossley and McNamara (2008) tentatively concluded that they were due to the larger corpus and to the more complex structures selected/created for intermediate texts written for adults. This study further demonstrates the ability of Coh-Metrix to analyze large amounts of texts automatically. In addition, the authors suggested that Coh-Metrix is a useful tool for textbook developers.

Based on the assumption that popular text readability indices such as Flesch reading ease (Flesch, 1948) and Flesch-Kincaid grade level (Kincaid, Fishburne, Rogers and Chissom, (1975) are based upon the analysis of 'shallow' textual features, Crossley, Greenfield, and McNamara (2008) investigated the use of Coh-Metrix to analyze 'deep' textual features to assess text readability. The study used a corpus of 31 short academic texts (the mean number of words = 269.28) on a variety of topics taken from a study by Bormuth (1971). Scores from cloze tests administered to 200 Japanese students to test comprehension of the same Bormuth (1971) texts from a previous study by Greenfield (1999) were used to assess the difficulty of each passages. Based on previous psycholinguistic research regarding reading, a lexical index, a syntactic index, and a meaning contruction index from Coh-Metrix were selected to assess readability. They chose *CELEX frequency score* as the lexical index, *syntactic similarity; sentence to sentence similarity, adjacent, mean* as the syntactic index, and *content word overlap* as the meaning-construction index. After each text was analyzed by Coh-Metrix, they used multiple regression analysis to determine to what degree the three predictors accounted for the variance in comprehension scores. The three predictors produced a multiple correlation of 0.93 and an R^2 of 0.86, indicating that they accounted for 86% of the variance in comprehension scores. When compared with other measures such as Flesch reading ease and Flesch-Kincaid reading level among others, the Coh-Metrix EFL index was significantly better at accounting for variance in EFL comprehension test scores. While this study was limited by the use of a single comprehension assessment measure, and was used solely with academic texts, it demonstrated Coh-Metrix's ability to improve on previous text analysis measures, even when limited by the number of indices used (in this case three out of more than 54 available).

Although comparing features of L2 language production against the standard of L1 production has been discouraged by leaders in the field of applied linguistics (e.g. Ortega, 2009), some such studies have provided insights into productive indices of writing. Crossley and McNamara (2009), for example, explored the use of Coh-Metrix indices to discriminate between English-language essays written by L2 writers and essays written on the same topics by L1 writers at a university in the USA. English essays written by Spanish L1 writers (taken from the International corpus of Learner English

(ICLE), n = 195) and the essays written by native speakers were controlled for length (between 500-100) words, age (early 20's), education (college students), and essay topics (the four most common topics in the ICLE). These essays were then divided into a training set (n = 201) and a test set (n = 202). All essays were analyzed by Coh-Metrix, and then an ANOVA was conducted on the training set to determine which variables had the highest effect sizes. These variables, *word hypernymy*, *word polysemy*, *argument overlap*, *number of motion verbs, CELEX written frequency*, *age of acquisition, locational nouns, LSA givenness*, *Colorado meaningfulness, and incidence of causal verbs* were then used as predictors in a discriminant function analysis of the test set in order to determine whether these predictors could accurately predict whether an essay was written in an L1 or an L2. The model using all ten predictors discriminated between L1 and L2 essay authorship with an accuracy of 78.11%. Nine more discriminant function analyses were conducted to determine if fewer predictors would produce a more accurate model. Predictors with the lowest effect sizes were removed one-by-one until only one predictor remained. The most accurate model, including the seven predictors *word hypernymy*, *word polysemy*, *argument overlap*, *number of motion verbs*, *CELEX written frequency*, *age of acquisition*, and *LSA givenness* was able to predict group membership with 79.1% accuracy. This study showed the ability of Coh-Metrix to automatically analyze a number of variables that can be used to significantly discriminate between texts. It also provided a number of potential indices that can be used to discriminate between texts in subsequent studies.

As many previous studies demonstrated the ability of Coh-Metrix to discriminate between a number of different types of texts by analyzing both shallow and deep features automatically, Coh-Metrix has also been used to discriminate between low and high quality essays. McNamara, Crossley, and McCarthy (2010) investigated whether Coh-Metrix indices of cohesion, syntactic complexity, lexical diversity, and the characteristics of words could be used to identify linguistic features of writing quality in the L1 context. They used argumentative essays (n = 120) written by university students taking a freshman composition course for their study controlled for essay length (500-1,000 words) and essay topic (4 topics used in the ICLE). The essays were rated by five experienced writing tutors using a holistic 6-point SAT writing rubric. All raters scored a training set of 20 essays and scores were re-evaluated until all inter-rater combinations reached acceptable reliability (*r* = .80). Each rater then rated 20 subsequent essays. The mean essay score was 3.26, and essays that received a score between 1-3 were considered low quality (n = 67), while essays that received a score between 4-6 were considered high quality. The results of an ANOVA statistic determined that essay topic did not significantly affect the holistic evaluation given by raters. The essays were then analyzed by Coh-Metrix and divided into a training set (n = 80) and a test set (n = 40) for further analysis. An ANOVA was run on the training set to determine which of the 53 measures included in Coh-Metrix 2.0 in the categories of coreference, connectives, syntactic complexity, lexical diversity, and word characteristics distinguished between high and low quality essays the most. The results of the ANOVA indicated that none of the coreference or connective measures significantly distinguished between the two sets of essays. The measures with the highest effect sizes for each of the three remaining categories were *number of words before the main verb*, *Measure of Textual and Lexical Diversity* (McCarthy, 2005), and *CELEX* (Baayen, Piepenbrock & Gulikers, 1995) *word frequency (logarithm including all words)*. The ability of these measures to predict the group membership of high and low quality essays was tested using discriminant function analysis, first with the training set and then with the test set. The model using the three predictors was able to significantly predict the group membership of 52 of the 80 training set essays (p < .01). It was also able to significantly the predict group membership of 28 of the 40 test set essays (p < .05), which indicated that the accuracy of the model was 67%. The three predictors were then used as variables in a stepwise multiple regression to determine whether they could predict the original essay scores (before they were divided into the dichotomous variables of high and low quality). The model was able to significantly predict original essays scores, and accounted for 22% of the variance. This study suggested that although cohesion has been shown to be an artifact of readability, it may not be an indicator of perceived writing quality. In addition, this study suggested that writers who produce texts perceived to be of higher quality use more complex language. Finally, this study indicated that Coh-Metrix 2.0, which is available for use over the internet, is able to analyze a large number of texts automatically and includes indices which can significantly discriminate between texts.

Building on McNamara, Crossley, and McCarthy (2010), the present study investigated the following research questions:

1.) Can the various text analysis functions of Coh-Metrix be used to create a model that can distinguish between high and low quality essays in the L2 context?

2.) If a model can be created, which objective measures are the strongest predictors of writing quality?

3.) Are objective predictors of writing quality constant across educational settings?

CHAPTER 3

This chapter outlines the methods used in each of the two experiments conducted. Included is information concerning the two corpora used, the holistic assessment, the objective assessment, and the analysis of the data. Aside from inter-rater reliability figures, all data will be absent from this chapter but provided in chapter four.

Corpus

Study 1

To investigate the first two research questions, I used a subset of the fledgling Intensive English Program Learner English Corpus (IEPLEC) that comprised of problem-solution essays written by intermediate-level students at an intensive English program at a state university (n = 63). The essays were written for a writing class that focused on transitioning from paragraph writing to essay writing. The problem-solution essay is the second type of essay taught in this class and is comprised of three drafts. In order to ensure a range of essay quality scores, the chosen essays came from all stages of the drafting process, though no essays from the same students were used. Although this approach may not meet rigorous standards for essay selection, approximately 70% of all final drafts are rated as passing (a score of 80%) by writing teachers at the particular IEP in which the essays were collected, preliminarily indicating that final drafts are quite homogeneous with regard to quality. The topic for all of these essays was “Write about one problem you had when you arrived in the United States and explain three solutions for that problem.” Problems identified by the writers included “homesickness,” “strange food,” and “boredom,” among others. Sample essays can be found in the appendix.

Study 2

To answer the third question, I used a subset of the International Corpus of Learner English, version 2 (ICLE) (Granger et al., 2009). The ICLE is a collection of essays that are mostly argumentative in nature that were collected from sixteen L1 groups. The essays were written on a number of topics, ranging from ‘Most university degrees are theoretical and do not prepare students for the real world. They are therefore of little value’ to ‘Poverty is the cause of the HIV/AIDS epidemic in Africa’. The most common essay topic was ‘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?’ (n = 491). The most frequent number of essays written on this topic were by Bulgarian L1 writers attending the same university (n = 147). I chose this subset of the ICLE to conduct my second experiment, but limited the essays used to those ranging from 400-700 words in length, which approximately matches the length of the essays used in the first experiment. This resulted in a data set of approximately the same number of essays (n = 64) as used in the first experiment (n = 63). According to the information provided by Granger et al. (2009), the essays were untimed and were not written as part of an examination. Sample essays can be found in the appendix.

Holistic Essay Assessment

Study 1

The essays were scored using a modified version TOEFL iBT independent writing task rubric which consisted of a scale from zero to five (half-points were allowed,) by two experienced writing instructors. The rubric, which can be found in appendix C, was chosen because it focuses on general writing quality and has been used in high-stakes testing to evaluate argumentative essays for some time. The first rater had one year of EFL writing instruction experience and approximately two years of ESL writing instruction experience. The second rater taught TOEFL writing for two and a half years in an EFL context and also taught ESL writing in the United States. Before the raters began scoring essays independently, a training session was conducted. The training session included an in-depth discussion of the essay topic and the rubric. The raters then scored one essay and discussed the rational behind their score, eventually coming to an agreement on the assigned score. The raters did this with three subsequent essays, by which time their scores agreed without discussion. Overall, inter-rater reliability was acceptable (α = .799).

The raters identified three essays that were very similar to others (potential plagiarism), and were therefore thrown out. Where holistic scores did not match, scores were averaged. The essays were then divided into two quality groups based on the mean scores. Approximately one-third of the essays received scored that were within one-quarter point of the mean. In order to create two distinct quality groups, these essays were excluded from the final analysis. Descriptive data can be found in Chapter 4.

Study 2

After I selected the texts, two raters scored the essays using the same TOEFL iBT independent rubric used in the first experiment. The first rater had two years of experience teaching college-level composition; one year with native-speaking students and another year with ESL students. The second rater was the same as the second rater from the first experiment. In order to maintain the consistency of the scores between raters and between the two experiments, a training session was conducted.

During the training session the rubric was explained, and then an essay from the first experiment was read and scored by each rater. The raters then shared their score, discussed their reasons for giving that score, and after a short discussion agreed on a score. During this discussion I acted as a mediator and ensured that their scores agreed with the scores assigned during the first experiment. This process was followed for two more essays from the first experiment, and then for four essays from the second experiment, at which time the raters felt as though they were scoring essays in a similar manner. During the scoring process raters had the opportunity to discuss any scores they were not sure about, which occurred on four occasions. Inter-rater reliability was very high ($\alpha = .915$).

The raters identified four essays that were not written on the appropriate topic, which were subsequently excluded from further analysis. Where holistic scores did not match, the rater scores were averaged. The essays were then separated into high and low quality groups based on their essay scores. Essays that received a score within one-quarter point of the mean score were discarded in order to ensure that the two groups were distinct. Descriptive data can be found in chapter 4.

Objective Assessment

Both corpora were analyzed by the 54 measures available in the online version of Coh-Metrix (Graesser et al. 2004). A full list of the measures included in the public-access Coh-Metrix 2.0, which was used in this study, can be found in appendix A. Below is a description of the measures identified by the selection process outlined later in this chapter. These descriptions have been adapted from the Coh-Metrix website (‘Coh-Metrix Demo,’ 2006)

Readability/Basic Counts

READNW

This is a simple word count for each essay.

READFRE

This measures Flesch Reading Ease, which provides an output of 1-100. The formula for Flesch Reading is is: READFRE= 206.835- (1.015 x average sentence length) – (84.6 x average number of syllables per word)

Latent semantic analysis

LSApssa

This index measures how semantically similar the words in each sentence are compared with all of the other sentences in the essay. The output includes the mean cosine of all of these combinations. (See chapter 2 for more information regarding latent semantic analysis)

LSAppa

This index measures how semantically similar the words in each paragraph are compared with the words in other paragraphs. The output includes the mean cosine of all of these combinations.

Connectives

CONLGpi

This index measures the incidence of positive logical connectives.

