Using corpus complexity analyses to refine a holistic ESL writing placement rubric

Jeremy R. Gevara
Pennsylvania State University

The purpose of this study is to determine if corpus analysis tools can identify linguistic features within writing placement samples that are significantly different between levels within a higher education language program. Although commercial tests are widely used for placement decisions, local performance assessments have become more common compliments that better adhere to communicative language teaching. At the university where this study was conducted, raters use a holistic rubric to score students' responses to one academic topic. The scoring process is fast when rates agree but too time consuming when raters search for information to resolve disagreements. Writing placement samples from 123 former students' essays at an Intensive English Program were used to compile a corpus. I divided the writing samples into four folders that correspond with the program levels and analyzed the folders using syntactic, lexical, and essay complexity analyzers. I utilized the robustness of the ANOVA to account for assumption violations. Data that violated the normality assumption were first analyzed using the Kruskal-Wallis Test. Those variables showing significant differences between levels were then analyzed using ANOVA and the appropriate post-hoc tests. Results show significant between group differences with lexical and word types and tokens, complex nominal, verb phrases, and ideas. I discuss the interpretation of these variables as well as show how administrators used this information to revise the rubric from Version I to Version II. Broader implications from this study are the use of corpus research tools to operationalize performance for the purposes of model building.

Key words: Writing, Placement, Holistic, Rubric, Corpus

Jeremy R. Gevara, The Pennsylvania State University Department of Applied Linguistics, 234 Sparks Building University Park, PA 16802.

Introduction

At an Intensive English Program (IEP) in the United States, a placement assessment consisting of two tests is given to all incoming students. One test is the English Placement Test (EPT), made by the English Language Institute (2006), and consists of selected response items that measure Listening, Grammar, Vocabulary, and Reading abilities. The second is an in-house writing test that acts as a close replication of classroom activities within the program. Because the writing test was developed from a task-focused theory of language knowledge (Bachman, 2007), it consists of a prompt for students to complete and a rubric for raters to give a score (Norris, Brown, Hudson, & Yoshioka, 1998). The tasks in the writing test have been developed over several years by IEP faculty, but the rubric has only undergone evaluation and revisions over the past two years.

The challenge with revising the writing test rubric is balancing administrators' two needs from the test. One is noted by Brown (2005) as the purpose of placement tests is to distinguish test takers that have a large range of the target ability. Because these tests are commonly perceived as low-stakes, many language programs require placement decisions to be made within a short amount of time (Green, 2012). At the IEP where this study was conducted, the second need it to make decisions within the same day that the assessment was administered. Brown (2012) describes time constraints as a reason for why many constructed response placement items are graded using a holistic rubric.

This kind of rubric has the advantage of requiring only one score, saving time, but is limited in providing information on what raters select as distinguishing test takers' performances. Using the first version of the writing placement rubric, **Appendix A,** as an example, a holistic rubric with only two levels for raters to score would require them to distinguish six pieces of information, two levels containing three descriptors. The rubric in **Appendix A**, however, has raters distill up to 14 pieces of information into one score that they must be agreed on by all raters. In this study, I will analyze the contents of essays organized by IEP levels, making a multilevel corpus, to identify any linguistic features that are significantly different among the four levels. The organization of the IEP corpus into level folders will be based on final placement decisions made for all test takers. I will analyze the four folders, one for each level in the IEP, with three corpus complexity tools, syntactic, lexical, and essay. By submitting the results of these complexity tools to tests of significant differences, I will be able to identify linguistic features within each tool that are different between pairs of levels.

These features will be information that is either confirmed to be present in version I rubric or needing to be added in version II, **Appendix D**.

Literature Review

Rubrics

Scoring rubrics are a systematic way for individuals to rate a performance by responding to one or more described observable variables. Holistic rubrics are one type that elicit a single score based on raters' overall impressions. The advantage of this rubric is that the assignment of scores can be done quickly without extensive attention given to all descriptions within categories. Harrington (1998) provides evidence that the use of a holistic rubric better matched teachers' expectations of student performance. Barkaoui (2007) supports the psychometric qualities of a holistic rubric by comparing holistic and analytic rubrics through calculated phi coefficients. A possible challenge to these conclusions, however, is that test developers are not sure if all information within the rubric helped raters to make a decision. This limitation would not be considered as long as the single scores are within a desired degree of agreement.