Coreference

CREFC1u

This index measures the proportion of content words that occur in adjacent sentences.

Syntax

SYNNP

This index measures syntactic complexity by calculating the mean number of modifiers per noun-phrase.

STRUTa

This index measures syntactic similarity by calculating the proportion of intersecting syntactic tree nodes between all adjacent sentences.

Word Characteristics

WORDCacw

This index measures the mean concreteness value for all words in a text that are included in the MRC Psycholinguistics Database (Coltheart, 1981). Values are based on human-evaluated concreteness scores. Words with a higher value are more concrete, while words with lower values are more abstract.

SPATC

This index calculates the mean of location and motion ratio scores. Location ratio scores (LCR) are calculated by dividing the number of location prepositions (LP) by the number of location prepositions plus the incidence of location nouns (LN). The motion ratio scores are calculated by dividing the number of motion prepositions (MP) by the number of motion verbs (MV). So, SPATC is calculated using the following equation: SPATC = (LP/(LP+LN)) / (MP/MV).

Analysis of Data

According to Leech and Barret (2008), three statistical methods- multiple regression, logistic regression, and discriminant function analysis can be particularly helpful when one wants to predict a dependent variable, such as essay quality, using a number of predictor (independent) variables, such as linguistic objective measures. Multiple regression is used when the dependent variable is an interval variable, such holistic scores ranging from 1 to 5. As the present study investigated a dichotomous variable (high and low quality), multiple regression was not appropriate. Logistic regression can be used with a categorical or dichotomous variable, but many suggest that there should be 20 cases for each predictor variable. As the logistic regression would limit the number of predictor variables to around 2 (IEPLEC = 38, ICLE = 37), I deemed it inappropriate for this study. Discriminate function analysis, then, a statistic that has been shown to be a useful statistic for discriminating between texts (e.g. Crossley et al., 2009, Hall et al., 2007; McNamara et al. 2010) was conducted to determine whether Coh-Metrix variables could accurately discriminate between essays perceived by human raters to be high and low quality. According to Leech and Barret (2008), the assumptions of discriminant function analysis are that ‘the relationships between all pairs of predictors must be linear, multi-variate normality must exist within groups, and the population covariance matrices for predictor variables must be equal across groups” (p. 114). Additionally, multicollinearity can be a problem with discriminant function analysis. If any of the variables are highly correlated with another variable in discriminate function analysis, it may lead to misleading results. (Leech and Barret, 2008).

As Coh-Metrix can process 54 indices and the accuracy of discriminate function analysis is also sensitive to the number of predictors used (If too many are used, ‘overfitting’ occurs. Conservative studies, such as Crossley et al. (2009) and McNamara et al. 2010 use one predictor for every 20 cases, however, the present study used a less conservative ratio of one predictor for ever 10 cases (three predictors) due to the small sample sizes.) the indices underwent a selection process partially outlined in Crossley et al. (2009) and McNamara et al. (2010). First, an analysis of variance (ANOVA) was conducted using Coh-Metrix measures as the dependent variables and high and low essay quality as the fixed factor. Second, the variable with the highest significant F-ratio from each Coh-Metrix category was preliminarily chosen as a predictor in the discriminate function analysis.

Next, I checked the other assumptions of discriminant function analysis. I conducted a Pearson correlation with the preliminary predictors as variables to check for multicollinearity. If any of the sets of preliminary predictors were correlated at the .35 level, the predictor with the larger F-ratio remained a predictor variable, while the predictor with the F-ratio was removed from the group of predictors. After a set of predictors that had no significantly correlated pairs was collected, the number of predictors was counted. If there were more than three predictors, the three predictors with the F-ratios remained.

After the predictors had been selected, scatterplots were created to test whether a linear relationship existed between the predictor variables and covariance matrices were equal across groups. If any variables appeared to violate these assumptions, they were replaced with the predictor with the next largest effect size and assumptions were tested again.

Finally, the assumption of multivariate normality was checked by checking the mean, median, mode, standard deviation and skewness of each variable.

Once all predictors had been tested for the assumptions of discriminate function analysis, a discriminate function analysis was conducted to determine whether the three predictor variables in each model could predict high and low quality group membership.

This process was followed in study one with the IEPLEC corpus. The sets of predictors identified in this corpus were then tested in study two on the ICLE corpus. Finally, the aforementioned selection process was conducted in order to determine whether a set of predictors could be found that discriminated between high and low quality essays better than the IEPLEC sets could.

These studies were conducted to answer the following research questions:

1.) Can the various text analysis functions of Coh-Metrix be used to create a model that can distinguish between high and low quality essays in the L2 context?

2.) If a model can be created, which objective measures are the strongest predictors of writing quality?

3.) Are objective predictors of writing quality constant across educational settings?

CHAPTER 4

This chapter includes the results of the statistical analyses I conducted. Any information not included in this chapter, such as predictor descriptives, correlation matrices, and scatterplot matrices can be found in appendix C. I first give the descriptive statistics for the first study, followed by results of the discriminant function analyses. I then provide descriptive statistics for the second study, followed by the results of the discriminant function analyses. A discussion of the results can be found in chapter five.

Study 1

Descriptives

Descriptive data for the IEPLEC essays concerning holistic scores, number of words per essay and number of sentences per essay can be found in table 4.1. The mean holistic score given was 3.43, while the median score was 3.5. The mean number of words per essay was 559.46, while the mode was 544. The mean number of sentences per essay was 35.98, while the mode was 34.

Table 4.1

Descriptive Statistics of IEPLEC Essays

	Essay Score	Number of Words per Essay	Number of Sentences per Essay
Mean	3.43	559.46	35.98
Median	3.50	544.00	34.00
Mode	4.00	429.00	28.00
Standard Deviation	.59	136.97	11.20
Skewness	-.47	.561	1.29

Discriminant Function Analysis

As the results of the multicollinearity test indicated that the variables with the largest F-ratios, number of words per text (READNW) and latent semantic analysis for all sentences (LSApssa), were significantly correlated, two sets of predictor variables were chosen for discriminant function analysis. The first set included READNW, incidence of positive logical connectives (CONLGpi), and the proportion of content words that overlap between adjacent sentences (CREFC1u). The second set included LSApssa, CONLGpi, and adjacent sentence syntax similarity (STRUTa). Descriptive data for these predictor variables as well as correlation matrices can be found in appendix C.

A discriminate function analysis was first conducted to determine whether the predictors

READNW, CONLGpi, and CREFC1u could accurately distinguish between high and low quality IEPLEC essays. The assumptions that the relationships between all pairs of predictors must be linear, multivariate normality must exist within groups, and the population covariance matrices must be equal across groups were checked and met. Scatterplots can be found in appendix C, and Box's M, which checks the assumption of equal population covariance matrices was not significant (p = .766), indicating that my visual assessment was correct. Wilks' lambda was significant, $\lambda = .588$, $\chi^2 = 18.337$, $p < .000$, partial $\eta^2 = .16$ (small effect size according to Cohen, 1988) which indicates that the model including these three variables was able to significantly discriminate between the two groups. Table 4.2 presents the standardized function coefficients, which suggests that READNW contributes the most to distinguishing between high and low quality essays, using these predictors. Table 4.3 displays the classification results which show that the model correctly predicts 78.9% of low quality essays and 78.9% of high quality essays. Overall, the model correctly predicted the group membership of 79.8% of the essays. The correlation coefficients in the table indicate the extent to which each variable correlates with the resulting discriminant function.

Table 4.2

IEPLEC Standardized Function Coefficients and Correlation Coefficients

	Standardized function coefficients	Correlations between variables and discriminant function
READNW	-.73	-.653
CONLGpi	.52	.48
CREFC1u	.50	.56

Table 4.3

IEPLEC Classification Results: Predicted Group Membership

		Correctly Predicted	Incorrectly Predicted	Total
Count	Low Quality	15	4	19
	High Quality	15	4	19
	Overall	30	8	38
Percentage	Low Quality	78.9	21.1	100.0
	High Quality	78.9	21.1	100.0
	Overall	78.9	21.1	100.0

A second discriminant analysis was conducted to determine whether the whether the three predictors LSApssa, CONLGpi, and STRUTa could accurately distinguish between high and low quality IEPLEC problem-solution essays. The assumptions that the relationships between all pairs of predictors must be linear, multivariate normality must exist within groups, and the population covariance matrices must be equal across groups were checked and met. Scatterplots can be found in appendix C, and Box's M was not significant (p = .313). Wilks' lambda was significant, $\lambda = .55$, $\chi^2 = 20.42$, $p < .000$, partial $\eta^2 = .18$ (small effect size according to Cohen (1988)) which indicates that the model including these four variables was able to significantly discriminate between low and high quality essays. Table 4.4 presents the standardized function coefficients, which suggest that LSApssa contributes most to distinguishing between low and high quality essays, using these predictors. Table 4.5 displays the classification results, which show that the model correctly predicted 73.7% of low quality essays and 84.2% of high quality essays. Overall, the model correctly predicted the group membership of 78.9% of the essays.

Table 4.4

IEPLEC Standardized Function Coefficients and Correlation Coefficients

	Standardized function coefficients	Correlations between variables and discriminant function
LSApssa	.86	.75
CONLGpi	.30	.44
STRUTa	.56	.39

Table 4.5

IEPLEC Classification Results: Predicted Group Membership

		Correctly Predicted	Incorrectly Predicted	Total
Count	Low Quality	14	5	19
	High Quality	16	3	19
	Overall	30	8	38
Percentage	Low Quality	73.7	26.3	100.0
	High Quality	84.2	15.8	100.0
	Overall	78.9	11.1	100.0

Study 2

Descriptive data for the ICLE essays concerning holistic scores, number of words per essay and number of sentences per essay can be found in table 4.6. The mean essay score was 3.26 with a standard deviation of .79. The mean number of words per essay was 517.94 with a standard deviation of 76.60. The mean number of sentences per essay was 28.73 with a standard deviation of 8.06.

Table 4.6

Descriptive Statistics of ICLE Essays

	Essay Score	Number of Words per Essay	Number of Sentences per Essay
Mean	3.26	517.94	28.73
Median	3.25	502	27.50
Mode	3.5	552	27
Standard Deviation	.79	76.60	8.06
Skewness	.382	.732	.794

Discriminant function analysis was conducted using the two sets of successful predictors identified in study one to determine whether either could accurately discriminate between ICLE argumentative essays. Descriptive data and correlation matrices can be found in appendix C.

Discriminant function analysis was first conducted using the three predictors READNW, CONLGpi, and CREFC1u. Scatterplots, which can be found in the appendix, were created to check the assumption of variance-covariance across groups. As the scatterplot matrices, which can be found in appendix C, appeared to be similar across the groups in all matrices, I proceeded with the analysis. Box's M, a statistic that checks the homogeneity of the covariance matrices was not significant (p = .107), indicating that my visual assessment of the scatterplot matrices was correct. Wilks' Lambda not significant λ = .87, χ^2 = 4.56, p = .207, partial η^2 = .04, indicating that the model using the three predictors was not able to significantly discriminant between high and low quality ICLE argumentative essays. Table 4.7 presents the standardized function co-efficients, which show how each predictor contributed to the model. Using the three predictors, READNW contributed the most to discriminating between high and low quality essays, while CREFC1u contributed very little. The classification results, which can be found in table 4.8, indicate that 75.0% of low quality essays were correctly grouped, while 70.6% of high quality essays were correctly grouped. Overall, the model accurately predicted the group membership of 73.0% of the essays.

Table 4.7

ICLE Standardized Function Coefficients and Correlation Coefficients

	Standardized function coefficients	Correlations between variables and discriminant function
READNW	.97	.85
CONLGpi	-.54	-.29
CREFC1u	-.06	-.20

Table 4.8

ICLE Classification Results: Predicted Group Membership

		Correctly Predicted	Incorrectly Predicted	Total
Count	Low Quality	15	5	20
	High Quality	12	5	17
	Overall	27	10	37
Percentage	Low Quality	75.0	25.0	100.0
	High Quality	70.6	29.4	100.0
	Overall	73.0	27.0	100.00

Discriminant function analysis was then conducted using the three predictors from the second IEPLEC model, LSApssa, CONLGpi, and STRUTa. Scatterplots, which can be found in the appendix, were created to check the assumption of variance-covariance across groups. The scatterplot matrices, which can be found in appendix C, appeared to vary somewhat across the groups in some of the matrices, indicating that the the homogeneity of variance-covariance may have been violated. In addition, some of the matrices appeared to violate the assumption of a linear relationship between all variables. Nonetheless, I continued with the analysis as discriminant function analysis is 'fairly robust to these assumptions' (Leech and Barret, 2008; p.114). Box's M, a statistic that checks the homogeneity of the covariance matrices was not significant (p = .817), indicating that my visual assessment of the scatterplot matrices may have been incorrect. Wilks' Lambda was not significant $\lambda = .860$, $\chi^2 = 5.07$, p = .167, partial $\eta^2 =$ .05, indicating that the model using the three predictors was not able to significantly discriminant between the groups. Table 4.9 presents the standardized function co-efficients, which show how each predictor contributed to the model. Using the three predictors, LSApssa to a large degree to discriminating between high and low quality essays, while CONLGpi and STRUTa did not. The classification results, which can be found in table 4.10, indicate that 60.0% of low quality essays were correctly grouped, while 52.9% of high quality essays were correctly grouped. Overall, the model accurately predicted the group membership of 56.8% of the essays.

Table 4.9

ICLE Standardized Function Coefficients and Correlation Coefficients

	Standardized function coefficients	Correlations between variables and discriminant function
LSApssa	.97	.99
CONLGpi	-.17	-.28
STRUTa	-.001	-.07

Table 4.10

ICLE Classification Results: Predicted Group Membership

		Correctly Predicted	Incorrectly Predicted	Total
Count	Low Quality	12	8	20
	High Quality	9	8	17
	Overall	21	16	37
Percentage	Low Quality	60.0	40.0	100.0
	High Quality	52.9	47.1	100.0
	Overall	56.8	43.2	100.00

I then followed the predictor selection variable process used in study one to determine whether it would yield more accurate predictors of writing quality for ICLE argumentative essays than those found for the IEPLEC problem-solution essays. As in the first study, a Pearson correlation, which was run to prevent multicollinearity indicated that the two potential predictor variables with the largest effect sizes correlated significantly with each other. Because two of the predictor variables with the largest F-ratios significantly correlated with each other, two sets of predictor variables were chosen for use in the discriminate function analysis. The first set of predictors included Flesch-Kinkaid reading ease (READFRE), the mean concreteness value for all content words (WORDCacw), and the mean of location and motion ratio scores (SPATC). The second set of predictors included the mean latent semantic analysis values between paragraphs (LSAppa), the mean number of modifiers per noun phrase (SYNNP), and WORDCacw. Descriptives and correlation matrices can be found in appendix C.

Discriminant function analysis was first conducted to determine whether the three predictors READFRE, WORDCacw, and SPATc could accurately distinguish between high and low quality ICLE argumentative essays. The assumption that the relationships between all pairs of predictors must be linear, multivariate normality must exist and the population covariance matrices for predictor variables must be equal across groups was checked and met. Scatterplots can be found in appendix C, and Box’s M was not significant (p = .375). Wilks’ lambda was significant, $\lambda = .64$, $\chi^2 = 15.05$, p = .002, partial $\eta^2 = .14$ (small according to Cohen (1988)), indicating that the model using the three predictors significantly discriminated between the high and low quality essays. Table 4.11 presents the standardized function coefficients, which suggest that READFRE contributed most to distinguishing between low and high quality essays, using these predictors. Table 4.12 displays the classification results, which show that the model correctly predicted group membership of 80% of low quality essays and 82.4% of high quality essays. Overall, the model accurately predicted group membership of 81.1% of the essays.

Table 4.11

ICLE Standardized Function Coefficients and Correlation Coefficients

	Standardized function coefficients	Correlations between variables and discriminant function
READFRE	.62	.74
WORDCacw	.42	.54
SPATC	-.52	-.63

Table 4.12

ICLE Classification Results: Predicted Group Membership

		Correctly Predicted	Incorrectly Predicted	Total
Count	Low Quality	16	4	20
	High Quality	14	3	17
	Overall	30	7	37
Percentage	Low Quality	80.0	20.0	100.0
	High Quality	82.4	17.6	100.0
	Overall	81.1	18.9	100.00

Discriminant function analysis was then conducted to determine whether the three predictors LSAppa, WORDCacw, and SYNNP could accurately distinguish between high and low quality ICLE argumentative essays. The assumption that the relationships between all pairs of predictors must be linear, multivariate normality must exist and the population covariance matrices for predictor variables must be equal across groups was checked and met. Scatterplots can be found in appendix C, and Box's M was not significant (p = .273). Wilks' lambda was significant, $\lambda = .61$, $\chi^2 = 16.52$, p = .001, partial $\eta^2 = .15$ (small according to Cohen (1988)), indicating that the model using the three predictors significantly discriminated between the high and low quality essays. Table 4.13 presents the standardized function coefficients, which suggest that SYNNP contributed most to distinguishing between low and high quality essays, using these predictors. Table 4.14 displays the classification results, which show that the model correctly predicted group membership of 85% of low quality essays and 88.2% of high quality essays. Overall, the model accurately predicted group membership of 86.5% of the essays.

Table 4.13

ICLE Standardized Function Coefficients and Correlation Coefficients

	Standardized function coefficients	Correlations between variables and discriminant function
LSAppa	.58	.66
WORDCacw	-.49	-.51
SYNNP	.61	.61

Table 4.14

ICLE Classification Results: Predicted Group Membership

		Correctly Predicted	Incorrectly Predicted	Total
Count	Low Quality	17	3	20
	High Quality	15	2	17
	Overall	32	5	37
Percentage	Low Quality	85.0	15.0	100.0
	High Quality	88.2	11.8	100.0
	Overall	86.5	13.5	100.00

CHAPTER 5

This chapter interprets the results found in chapter four with a focus on answering the questions that guided this study, including an objective measure profile of essays holistically evaluated to be high quality. This chapter also describes how the current study is also situated within the context of relevant scholarship. Pedagogical implications, and directions for future research are discussed, as well as the limitations of this study.

Answers to Research Questions

The present study was guided by the following research questions:

1.) Can the various text analysis functions of Coh-Metrix be used to create a model that can distinguish between high and low quality essays in the L2 context?

2.) If a model can be created, which objective measures are the strongest predictors of writing quality?

3.) Are objective predictors of writing quality constant across educational settings?

Research Question One

In this study, I used four models using only three predictors each that were able to significantly discriminate between L2 essays that were holistically evaluated to be of high and low quality. This indicates that usefulness of Coh-Metrix 2.0 in identifying textual features of high and low quality writing in the L2 context.

The first Intensive English Learner English Corpus (IEPLEC) predictor set, which took into account the number of words per essay (READNW), the number of positive logical connectives per essay (CONLGpi), and the proportion of content words that occurred in adjacent sentences in each essay (CREFC1u), was able to predict whether an essay was of high or low quality with an accuracy of 79%. The second IEPLEC predictor set, which evaluated the semantic relatedness between all combinations of sentences in each essay using latent semantic analysis (LSApssa), the incidence of positive logical connectives (CONLGpi), and the syntactic similarity of adjacent sentences in each essay (STRUTa), was also able to predict high and low quality group membership with an accuracy of 79%.

The first International Corpus of Learner English (ICLE) predictor set, which took into account Flesch Reading Ease scores of each essay (READFRE), the concreteness of the words used in each essay (WORDCacw), and the mean of location and motion ration scores for each essay (SPATC), was able to predict the quality of essays with an accuracy of 81%. The second predictor set, which proved to be the most accurate tested, included the semantic relatedness of the words in each paragraph in each essay using latent semantic analysis (LSAppa), the concreteness of all words in each essay (WORDCacw), and the mean number of modifiers per noun-phrase (SYNNP) was able to predict whether essays were holistically evaluated to be of high or low quality with an accuracy of 87%. See chapter 4 for a full description of the results, including the relative strength of each predictor in each model.

Research Question Two

The strongest predictors of writing quality for the IEPLEC essays were the number of words per essay (READNW), the semantic relatedness of all combinations of sentences in each essay as measured by latent semantic analysis (LSApssa), the incidence of positive logical connectives CONLGpi), the syntactic similarity between adjacent sentences STRUTa, and the proportion of content words that occur in adjacent sentences in each essay (CREFC1u).

Within the first IEPLEC predictor set (READNW, CONLGpi, and CREFC1u), the number of words per essay was the strongest predictor of writing quality (READNW), although the incidence of positive logical connectives (CONLGpi) and the proportion of words used in adjacent sentences (CREFC1u) also contributed substantially to the model.

Within the second IEPLEC predictor set (LSApssa, CONLGpi, and STRUTa), the semantic relatedness between all combinations of sentences as measured by latent semantic analysis (LSApssa) was the strongest predictor of writing quality. The syntactic similarity between adjacent sentences also was a strong predictor (STRUTa) while the incidence of positive logical connectives (CONLGpi) played a lesser role.

The strongest predictors of writing quality for the ICLE essays were Flesch Reading Ease scores (READFRE), the semantic relatedness of words of all paragraphs in each essay (LSAppa), the concreteness of words in each essay (WORDCacw), the location and motion ratio scores (SPATC), and the number of modifiers per noun phrase (SYNNP).

Within the first predictor set, Flesch Reading Ease (READFRE) was the strongest predictor of writing quality, followed by location and motion ratio scores (SPATC), and finally by the concreteness of words in each essay (WORDCacw).

Within the second predictor set, the number of modifiers per noun phrase (SYNNP) and the semantic relatedness of words in all paragraphs in each essay

(LSAppa) were the strongest predictors of writing quality (with SYNNP being slightly stronger). The concreteness of words (WORDCacw) was also strong predictor of essay quality, though not as strong as SYNNP and LSAppa. See chapter 4 for a full description of the results, including the relative strength of each predictor in each model.

Research Question Three

Neither of the models that significantly discriminated between high and low quality IEPLEC essays (which were written by students learning English as a second language) were able to discriminate between high and low quality ICLE essays (which were written by students learning English as a foreign language). The first IEPLEC predictor set (READNW, CONLGpi, and CREFC1u) was able to accurately predict the relative quality of 73% of ICLE essays, while the second IEPLEC predictor set (LSApssa, CONLGpi, and STRUTa) was only able to accurately predict the relative quality of 57% of ICLE essays. Interestingly, although the mean holistic scores of ESL and EFL essays were quite similar (3.43 and 3.26), they were quite different with regard to textual features identified in previous studies (eg. McNamara et al., 2010) to contribute to essay quality, such as lexical diversity. The results of a one-way ANOVA revealed that there were significant differences between the ESL and EFL essays on 43 of the 54 indices that Coh-Metrix measures (all but three indices (which were significant at the $p < .01$ level) were significant at the $p < .01$ level). In fact, a model using lexical frequency as the sole predictor was able to predict whether an essay was written in the ESL or the EFL context with an accuracy of 96% using discriminant function analysis.

Despite these differences, however, a model using four predictors, the number of words per essay, the average number of syllables per word, the concreteness of words, and the mean of location and ratio scores were able to predict the quality of essays regardless of educational context with an accuracy of 76%, indicating that these textual features may transcend the EFL/EFL boundary. Future research, however, is needed to fully explore the differences between ESL and EFL writing.

Although the specific models that were successful in discriminating IEPLEC essays could not successfully discriminate between ICLE essays, some general trends were observed. Latent semantic analysis, for example, played a large role in discriminating between high and low quality essays from both corpora, although the trends varied (see below). Readability/basic counts also played a large role in discriminating between high and low quality essays from both corpora.

A Profile of High and Low Quality Essays by Predictor

This section outlines the features of high and low quality essays based on the successful predictor models for each corpus. Table 5.1 displays the mean score for each IEPLEC predictor, while table 5.2 displays the mean for each ICLE predictor. Refer to chapter 2 for an explanation of each predictor.

IEPLEC essays

Table 5.1

Mean predictor values for IEPLEC high and low quality essays

Predictor	High Quality Essay Mean Value	Low Quality Essay Mean Value
READNW Number of Words	642.21	495.42
CONLGpi Incidence of positive logical connectives	36.14	45.95
CREFC1u Proportion of content words that overlap between adjacent sentences	.17	.21
LSApssa LSA, sentences, all combinations, mean	.22	.30
STRUTa Sentence syntax similarity, adjacent	.10	.30

READNW

High quality essays tended to be longer, with a mean length of 642.21 words, while low quality essays tended to be shorter, with a mean length of 495.42 words.