Hamp-Lyons (1991) challenges the use of holistic rubrics by arguing that the score given is conformity to a single point of information. If using a placement test to determine whether a student should be enrolled in a course or not, a single point of information could be sufficient. A single piece of information, however, will not suffice for the purposes of dividing test takers into more than two categories. Several researchers argue that constructed response tasks offer a large variety of information in their responses (McNamara, 1996; Norris et al., 1998), but raters may not adequately consider this when assigning a single score to the essay. Charney (1984) shows that raters were more significantly influenced by superficial features of essays, e.g., length of essay or spelling mistakes. If raters are supposed to use holistic rubrics for global impressions of students' performances, there should be multiple descriptors that are concise.

Analytic rubrics contain descriptors that break test takers' performances into several parts that raters can distinguish from each other. These descriptors are normally scored individually, but Brown (2012) shows several rubrics that contain several descriptors that are scored holistically. Readers will note that version I of the writing placement rubric looks analytic although IEP administrators score it holistically. Weigle (2002) notes that the use of analytic rubrics in writing performances helps raters by providing more detailed information to justify assigned scores. Hamp-Lyons (2007)

further supports this statement by discussing the use of holistic rubrics creating issues for standardizing raters.

My review of the literature so far supports the idea of having several descriptors within the writing test rubric to help raters distinguish performance between levels. Hamp-Lyons (2007) notes that more descriptors resulted in an extended discussion between raters she observed. This is good for resolving differences in rating, but the additional information could result in raters taking more time to score. My solution for this problem is to refine the descriptors to linguistic features that are present in different frequencies across the four IEP levels. Similar to standard setting (Brown & Hudson, 2002), two methods are available for identifying linguistic features within students' essays. One method is test-centered, asking raters as subject matter experts to identify linguistic features they feel are important. The other method is student-centered, analyzing the essays for features that are most salient and different amongst the levels in the program.

Student-centered rubric design through corpus analysis

Flowerdew (2009) describes corpus-based approaches as becoming increasingly researched for language teaching purposes. One reason for this emergence are developments in personal computing. Current personal computers now allow researchers to compile or access corpora containing millions of words that can be queried within seconds. An example of this can be found with the Corpus of Contemporary English (COCA). Created by Davies (2008), this corpus currently contains 450 million words and is free to use from any computer.

Another reason for Flowerdew's description of corpus-based approaches as emerging in language teaching contexts is that the various analysis methods are still being understood. One method is qualitative data analysis, querying the corpus by tags to identify and discuss emerging patterns. Hyland (2010) uses this method to examine a corpus of academic texts organized by genre. One advantage to using this method is that a large amount of information can be gained from having a computer organize the corpus by desired categories. Hyland used a qualitative corpus analysis method to identify words, sentences, and passages that support his study of proximity, defined as the writer's control of rhetorical features to display authority and personal position (p. 117). One disadvantage to using a qualitative corpus analysis method, however, is that tagging texts is not automatically done by a computer. Programs such as AntConc (Anthony, 2014) are able to tag texts that are imported, but these tags are general and may not organize information for novel interpretations. Dryer's (2013) study is both an

example of corpus analysis for testing purposes and how manual tagging can be a challenge to analyzing a large corpus. The focus of Dryer's study is on understanding writing ability through the analysis of scoring rubrics. The amount of rubrics included in the corpus, 83, is significant but illustrates the limitations of manual coding.

The alternative to analyzing a corpus through qualitative methods is through quantitative methods, or analyzing a sample of text or speech for differences or patterns that could be generalized to a larger population of interest. Researchers are able to successfully use quantitative data analysis methods by running a computer program that tags the corpus by parts of speech. Banerjee and Yan (2015) conducted a discriminant function analysis using the results from Coh-Metrix (McNamara, Louwerse, Cai, & Graesser, 2013), an online program that quantifies a corpus of interest. Coh-Metrix is a popular tool for quantitative corpus analysis, but it has limitations both in its function and analysis. In terms of function, the current version of Coh-Metrix is only able to analyze one text at a time. This limitation would impact the size of the corpus that could be analyzed due to time constraints. Another limitation with using Coh-Metrix is that research is still ongoing about how to best quantify the features within a corpus. Lu's (2010, 2012) two studies compared several measures of syntactic and lexical complexity proposed by various researchers. Lu tested each measure by determining whether the results could separate college essays divided into four levels of English ability. Lu validated the measures with a sample that received the same instruction, whereas this study will test them with a sample that is more diverse in prior education.