CONLGpi

High quality essays tended to have fewer positive logical connectives, with a mean of 36.14 logical positive connectives per essay, while lower quality essays tended to contain more positive logical connectives (mean of 45.95). Logical connectives include but are not limited to *or*, *actually*, and *if*.

CREFC1u

High quality essays tended to have fewer words that overlapped between adjacent sentences (mean proportion of .17 overlapping words per sentence) than low quality essays (mean proportion of .21 overlapping words per sentence).

LSApssa

High quality essays tended to have words in each sentence that were less semantically related (a mean cosine of .22 between all sentences) than those of low quality essays (a mean cosine of .30 between all sentences).

STRUTa

High quality essays tended to have less similar adjacent sentence structure (mean proportion of intersecting syntactic tree nodes between adjacent sentences of .10) than low quality essays (mean proportion of intersecting syntactic tree nodes between adjacent sentences of .12).

Overall, these findings indicate essays written in the ESL context tend to be perceived by human raters to be of higher quality if they are longer, employ more diverse syntactic structures, and have fewer cohesive features such as connectives, co-reference, and semantically related lexis.

ICLE essays

Table 5.2

Mean predictor values for ICLE high and low quality essays

Predictor	High Quality Essay Mean Value	Low Quality Essay Mean Value
READFRE Flesch Reading Ease Score (0-100)	50.07	60.27
WORDCacw Concreteness, mean for content words	345.46	357.43
SPATC Mean of location and motion ratio scores	.52	.48
LSAppa LSA, Paragraph to Paragraph, mean	.33	.25
SYNNP Mean number of modifiers per noun-phrase	.80	.75

READFRE

High quality essays tended to have a lower Flesch Reading Ease score (more difficult to comprehend), with a mean score of 50.07, while low quality essays tended to have a higher Flesch Reading Ease score (easier to comprehend), with a mean score of 60.28.

WORDCacw

High quality essays tended to use less concrete words (mean concreteness score of 345.46) than low quality essays (mean concreteness score of 357.43).

SPATC

High quality essays tended to have higher proportion of location scores to motion scores (mean of location and motion ratio scores = .52) than low quality essays (mean of location and motion ration scores = .48).

LSAppa

Paragraphs in high quality essays tended to be more semantically related to each other (mean LSA cosine of .33) than paragraphs in low quality essays (mean LSA cosine of .25).

SYNNP

High quality essays tended to have more modifiers per noun phrase (mean of .86) than low quality essays (mean of .75).

Overall, these findings indicate that EFL essays tend to be perceived by human raters to be of higher quality if they use longer sentences with longer words, use less concrete language, use more complex sentences, and have semantically related words in each paragraph.

How the findings of this study compare with extant research

The present study found that higher quality essays tended to have either more complex or diverse syntax, which generally corresponds with human-rated objective measures studies such as Larsen-Freeman (1978), Flahive and Snow (1980), and Homburg (1984), which found that essays with more complex syntactic structures tended to receive higher holistic scores.

The general findings of the present study also correspond with studies such as McNamara et al. (2010), which found that higher quality L1 essays tended to have more complex syntax (measured by the number of words before the main verb), more diverse vocabulary, and less common words. The present study diverges from McNamara et al., however, with regard to cohesion. Using the same Coh-Metrix 2.0 indices of cohesion as the present study, McNamara et al. found no significant differences between high and low quality essays with regard to cohesion, while the current study found that high quality ESL essays tended to be less cohesive when measured by measures of latent semantic analysis, connectives, and co-reference, while EFL high quality essays tended to have more spatial cohesion and were more cohesive as measured by latent semantic analysis.

Although it has been presumed that cohesion is an artifact of coherence (eg. Halliday and Hasan, 1976), and studies have indicated a relationship between cohesive textual features and coherence (eg. Foltz et al., 1998; Leon et al., 2006), the present study seems to indicate that cohesion is not neccesarily an artifact of coherence in ESL argumentative essays. In the ESL context, essays identified by human raters using a holistic scale that emphasized coherence as an artifact of writing quality to be of high quality had fewer cohesive features than those considered to be of low quality. These findings generally support Carrell (1982), who cautioned that teaching students to use cohesive ties in their writing would not necessarily guarantee cohesive texts. In the EFL context, however, essays that were more cohesive tended to earn higher holistic scores. Clearly, the relationship between coherence and cohesion, especially in various L2 contexts is an area of research that needs to be explored further.

One issue in identifying a set of linguistic predictors of writing quality lies in the possibility that the construct of writing quality may be more complex than researchers have assumed. Although most objective measure studies that have been conducted including the present one have tried to find a linear relationship between various predictors and writing quality, Jarvis, Grant, Bikowski, and Ferris (2003) presented a different way at looking at the issue. Using two sets of highly-rated L2 essays and by measuring 21 textual features, they were able to identify a number of writing profiles that characterized groups of highly proficient writers. They suggested that writers were able to use a set linguistic features in a complimentary way and/or are able to compensate for linguistic weakness with linguistic strengths in order to produce an effective text, but that the linguistic features used may vary from writer to writer. This phenomenon could explain the difficulty in identifying a global set of linguistic features that would describe quality writing and deserves the attention of future research.

Pedagogical Implications

This study supports the idea that an increase in syntactic diversity positively affects the perceived quality of argumentative writing in the ESL context, suggesting that sentence structure instruction is valuable in the ESL writing classrooms. It also, however, challenges the idea that teaching cohesive textual features will necessarily increase the quality of student writing. Future research should focus on the relationship between explicit instruction of cohesive ties and writing quality in the L2 context. Furthermore, future research should investigate the relationship that cohesive textual features have with other writer-focused factors, such as language proficiency and writing experience.

This study also supports the idea that complex language, both with regard to syntax and vocabulary positively affects perceived writing quality in the EFL context, suggesting that sentence-level and word-level instruction is valuable in EFL classrooms. It also indicates that EFL essays that include semantically related words in each paragraph are perceived to be of higher quality than those that have words that are less semantically related. In addition, these findings indicate that essays that exhibited more spatial cohesion were perceived to be of higher quality. Little, if any previous research has been conducted on the relationship between spatial cohesion and writing, opening possibilities for future research.

Finally, this study suggests that objective measures of essay quality are not generalizable across learning contexts. The EFL argumentative essays analyzed in this study had more lexical diversity and less common lexical items than the ESL argumentative essays, and yet received a mean holistic quality score than the ESL essays. Further research is needed to explore how educational context affects different aspects of writing quality, but it is clear that lexis is only one piece of the writing quality pie that needs to be evaluated within the context of other variables.

Limitations

As with all studies, the present one has some limitations. The sample texts are relatively low in number, raising the question of generalizability. In addition, although previous studies using Coh-Metrix have indicated that writing prompt does not significantly affect the analysis of textual features (McNamara et al. 2010), it is unclear whether this is generalizable to the L2 context, potentially making the relationship between writing in the ESL and EFL contexts unclear. Furthermore, many early objective measure studies (e.g. Larsen-Freeman, 1978; Perkins, 1980; Homburg, 1980) found that error played a role in the perceived quality of essays written by English-learners. Coh-Metrix 2.0 is not able to take errors into account, potentially obscuring the construct of writing quality.

In addition, essays collected for the IEPLEC corpus were taken from various steps in the drafting process, while ICLE essays were taken from a single draft, further bringing in to question the comparison made between the two essay types. In addition, the inherent nature of English for Academic Purposes (EAP) courses in an ESL setting is problematic, as the length of stay in the second-language country and the length of time taking classes at an IEP can vary greatly from student to student. My corpus collection procedure did not take into account how ‘ESL’ our ESL students were. It is possible that some of the student writers whose writings were collected in the ESL corpus had very little experiential and educational background differences from the student writers whose writings were collected for the EFL corpus. At the same time, their differences may have been great. Future research should focus on studying two writing corpora that are exhaustively comparable except with regard to educational setting.

Conclusion

This study has demonstrated the usefulness of Coh-Metrix in examining the construct of writing quality. Coh-Metrix indices were used to create a number of models that were able to significantly predict whether L2 essays were perceived as high or low quality by human raters with an accuracy of up to 86.5%. Although these models were significant predictors of writing quality within a certain educational setting (ESL or EFL), they were not robust across educational settings. The results of this study support the idea that essays containing complex language are generally perceived to be of high quality. The results of this study question, however, the idea that coherent texts have more cohesive textual features, at least in the ESL setting. Furthermore, this study highlights the differences between ESL and EFL writing, though these differences may prove to be related to factors other than educational setting due to some limitations of this study. Future research should focus on the differences between ESL and EFL writing. Continuing the work of Jarvis et al. (2003) in multiple profiles of writing quality may also be a worthwhile pursuit and may explain the linguistic differences between ESL and EFL writing.

REFERENCES

Anonymous. (2006). *Coh-Metrix Demo.* Retrieved from http://cohmetrix.memphis.edu/CohMetrixDemo/demo.htm

Attali, Y. & Burnstein, J. (2005). *Automated essay scoring with e-rater® v.2.0 (RR-04-45).* Princeton, NJ: ETS.

Attali, Y. & Burstein, J. (2006). Automated essay scoring with e-rater® v.2.0. *The Journal of Technology, Learning and Assessment, 4(3),* (np).

Arena, Louis A. (1982). The language of corporate attorneys. *Linguistics and the Professions,* Robert J. Dipietro (Ed.), 143-154. Norwood, NJ: Ablex.

Baayen, R. H., Piepenbrock, R., & Gulikers, L. (1995). *The CELEX lexical database*. Philadelphia: University of Pennsylvania.

Bormuth, J. R. (1971). *Development of standards of readability: Toward a rational criterion of passage performance*. U. S. Department of Health, Education and Welfare (ERIC Doc. No. ED O54 233).

Britton, B.K., and Gulgoz, S. (1991). Using Kintsch's computational model to improve instructional text: Effects of repairing inference calls on recall and cognitive structures. *Journal of Educational Psychology, 83*, 329-345.

Brodkey, D., & Young, R. (1981). Composition correctness scores. *TESOL Quarterly, 15* (2), 159-167.

Brown, J. D., & Bailey, K. M. (1984). A categorical instrument for scoring second language writing skills. *Language Learning*, *34*(4), 21-42.

Carrell, P. L. (1982). Cohesion is not coherence. *TESOL Quarterly, 16*(4), 479-488.

Charney, D. (1984). The validity of using holistic scoring to evaluate writing: A critical overview. *Research in the Teaching of English, 18*(1), 65-81.

Chodorow, M. & Burnstein, J. (2004). *Beyond essay length: Evaluating e-rater®'s performance on TOEFL® essays (Report No. 73).* Princeton, NJ: ETS.

Coltheart, M. (1981). The MRC psycholinguistic database quarterly. *Journal of Experimental Psychology, 33A*, 497-505.

Cobb, T. (2010). *Compleat Lexical Tutor*. Retrieved from http://www.lextutor.ca/

Corrigan, R., & Surber, J. R. (2010). The reading level paradox: Why children's picture books are less cohesive than adult books. *Discourse Processes, 47*(1), 32-54.

Crossley, S. A., & McNamara, D. S. (2008). Assessing L2 reading texts at the intermediate level: An approximate replication of crossley, louwerse, McCarthy & McNamara (2007). *Language Teaching, 41*(3), 409-429.

Crossley, S. A., & McNamara, D. S. (2009). Computational assessment of lexical differences in L1 and L2 writing. *Journal of Second Language Writing, 18*(2), 119-135.

Crossley, S. A., Greenfield, J., & McNamara, D. S. (2008). Assessing text readability using cognitively based indices. *TESOL Quarterly*, *42*(3), 475-493.

Crossley, S. A., Louwerse, M. M., McCarthy, P. M., & Mcnamara, D. S. (2007). A linguistic analysis of simplified and authentic texts. *The Modern Language Journal*, *91*(1), 15-30.

Crossley, S. A., Salsbury, T., & McNamara, D. S. (2010). The development of semantic relations in second language speakers: A case for latent semantic analysis. *VIAL - Vigo International Journal of Applied Linguistics, 7*, 55-74.

Cumming, A., Kantor, R., Baba, K., Erdosy, U., Eouanzoui, K., & James, M. (2005). Differences in written discourse in independent and integrated prototype tasks for next generation TOEFL. *Assessing Writing, 10*(1), 5-43.

Greenfield, J. (2004). Readability formulas for EFL. *JALT Journal, 26*, 5–24.

Duran, N. D., Hall, C., McCarthy, P. M., & McNamara, D. S. (2010). The linguistic correlates of conversational deception: Comparing natural language processing technologies. *Applied Psycholinguistics, 31*(3), 439-462.

Endicott, A. L. (1973). A proposed scale for syntactic complexity. *Research in the Teaching of English, 7*(1), 5-12.

Engber, C. A. (1995). The relationship of lexical proficiency to the quality of ESL compositions. *Journal of Second Language Writing, 4*(2), 139-155.

Enright, M. K., & Quinlan, T. (2010). Complementing human judgment of essays written by english language learners with e-rater(R) scoring. *Language Testing, 27*(3), 317-334.

Flahive, E. Douglas, and Becky G. Snow. 1980. Measures of syntactic complexity in evaluating ESL compositions. *In Research in language testing*, John W. Oiler, Jr., and Kyle Perkins (Eds.), 171-176. Rowley, Massachusetts: Newbury House.

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

Foltz, P. W., Kintsch, W., & Landauer, T. K. (1998). The measurement of textual coherence with latent semantic analysis. *Discourse Processes, 25*(2-3), 285-307.

Freedman, S.W. (1979). How characteristics of student essays influence teachers' evaluations. *Journal of Educational Psychology, 71*(3), 328-338.

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.

Granger, S., Dagneaux, E.., Meunier, F., Paquot, M. (Eds.) (2009). *International corpus of learner english. version 2.* Belgium: Presses universitaires de Louvain.

Greenfield, G. (1999). *Classic readability formulas in an EFL context: Are they valid for Japanese speakers?* Unpublished doctoral dissertation, Temple University, Philadelphia, PA, United States. (University Microfilms No. 99–38670).

Halliday, M. A. K., & Hassan, R. (1976). *Cohesion in English*. London: Longman.

Hall, C., Lewis, G. A., McCarthy, P. M., Lee, D. S., & McNamara, D. S. (2007). A Coh-Metrix assessment of American and English/Welsh Legal English. *Coyote Papers: Psycholinguistic and Computational Perspectives. University of Arizona Working Papers in Linguistics, 15*, 40-54.

Hamp-Lyons, L., & Henning, G. (1991). Communicative writing profiles: An investigation of the transferability of a multiple-trait scoring instrument across ESL writing assessment contexts. *Language Learning, 41*(3), 337-373.

Hamp-Lyons, L. (1995). Rating nonnative writing: The trouble with holistic scoring. *TESOL Quarterly, 29*(4), 759-762.

Homburg, T. J. (1984). Holistic evaluation of ESL compositions: Can it be validated objectively? *TESOL Quarterly, 18*, 87-107.

Hunt, K. (1965). *Grammatical structures written at three grade levels.* Champaign, IL: National Council of Teachers of English.

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.

Jorge-Botana, G., Leon, J. A., Olmos, R., & Escudero, I. (2010). Latent semantic analysis parameters for essay evaluation using small-scale corpora. *Journal of Quantitative Linguistics, 17*(1), 1-29.

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

Landauer, T. K., Foltz, P. W., & Laham, D. (1998a). An introduction to latent semantic analysis. *Discourse Processes, 25*(2-3), 259-284.

Landauer, T. K., Foltz, P. W., & Laham, D. (1998b). *Latent semantic analysis passes the test: Knowledge representation and multiple choice testing.* Unpublished manuscript.

Larsen-Freeman, D. (1978). An ESL index of development. *TESOL Quarterly, 12*(4), 439-448.

Laufer, B., & Nation, P. (1995). Vocabulary size and use: Lexical richness in L2 written production. *Applied Linguistics, 16*(3), 307-322.

Lee, Y., Gentile, C., & Kantor, R. (2010). Toward automated multi-trait scoring of essays: Investigating links among holistic, analytic, and text feature scores. *Applied Linguistics (Oxford), 31*(3), 391-417.

Leech, N. L., Barrett, K. C. (2008). *SPSS for intermediate statistics: use and interpretation*. *3rd ed.* New York: L. Erlbaum/Taylor & Francis.

Leon, J. A., Olmos, R., Escudero, I., Canas, J. J., & Salmeron, L. (2006). Assessing short summaries with human judgments procedure and latent semantic analysis in narrative and expository texts. *Behavior Research Methods, 38*(4), 616-627.

Li, Y. (2000). Assessing second language writing: The relationship between computerized analysis and rater evaluation. *ITL, Review of Applied Linguistics, 127-128*(Jan), 37-51.

McCarthy, P. M. (2005). An assessment of the range and usefulness of lexical diversity measures and the potential of the measure of textual, lexical diversity (MTLD). *Dissertation Abstracts International, 66*, 12. (UMI No. 3199485)

McCarthy, P. M., Hall, C., Duran, N. D., Doiuchi, M., Fujiwara, Y., Duncan, B., & McNamara, D. S. (2009). Analyzing journal abstracts written by japanese, american, and british scientists using COH-metrix and the gramulator. *The ESPecialist, 30*(2), 141-173.

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

McNamara, D. S., Louwerse, M. M., McCarthy, P. M., & Graesser, A. C. (2010). Coh-metrix: Capturing linguistic features of cohesion. *Discourse Processes, 47*(4), 292-330.

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

Nas, Gerard. 1975. *Determining the communicative value of written discourse produced by L2 learners*. Utrecht, The Netherlands: Institute of Applied Linguistics.

Ortega, L. (2009). *Understanding second language acquisition*. London: Hodder Education.

Perkins, K. (1980). Using objective methods of attained writing proficiency to discriminate among holistic evaluations. *TESOL Quarterly, 14*(1), 61-69.

Polio, C. G. (1997). Measures of linguistic accuracy in second language writing research. *Language Learning, 47*(1), 101-143.

Sparks, J. (1988). Syntactic complexity, error and the holistic evaluation of ESL student essays. *The ORTESOL Journal, 9,* 35-49.

Yu, G. (2010). Lexical diversity in writing and speaking task performances. *Applied Linguistics (Oxford), 31*(2), 236-259.

APPENDIX

Appendix A

Table A

Coh-Metrix Category Distinctions

1. Coh-Metrix General Categories

Output number	Description	Measure Abbreviation	Full description
1	Title	N/A	N/A
2	Genre	N/A	N/A
3	Source	N/A	N/A
4	JobCode	N/A	N/A
5	LSASpace	N/A	N/A
6	Date	N/A	N/A

2. Readability

59	Flesch Reading Ease	READFRE	Flesch Reading Ease Score (0-100)
60	Flesch-Kincaid	READFKGL	Flesch-Kincaid Grade Level (0-12)

3. General Word and Text Information

3.1 Basic Count

53	No. of words	READNW	Number of Words
54	No. of sentences	READNS	Number of Sentences
55	No. of paragraphs	READNP	Number of Paragraphs
56	Syllables per word	READASW	Average Syllables per Word

57	Words per sentence	READASL	Average Words per Sentence
58	Sentences per paragraph	READAPL	Average Sentences per Paragraph

3.2 Frequencies

40	Raw freq. content words	FRQCRacw	Celex, raw, mean for content words (0-1,000,000)
41	Log freq. content words	FRQCLacw	Celex, logarithm, mean for content words (0-6)
42	Min. raw freq. content words	FRQCRmcs	Celex, raw, minimum in sentence for content words (0-1,000,000)
43	Log min. freq. content words	FRQCLmcs	Celex, logarithm, minimum in sentence for content words (0-6)

3.3 Concreteness

44	Concreteness content words	WORDCacw	Concreteness, mean for content words
45	Min. concreteness content words	WORDCmcs	Concreteness, minimum in sentence for content words

3.4 Hypernymy

46	Noun hypernym	HYNOUNaw	Mean hypernym values of nouns
47	Verb hypernym	HYVERBaw	Mean hypernym values of verbs

4. Syntax Indices

4.1 Constituents

48	Negations	DENNEGi	Number of negations, incidence score
49	NP incidence	DENSNP	Noun Phrase Incidence Score (per thousand words)
50	Modifiers per NP	SYNNP	Mean number of modifiers per noun-phrase
51	Higher level constituents	SYNHw	Mean number of higher level constituents per word
52	Words before main verb	SYNLE	Mean number of words before the main verb of main clause in sentences

4.2 Pronouns, Types, Tokens

17	Personal pronouns	DENPRPi	Personal pronoun incidence score
18	Pronoun ratio	DENSPR2	Ratio of pronouns to noun phrases
19	Type-token ratio	TYPTOKc	Type-token ratio for all content words

4.3 Connectives

29	All connectives	CONi	Incidence of all connectives
30	Conditional operators	DENCONDi	Number of conditional expressions, incidence score
31	Pos. additive connectives	CONADpi	Incidence of positive additive connectives
32	Pos. temporal connectives	CONTPpi	Incidence of positive temporal connectives
33	Pos. causal connectives	CONCSpi	Incidence of positive causal connectives
34	Pos. logical connectives	CONLGpi	Incidence of positive logical connectives
35	Neg. additive connectives	CONADni	Incidence of negative additive connectives
36	Neg. temporal connectives	CONTPni	Incidence of negative temporal connectives
37	Neg. causal connectives	CONCSni	Incidence of negative causal connectives
38	Neg.logical connectives	CONLGni	Incidence of negative logical connectives

4.5 Logical Operators

39	Logic operators	DENLOGi	Logical operator incidence score (and + if + or + cond + neg)

4.6 Sentence Syntax Similarity

24	Syntactic structure similarity adjacent	STRUTa	Sentence syntax similarity, adjacent

25	Syntactic structure similarity all-1	STRUTt	Sentence syntax similarity, all, across paragraphs
26	Syntactic structure similarity all 2	STRUTp	Sentence syntax similarity, sentence all, within paragraphs

5 Referential and Semantic Indices

5.1 Anaphor

7	Adjacent anaphor reference	CREFP1u	Anaphor reference, adjacent, unweighted
8	Anaphor reference	CREFPau	Anaphor reference, all distances, unweighted

5.2 Co-reference

9	Adjacent argument overlap	CREFA1u	Argument Overlap, adjacent, unweighted
10	Argument overlap	CREFAau	Argument Overlap, all distances, unweighted
11	Adjacent stem overlap	CREFS1u	Stem Overlap, adjacent, unweighted
12	Stem overlap	CREFSau	Stem Overlap, all distances, unweighted
13	Content word overlap	CREFC1u	Proportion of content words that overlap between adjacent sentences

5.3 LSA

14	LSA sentence adjacent	LSAassa	LSA, Sentence to Sentence, adjacent, mean
15	LSA sentence all	LSApssa	LSA, sentences, all combinations, mean
16	LSA paragraph	LSAppa	LSA, Paragraph to Paragraph, mean

6. Situational Model Dimensions

6.1 Causal Dimension

20	Causal content	CAUSVP	Incidence of causal verbs, links, and particles

21	Causal cohesion	CAUSC	Ratio of causal particles to causal verbs (cp divided by cv+1)

6.2 Intentional Dimension

22	Intentional content	INTEi	Incidence of intentional actions, events, and particles.
23	Intentional cohesion	INTEC	Ratio of intentional particles to intentional content