Motivation for the Current Study

Because my reason for conducting this study is to better understand what linguistic differences are present among the texts divided into four IEP levels, I will use a quantitative corpus analysis with a variety of ways to measure complexity. In addition, the programs created by Lu for lexical and syntactic complexity have batch run options, meaning that all the texts within the corpus can be run with one command. The results from these two analyses would provide meaningful information about which features of grammatical knowledge (Bachman & Palmer, 1996) are significantly different between each pair of IEP levels, but results would only address the grammar and vocabulary criterion of rubric version I in **Appendix A**. I propose measuring the organization and content criterion by operationalizing textual knowledge. I define textual knowledge as the connection of two or more utterances or sentences. The program that I will use to measure essay complexity is the Computer Propositional Idea Density Rater (CPIDR).

The CPIDR (Covington, 2012) measures textual knowledge through the tagging of Ideas and the calculation of Idea Density. Ideas is broadly defined as a count of verbs, adjectives, adverbs, conjunctions, and prepositions within a text. Idea density is a value calculated from the ratio of ideas and the total number of words within a text. Previous research supports the use of the program to identify differences between written text divided into multiple performance levels (Elvevaag, Wynn, & Covington, 2011; Nicholson, 2009). If textual knowledge is operational, I would expect to see idea density increase significantly from the lowest IEP level to the highest. In other words, students should be able to discuss each idea in more detail before moving on to the next one. Along with being able to measure variables of textual knowledge, this program also has the advantage of having a batch run option.

Using the three corpus analysis tools, lexical complexity, syntactic complexity, and essay complexity, described above, I will answer the following research questions:

RQ1: Is there a significant difference among the IEP levels using any of the lexical complexity measures?

RQ2: Is there a significant difference among IEP levels using any of the syntactic complexity measures?

RQ3: Is there a significant different between IEP levels using the two essay complexity measures?

Methods

Participants

Essays used in this study were collected from an IEP at a large research university in the northeastern United States. Students first entering the program are required to take a placement assessment which includes the EPT and a writing performance test. I collected 123 essays from the Autumn semester of 2012 and organized into program levels based on overall placement results. Thirty-two students were placed in Level 1 (Beginner), 32 were placed in Level 2 (Intermediate), 32 were placed in level 3 (High-intermediate), and 27 students were placed in Level 4 (Advanced). The students who enroll in this program come from a wide range of language backgrounds, e.g., Arabic, Chinese, and Korean. Their reasons for enrolling in the program are to seek admission into the university majoring in subjects such as business, engineering and education. The majority of students are seeking an

undergraduate admission and range in age from 18-24 years. The only criterion used to eliminate participants from inclusion in the study was that some Level 1 students did not make an attempt at the writing performance section due to low language ability. Because the typical amount of time spent in the IEP is 1.5 years, this sample is the majority of students currently in the program.

Materials

For the data collection instrument, the writing performance section of the placement test consists of a task prompt and the rubric raters use to assign a score. The task prompt is selected by the teacher in charge of placement testing and comes from a task bank developed by the program. The topics are intended to elicit an argumentative essay that students compose using information from personal experiences. Raters use rubric version I, **Appendix A**, to give one score from a four point Likert scale that corresponds to the IEP levels.

In order to prepare the corpus of students' essays for complexity analyses, I ran several programs on the corpus using the Linux operating system. The text files were annotated using the Stanford Part-of-Speech Tagger (Toutanova, Klein, Manning, & Singer, 2003) with the Wall Street Journal model, Stanford Parser (Klein & Manning, 2003), and the Morpohological Processing of English (Humphreys, Carroll, & Minnen, 2002). When running the Morphological Processing of English program, I chose the –t option to ensure the Part-of Speech tags were printed with the lemmas in the output. **Table** 1 below shows some descriptive statistics for the essays that were organized in each level folder of the IEP corpus.

Table 1. Descriptive statistics for essays within each level folder

Level	Average Words	Min. Words	Max. Words
One	130.59	51	263
Two	189.22	95	380
Three	233.84	144	403
Four	287.44	142	475

Procedures

The order of the placement assessment at the IEP is first the writing performance test followed immediately by the EPT. For the writing performance section, students are given the one prompt and instructed to write as much as they can in 50 minutes. Proctors are permitted to clarify instructions for students but no help is provided in

regards to the content of the essay. During the administration of the writing test, raters are given the training protocol by a senior instructor for standardization and clarification purposes. Every essay is given a holistic score by two raters with a third rater included when there is not agreement between the first two. Ratings are typically completed by the end of the EPT, about 80 minutes. Reliability is calculated after each writing test administration in order to report the results to an accreditation agency. Overall placement is calculated and discrepancies are addressed at the end of testing. These discrepancies could be differences in ratings for the writing test or placement differences between the EPT and writing test.

Once the annotation of the corpus was completed, the three corpus analysis tools were used to answer the research questions. Lu's (2012) system of lexical richness is comprised of 25 metrics that measures lexical density, sophistication, and variation. **Appendix B** describes the 25 measures and how they are calculated. I used this system to identify significant aspects of the Vocabulary category of the rubric version I. Currently, teachers use intuition to identify complex and academic vocabulary words. Lu's system compares every word in the essays to a corpus of *Wall Street Journal* articles. The system for measuring syntactic complexity comes from Lu (2010). Lu's system consists of 23 measures that are defined and described formulaically in **Appendix C**. I should note here that some of the metrics are transformations of another one already given. I did not exclude these from the results because there is still discussion about which metric best measures the reported linguistic feature. I used the syntactic complexity system to better understand significant aspects of the Grammar and Sentence Structure category. Lu's program analyzes several grammatical features for both simple and complex sentences.

For understanding the Content and Organization category, I used the CPIDR (Covington, 2012) to calculate the number of ideas and idea density, essay complexity. Ideas are similar to measuring the length of an essay, but Idea density shows differences in the amount of information placed in essays of the same length. Although this program is not an exact match to the Content and Organization category, it can provide some information on descriptions of idea and subtopics within the rubric. Being a stand-alone program, CPIDR takes the original texts and annotates the corpus using programs similar to the method described above. Several adjustment rules were also added to the program by Covington in order to increase the program's accuracy when compared to human raters. An example of ideas and idea density can be given from a simple sentence, "A small pug walked across the street." After applying annotation and adjustment rules, this sentence contains three ideas and an idea density

of .429, three ideas divided by seven total words. There is no standard for comparing texts against the two measures, but differences can be tested across levels.

Analysis

I answered all three research questions by using a one-way ANOVA to determine significant between group differences amongst the four levels in the IEP. Although a MANOVA is possible with this data set, not all dependent variables met the more stringent assumptions. I utilized the robustness of the ANOVA to compare all the measures by the same analysis despite many violating required assumptions. For dependent variables violating the normality assumption, the nonparametric Kruskal-Wallis test was first used. Using the Bonferroni adjustment, I included dependent variables showing a significant difference, $\alpha = .01$, in the one-way ANOVA along with those that met the normality assumption. Post-hoc analyses of significant differences were run using Tukey's HSD and Games-Howell. The Games-Howell post-hoc analysis is used for variables that violate the homogeneity of variance assumption (Larson-Hall, 2010).

Results

Before submitting the data to analysis for answering the research questions, I calculated reliability to ensure that the differences in variance can be attributed to linguistic features. Several measures of reliability are available depending on how many sources of variance exist within the instrument. Because the writing placement test in this study has two sources of variance, raters and test takers, I can calculate a Classical Test Theory reliability. The task cannot be a source of variance because test takers are only given one to answer.

Using Spearman-Brown's prophecy formula, suggested by Brown (2005), the interrater reliability for the ratings in this study's corpus is 0.92. This number is above Brown and Hudson's (2002) recommendation of 0.80 for language tests. The interpretation of this reliability is that the majority of the variance can be attributed to the variables I am investigating in this study. Using the kappa coefficient formula (Brown, 2005), the agreement between placement decisions using the writing test and the EPT is 0.89. This value, interpreted the same as reliability, is also acceptable but lower than the writing test interrater reliability, likely due to borderline students. Borderline students are individuals who show language ability that fits into more than one IEP level.

Lexical Complexity

Descriptive statistics and tests of normality for the 25 measures of lexical complexity are provided in **Table 2**. In addition, the results of the measures analyzed using the Kruskal-Wallis Test are given in the table. Values are shown for the results that support an adjusted significant difference between levels.