6.3 Temporal Cohesion

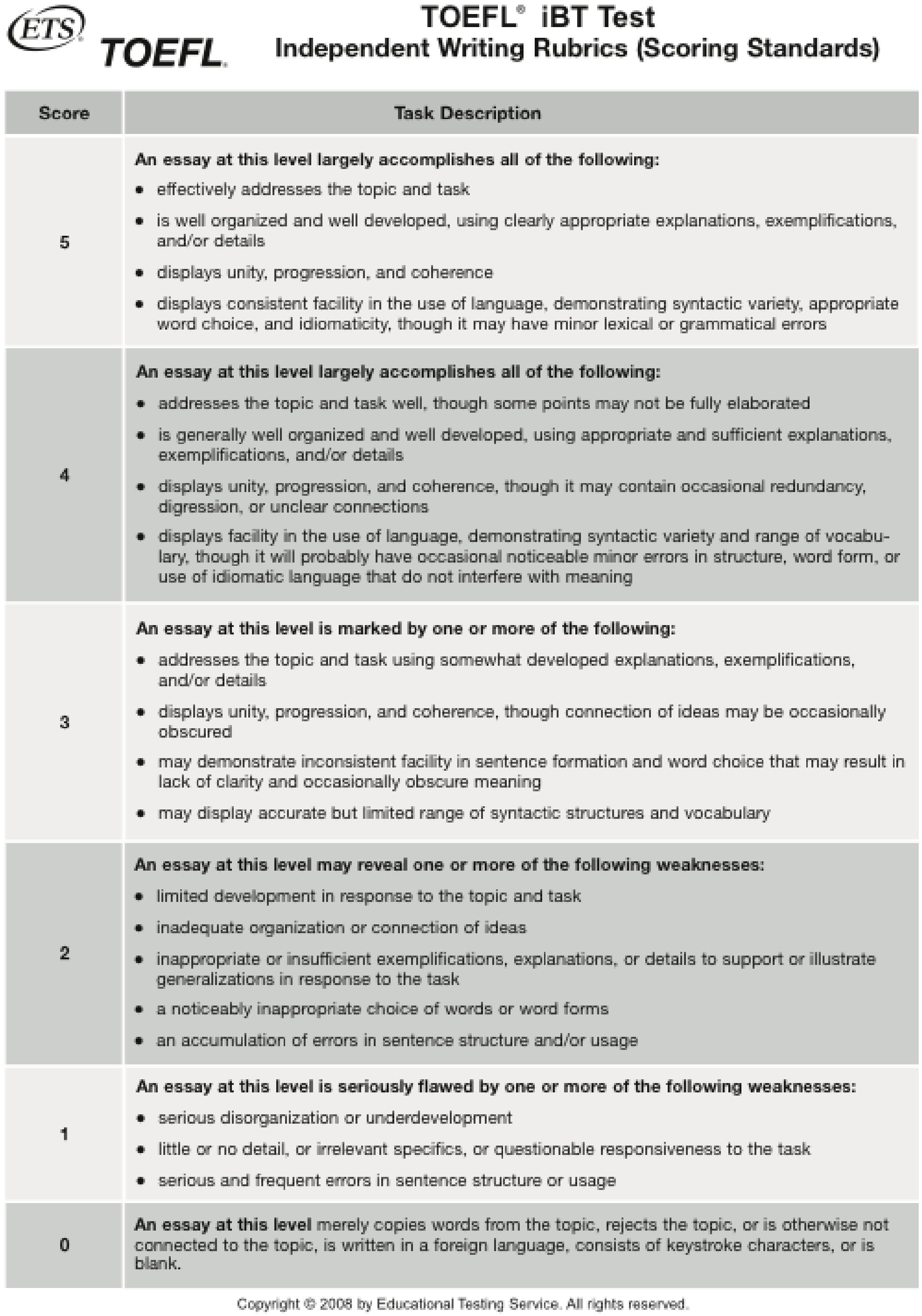
27	Temporal cohesion	TEMPta	Mean of tense and aspect repetition scores

6.4 Spatial Cohesion

28	Spatial cohesion	SPATC	Mean of location and motion ratio scores.

Appendix B

Appendix B1: TOEFL iBT Rubric



ETS TOEFL

TOEFL® iBT Test
Independent Writing Rubrics (Scoring Standards)

Score	Task Description
5	**An essay at this level largely accomplishes all of the following:** • effectively addresses the topic and task • is well organized and well developed, using clearly appropriate explanations, exemplifications, and/or details • displays unity, progression, and coherence • displays consistent facility in the use of language, demonstrating syntactic variety, appropriate word choice, and idiomaticity, though it may have minor lexical or grammatical errors
4	**An essay at this level largely accomplishes all of the following:** • addresses the topic and task well, though some points may not be fully elaborated • is generally well organized and well developed, using appropriate and sufficient explanations, exemplifications, and/or details • displays unity, progression, and coherence, though it may contain occasional redundancy, digression, or unclear connections • displays facility in the use of language, demonstrating syntactic variety and range of vocabulary, though it will probably have occasional noticeable minor errors in structure, word form, or use of idiomatic language that do not interfere with meaning
3	**An essay at this level is marked by one or more of the following:** • addresses the topic and task using somewhat developed explanations, exemplifications, and/or details • displays unity, progression, and coherence, though connection of ideas may be occasionally obscured • may demonstrate inconsistent facility in sentence formation and word choice that may result in lack of clarity and occasionally obscure meaning • may display accurate but limited range of syntactic structures and vocabulary
2	**An essay at this level may reveal one or more of the following weaknesses:** • limited development in response to the topic and task • inadequate organization or connection of ideas • inappropriate or insufficient exemplifications, explanations, or details to support or illustrate generalizations in response to the task • a noticeably inappropriate choice of words or word forms • an accumulation of errors in sentence structure and/or usage
1	**An essay at this level is seriously flawed by one or more of the following weaknesses:** • serious disorganization or underdevelopment • little or no detail, or irrelevant specifics, or questionable responsiveness to the task • serious and frequent errors in sentence structure or usage
0	**An essay at this level** merely copies words from the topic, rejects the topic, or is otherwise not connected to the topic, is written in a foreign language, consists of keystroke characters, or is blank.

Appendix B2: IEPLEC Sample Essay – Low Quality (Score of 2.5)

Sample 37

Entering a new country is excited, but many people will be homesick. Now, I live in Fort Collins while I'm studying English in study abroad. Study abroad makes me happy and excited, but sometime it makes me sad. Especially, I worry about habit in USA such as different language. I have to become accustomed to USA such as be studying English better, listening American music and talking American or foreigners in English.

I can't speak English very well because I need to study English better in IEP. First, I have to go school and learn how to write the essay or speak my presentation. IEP teacher teach us everything because I have to study hard. I must not forget my homework every day. Second, I try to participate in the class. Japanese is shy and quiet because we look like not participating. In the class, I always worry because when I have to answer the question, I don't have confidence. Because I have to have my confidence in the class and participate. If I can it, I will be able to speak English better.

I like listening American music because it helps me grown my hearing. I always listen to American music in USA. When I live in Japan, I often listened to it, but I couldn't understand the lyrics. Now, I thought to listen to music helps me English studying. Because I always listen to American music, and I try to understand the lyrics. Sometimes, I can understand it, but almost I can't understand. I have to try more. Especially, I want to understand the meaning. If I can it, my hearing will be better.

I like talking American or foreigners about friends, hobbies or classes. First, I think to talking is very important things for me and it helps my listening and speaking. For example, I can understand American saying, but Arabic or another country people's pronunciation is difficult for me. Sometimes, I can't understand their English. I often think sorry. But I think it is good practice. I always think Arabic or another country people can speak English better because I respect their. But some teacher says "They can speak English better than you, but their grammar isn't better than you." Because I thought leaning speed or timing from they or American. I want to talk American or foreigners more. It helps me to be good speaker. I try to talk and make many friends.

To use different language makes me worry or sad, but it makes me grown mind. It has some solution such as be studying, listening American music and talking American or foreigners. If you have a problem about different language, you don't have to worry. If you can't have confidence about your English skills, you should study English better in your English classes. And if you like listening America music, you should keep it. You will get a good hearing skill. Finally, if you want to talk American or foreigners, you can try it. You will be good speaker. Conclusion, I think entering new country is difficult for you, but it is good challenge for you. If you want to grow your mind, you try to study abroad. How do you think if you go to new country?

Appendix B3: IEPLEC Sample Essay – High Quality

Sample 49

Do you want to know how to learn English effectively? If so, you are not alone. In fact, there are many people out there today who are working to learn English as a second language. No matter what the reason is that you are learning English, you want to make sure that you can learn it quickly and effectively as well. Whether you are learning English in a class, on your own, or with language teaching software, there are certain things that you can do to make sure that you learn the language effectively like **surround yourself with English by using all recourses, watch English films and television, and do exercises and take tests**

First of all, surround yourself with English as much as you can. The absolute best way to learn English is to surround yourself with it. Take notes in English, put English books around your room, listen to English language radio broadcasts, watch English news, movies and television. Speak English with your friends whenever you can. The more English material that you have around you, the faster you will learn and the more likely it is that you will begin "thinking in English."Also **Use all of your resources**. Even if you study English at a language school it doesn't mean you can't learn outside of class. Using as many different sources, methods and tools as possible, will allow you to learn faster. There are many different ways you can improve your English, so don't limit yourself to only one or two. For example, the internet is a fantastic resource for virtually anything, but for the language learner it's perfect. Focusing in English it is a very powerful way to learn it fast.

In addition, watch English films and television is a helpful way. This is not only a fun way to learn but it is also very effective. By watching English films especially those with **English** subtitles you can expand your vocabulary and hear the flow of speech from the actors. If you listen to the news you can also hear different accents. Moreover, **listening to English music it is helpful way also**. Music can be a very effective method of learning English. In fact, it is often used as a way of improving comprehension. The best way to learn though is to get the words to the songs you are listening to and try to read them as the artist sings. There are several good internet sites where one can find the words for most songs. This way you can practice your listening and reading at the same time.

Last but not least, **doing exercises and take tests it is effective way to learn fast.** Many people think that exercises and tests aren't much fun. However, by completing exercises and taking tests you can really improve your English. One of the best reasons for doing lots of exercises and tests is that they give you a benchmark to compare your future results with. For example, every Saturday I usually take test on the internet and I realize that my English language it is improved per weeks. In addition, **record yourself**. Nobody likes to hear their own voice on tape but like tests, it is good to compare your tapes from time to time. For example, Every Friday I have to record one audio diary in my listening speaking class. You may be so impressed with the progress you are making that you may not mind the sound of your voice as much but it success and effective way.

In cancelation, **these are some tips which may help you to master the English language and** you can do to make sure that you learn the language effectively. **Surround yourself with English by using all recourses, watch English films and television, and do exercises and take tests. These are many other effective ways like speaking without fear and listen to native speakers as much as possible and** only by studying things like grammar and vocabulary and doing exercises, can you really improve your knowledge of any language.

Appendix B4: ICLE Sample Essay – Low Quality (Score of 2.5)

Sample 9

I believe this statement is true. We should, however, be careful what we mean by "dreaming" and "imagination". Usually people connect these two words with something close to nature and in this sense, in our modern world, there is really no place for them. I would like to make it clear that this does not mean I am against technology or science. I am not. They have definitely made many spheres of our lives much easier, but, as everything in this world, this also has its negative sides.

First, one of the main reasons why dreaming and imagination have lost their way in our modern world, is because industrialisation has created a big rush. There seem to be so many things to do nowadays, everyone is in a hurry, people constantly have work to do. Technology works fast and, respectively, we have to keep up with it. We do not feel comfortable if we just sit down and dream, we simply feel we are wasting our time. As I mentioned above, dreaming and imagination are closely connected with nature. I have often realised that when I am away on holiday somewhere in the mountains, I totally lose track of time, I have nowhere to hurry to and I find it very easy to just sit down and dwell on the scenery, or just stare at the sky without feeling that I am wasting my time. Once you are back among technology, however, the noise and the everyday hustle make you run to a tight schedule and practically nothing can stop you, except maybe a world-wide short circuit, which would not be very recommendable.

Another major hindrance of imagination is devices such as computers and television.

Children do not have to make up games anymore, whatever fun they are looking for is right there in the computer, waiting for them to turn it on. I must mention, though, that there is some kind of room for imagination when it comes to computers (not television). It opens a different kind of imagination. It makes you imagine different worlds, galaxies, spaces, things, which if we had not been in touch with technology, would have been quite impossible. Nevertheless, there is no doubt, at least to my mind, that just sitting and watching something readily made, does not inspire you to do very much by yourself, your mind is not really challenged and this wastes away your imagination. So, no matter how educational television may be, it is an absolute destructor of imagination. I should mention, nontheless, that it probably gives huge ground for dreaming, after watching all those fantastic movie stars!

The irony of the whole thing is that if it was not for someone's free imagination, we would probably not have had all these technologies today. In a sense, technology ruins its own creator: imagination.

Appendix B5: ICLE Sample Essay – High Quality (Score of 4.5)

Sample 43

What does it mean to live? To live in a world of fantasy or in a real world? Maybe everyone can give an individual answer to this question. An old problem which still seems valid today: does a real world really exist or do we retire into own shrouded worlds, into our own "whispering chambers of imagination", do we surround ourselves with our own virtues, values, measures?

As for me, I am quite willing to believe that everyone lives " in their heads". Living in the last decade of the 20th century. One feels more and more confined by the severe imprisonment of mechanisation, industrialization, science, technology. One is energetically and emotionally involved in the crazy, hectic rat-race of our modern world. And this rat-race squeezes our emotions, deadens our real human feelings and creates a 'clock-mad mechanism out of a human being. So, welcome to the machine.

In this world of science, technology, industrialization most of our activities run through a kind of narrow neck of a funnel. But out thoughts, our minds, our imagination cannot be confined.

Because, as the matter of fact, industrialization is in a way a product of human imagination. This maze of plans, schedules, machines is founded on human imagination. On the other hand, eventhough we live in a kind of vicious circle, eventhough we are absorbed in numerous, meaningless, even absurd activities, we are free to let our fancies roam. Human strivings for knowledge are endless; human thoughts can escape from this rapidly changing world of new technologies, strange devices, inventions of all kinds, world of crazy experiments, complex apparatus, virtual reality, mechanisms, machinery, clock-works. One's mind is free, one can withdraw in a fictitious, fantasy world, in a world of dreams, an internal world where there is no verbal clarity.

The problem, however, is that man lives by force of habit, because we get into the habit of living first, and later into the habit of thinking. When one gets into the habit of thinking, one understands the absurdity of all these technologies, scientific achievements. That is why at a certain point of realization one becomes aware of the meaningless of the hectic, dusty, noisy electronic world surrounding us. A person who has become aware of their absurd existence tries to find the Real Thing in life, to start a new, active life devoted to self-perfection. Then life becomes a kind of rebellion against the harsh reality of life. Man's power is in his conscious rebellion against a reality which is much stronger than him. Maybe it is through imagination that one can escape from this harsh reality of life and enter a world of freedom, a world independent of time, society, space, technology, industrialization. Therefore, once we realize the absurdity of this humanized world we rebel against it using our imagination.

It is also because of the existence of art that we can safely assume that there is time for dreaming and imagination. It is in art where one can give the reins to one's imagination. One can transform the modern, industrialized world into a piece of art; because art is deeply influenced by one's inner emotions, thoughts, dreams, experiences. In fact, we can safely say that modern art is on the verge of unconsciousness, dreaming, knowing this. I nurse the hope that there will always be "some unknown regions preserved as hunting grounds of poetic imagination."

Appendix C

Table C1

Descriptive Statistics of IEPLEC Essay Predictors

	READNW	CONLGpi	CREFC1u	STRUTa	LSApssa
Mean	568.82	41.05	.19	.11	.26
Median	536.50	42.25	.19	.11	.24
Mode	502.00	15.44	.14	.11	.22
Standard Deviation	155.01	13.41	.05	.03	.08
Skewness	.59	.18	.44	.88	1.07

Table C2

Correlation Matrices for Study 1a

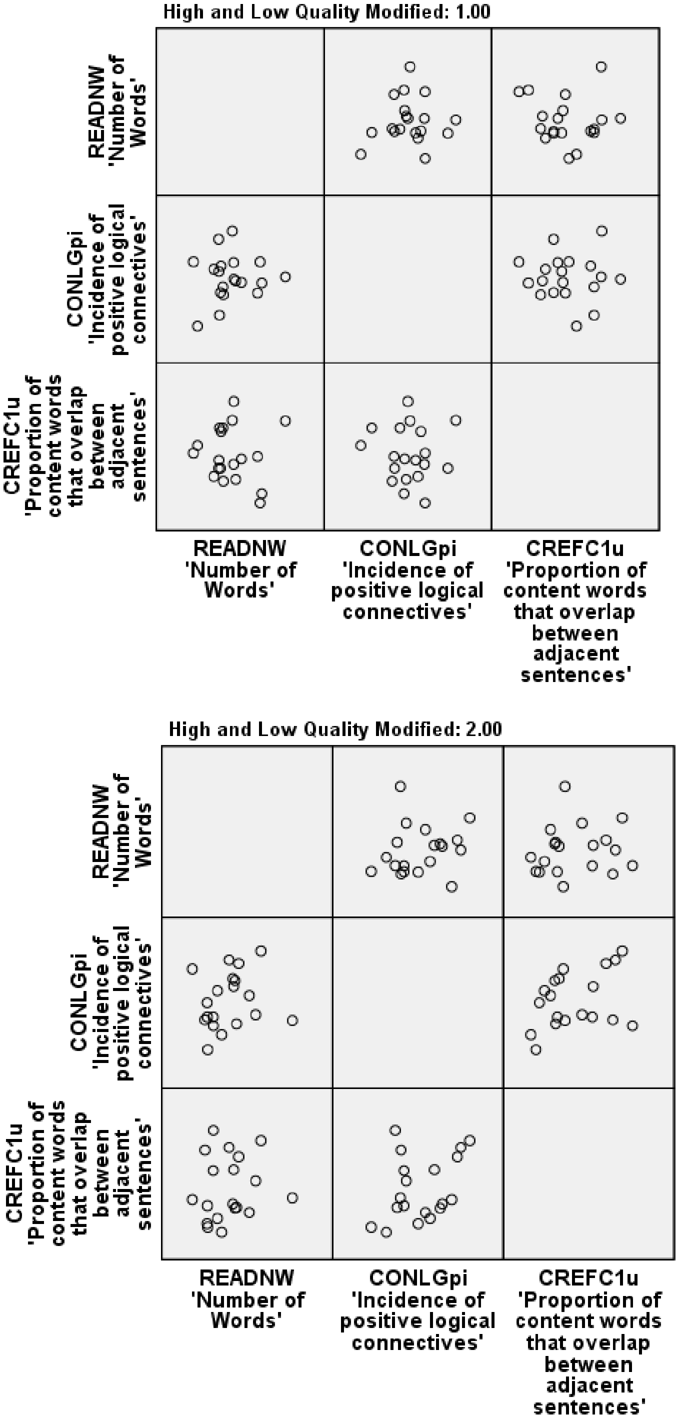
Predictor Variable		READNW	CONLGpi	CREFC1u
	Pearson Correlation	1	-.068	-.199
READNW	Sig. (2-tailed)		.687	.231
	N	38	38	38
CONLGpi	Pearson Correlation	-.068	1	.258
	Sig. (2-tailed)	.687		.118
	N	38	38	38
CREFC1u	Pearson Correlation	-.199	.258	1
	Sig. (2-tailed)	.231	.118	
	N	38	38	38

Table C3

Correlation Matrices for Study 1b

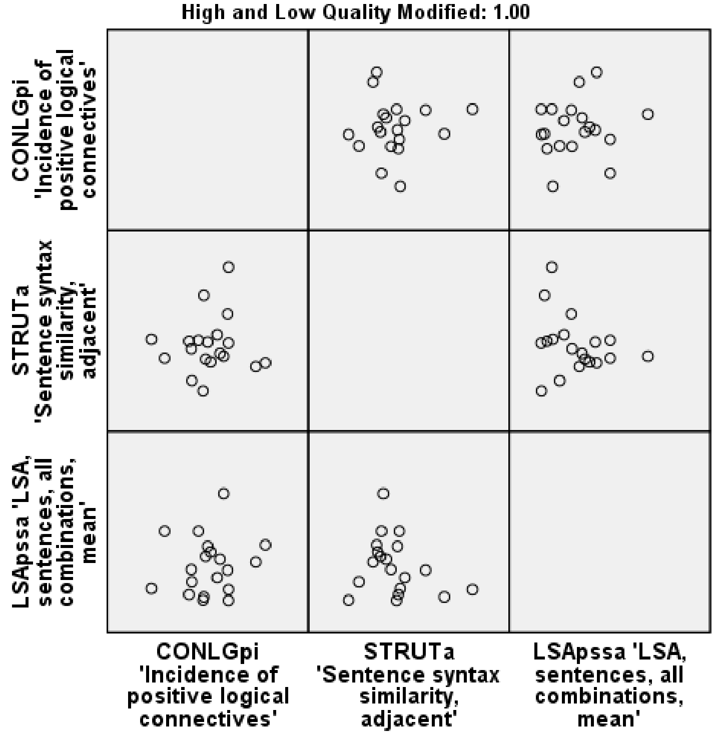
Predictor Variable		LSApssa	CONLGpi	STRUTa
	Pearson Correlation	1	.265	.001
LSApssa	Sig. (2-tailed)		.107	.995
	N	38	38	38
CONLGpi	Pearson Correlation	.265	1	.242
	Sig. (2-tailed)	.107		.143
	N	38	38	38
STRUTa	Pearson Correlation	.001	.242	1
	Sig. (2-tailed)	.995	.143	
	N	38	38	38

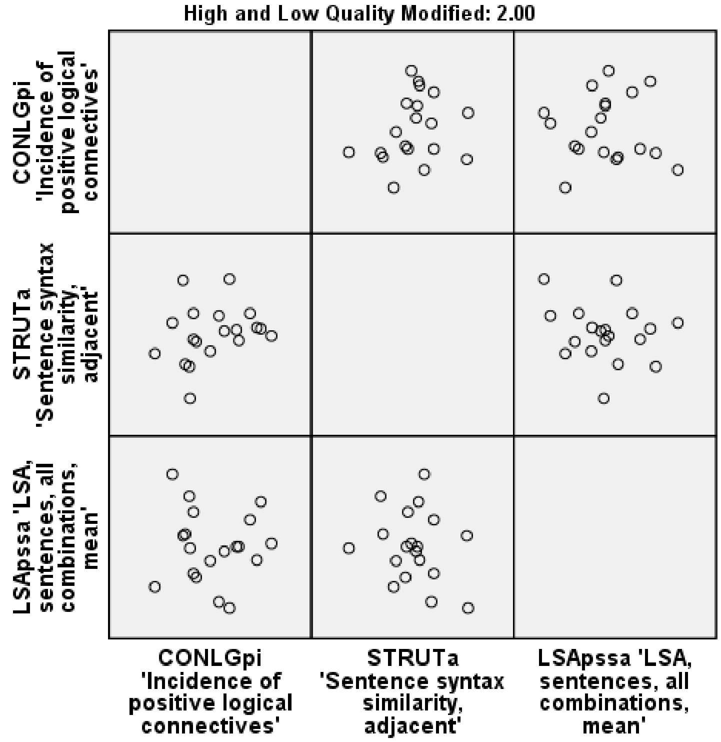
Table C4: Scatterplots for Study 1a



Note: Low Quality = 1.00; High Quality = 2.00

Table C5: Scatterplots for Study 1b





Note: Low Quality = 1.00; High Quality = 2.00

Table C6

Descriptive Statistics of IEPLEC Essay Predictors in ICLE

	READNW	CONLGpi	CREFC1u	STRUTa	LSApssa
Mean	518.32	28.39	.09	.08	.15
Median	507.00	27.31	.08	.08	.15
Mode	462.00	8.66	.07	.09	.16
Standard Deviation	67.89	11.33	.025	.02	.05
Skewness	.846	.46	.40	.50	.53

Table C7

Correlation Matrices for Study 1a

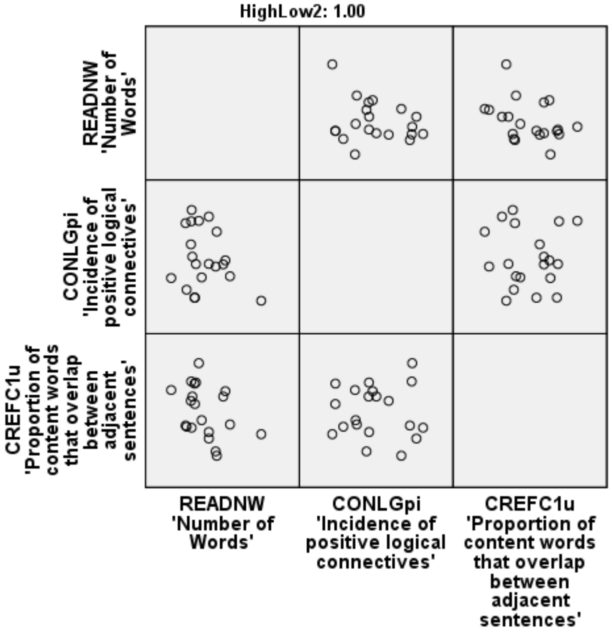
Predictor Variable		READNW	CONLGpi	CREFC1u
	Pearson Correlation	1	.20	-.21
READNW	Sig. (2-tailed)		.236	.205
	N	37	37	37
CONLGpi	Pearson Correlation	.20	1	-.09
	Sig. (2-tailed)	.236		.601
	N	37	37	37
CREFC1u	Pearson Correlation	-.21	-.09	1
	Sig. (2-tailed)	.205	.601	
	N	37	37	37

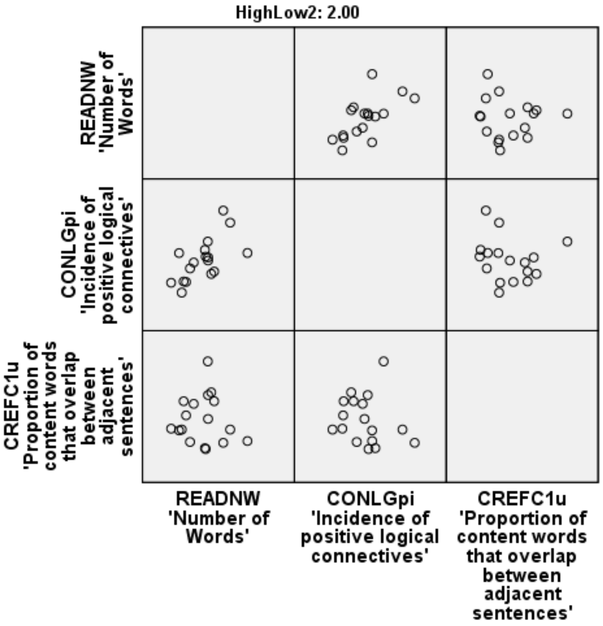
Table C8

Correlation Matrices for Study 2b

Predictor Variable		CONLGpi	LSApssa	STRUTa
CONLGpi	Pearson Correlation	1	-.15	-.15
	Sig. (2-tailed)		.392	.387
	N	37	37	37
LSApssa	Pearson Correlation	-.15	1	-.10
	Sig. (2-tailed)	.392		.577
	N	37	37	37
STRUTa	Pearson Correlation	-.15	-.10	1
	Sig. (2-tailed)	.387	.577	
	N	37	37	37

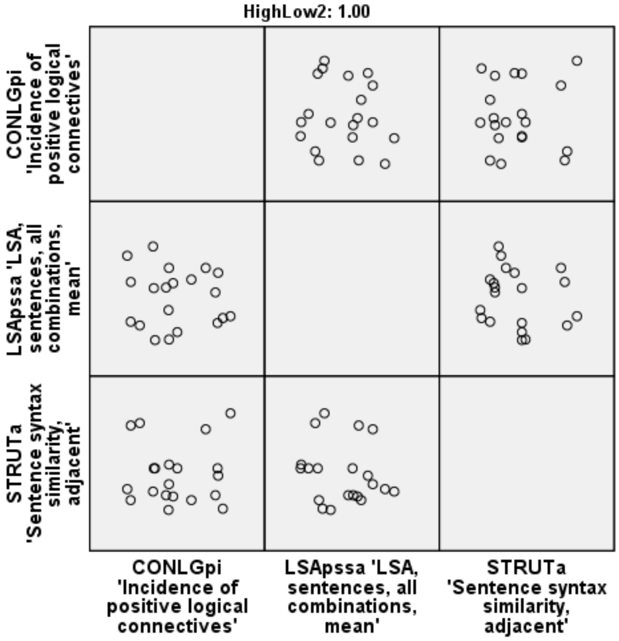
Table C9: Scatterplots for Study 2a

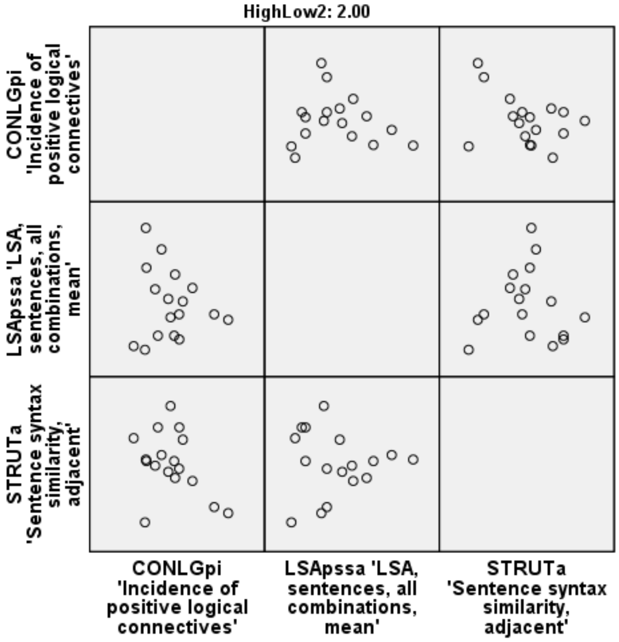




Note: Low Quality = 1.00; High Quality = 2.00

Table C10: Scatterplots for Experiment 2b





Note: Low Quality = 1.00; High Quality = 2.00

Table C11

Descriptive Statistics ICLE Essay Predictors

	READFRE	WORDCacw	SPATC	LSAppa	SYNNP
Mean	55.59	351.93	.50	.29	.80
Median	57.94	347.84	.50	.28	.79
Mode	31.41	325.68	.50	.09	.65
Standard Deviation	10.61	16.14	.05	.08	.14
Skewness	-.16	.58	-.04	-.07	.24

Table C12

Correlation Matrices for Study 2c

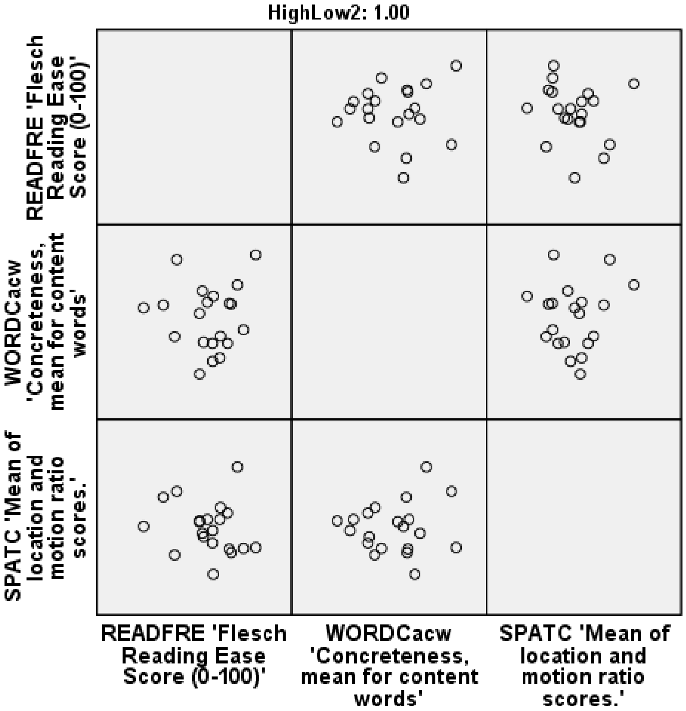
Predictor Variable		READFRE	WORDCacw	SPATC
READFRE	Pearson Correlation	1	.29	-.31
	Sig. (2-tailed)		.083	.060
	N	37	37	37
WORDCacw	Pearson Correlation	.29	1	-.23
	Sig. (2-tailed)	.083		.180
	N	37	37	37
SPATC	Pearson Correlation	-.31	-.23	1
	Sig. (2-tailed)	.060	.180	
	N	37	37	37

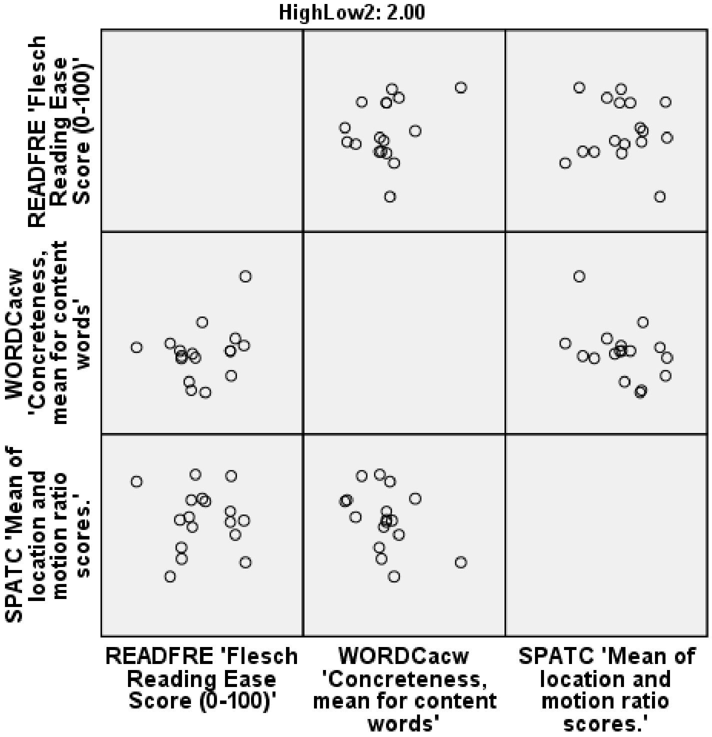
Table C13

Correlation Matrices for Study 2d

Predictor Variable		WORDCacw	LSAppa	SYNNP
	Pearson Correlation	1	-.26	-.11
WORDCacw	Sig. (2-tailed)		.125	.520
	N	37	37	37
LSAppa	Pearson Correlation	-.26	1	.25
	Sig. (2-tailed)	.125		.136
	N	37	37	37
SYNNP	Pearson Correlation	-.11	.25	1
	Sig. (2-tailed)	.520	.136	
	N	37	37	37

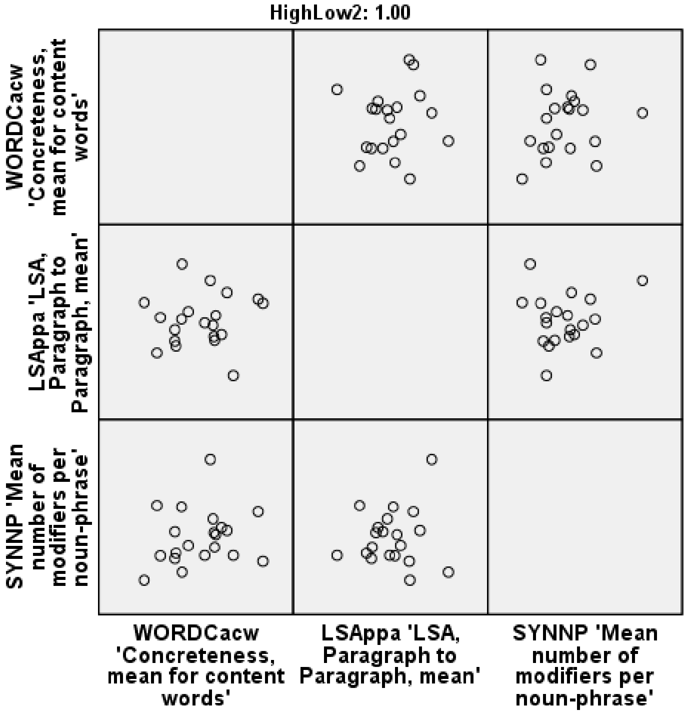
Table C14: Scatterplots for Experiment 2c

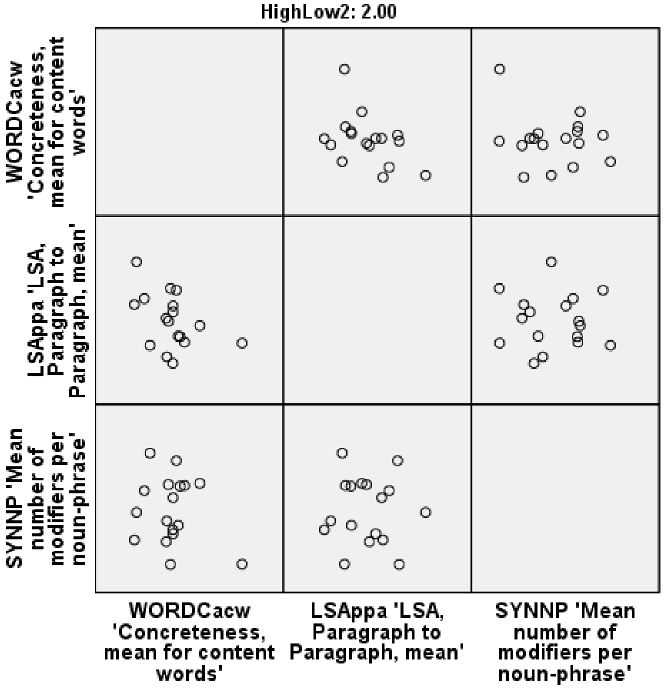




Note: Note: Low Quality = 1.00; High Quality = 2.00

Table C15: Scatterplots for Experiment 2a





Note: Low Quality = 1.00; High Quality = 2.00