Table 2. Descriptive statistics and nonparametric between group differences of Lexical complexity measures

Measure	Mean	SD	Min	Max	Std. Skewness	Std. Kurtosis	Kruskal-Wallis
sentences	12.09	7.12	1.00	43.00	7.16	9.18	21.66
wordtypes	100.58	32.83	29.00	192.00	2.43	-.04	55.03
swordtypes	20.03	9.28	1.00	53.00	3.08	1.96	19.49
lextypes	62.90	24.36	12.00	135.00	3.30	.56	51.81
slextypes	16.75	8.38	1.00	47.00	3.43	2.06	18.70
wordtokens	209.77	87.70	51.00	477.00	3.25	.76	56.03
swordtokens	25.76	12.02	1.00	63.00	2.36	.12	21.53
lextokens	100.42	44.68	21.00	250.00	3.99	1.29	57.58
slextokens	20.64	10.55	1.00	54.00	3.06	.94	23.56
vs1	.13	.10	.00	.60	5.92	9.02	-
vs2	.54	.64	.00	3.60	9.59	13.29	-
cvs1	.43	.30	.00	1.34	2.20	-.18	-
ndw	100.58	32.83	29.00	192.00	2.43	-.04	55.03
ndwz	37.02	3.44	28.00	45.00	-1.07	-.56	-
ndwerz	37.32	2.48	28.90	42.30	-2.40	.71	-
ndwesz	37.36	2.60	28.30	43.30	-2.88	1.43	-
msttr	.74	.05	.56	.85	-3.36	2.82	-
cttr	4.92	.75	2.87	6.98	1.24	-.16	29.18
rttr	6.96	1.07	4.06	9.88	1.25	-.15	29.24
uber	17.86	3.23	11.60	30.29	2.82	1.47	-
svv1	12.00	5.57	3.12	31.37	5.66	4.03	21.32
cvv1	2.39	.54	1.25	3.96	2.84	.82	21.33
advv	.02	.01	.00	.07	2.79	1.03	-
modv	.15	.04	.07	.28	2.24	-.07	-

I divided the analyses of lexical complexity into smaller categories,dependent variables that are only frequency based and those that are formula based measures. The frequency based variables are sentences, wordtypes, swortypes, lextypes, slextypes, wordtokens, swortokens, lextokens, slextokens, and ndw. All these variables show significant differences from the results of the Kruskal-Wallis Test, a Chi-Squared value that has three degrees of freedom in this study. Because I am looking for measures that

show significant differences among the four levels, a larger value from the Kruska-Wallis test tells me there is more separation among the four IEP levels. Because the Kruskal-Wallis Test does not have a post-hoc test, I chose the variables of wordtypes, lextypes, wordtokens, lextokens, and ndw for further analysis due to a large X^2 value and noticeable differences between each level. **Table 3** shows the results of the one-way ANOVA calculated for the four variables chosen.

Table 3. Results of one-way ANOVA with effect sizes for frequency based measures

Measure	F(3,119)	Sig.	Partial Eta²	Adjusted R²
wordtypes	29.74	<.001	.428	.414
lextypes	27.90	<.001	.413	.398
wordtokens	29.48	<.001	.426	.412
lextokens	32.51	<.001	.450	.437
ndw	29.74	<.001	.428	.414

Games-Howell post-hoc analyses were run on all four dependent variables. Significant differences for wordtypes are found for all pairwise comparisons in level 1 and 2-4. For lextypes, significant pairwise comparisons are observed for level 1, 2-4, and 3-4. Significant pairwise comparisons for wordtokens are found for all levels but 3-4. Significant pairwise comparisons for lextokens are found for all levels except 2-3. Finally, the ndw variable shows significant pairwise differences between all levels except 2-3 and 2-4. Adjusted R squared values are reported along with partial eta squared values to account for issues of linearity in the data. All effects sizes shown in **Table 3** are considered large (Cohen, 1988).

The formula based measures, vs1, vs2, cvs1, ndwesz, msttr, cttr, rttr, uber, svv1, cvv1, advv, and modv, were analyzed using the Kruskal-Wallis Test. Results from the Kruskal-Wallis Test show significant between group differences for cttr, rttr, svv1, and cvv1, shown in **Table 2** with significance values also at p<.001. Using the same criteria for selection as with the frequency count variables, none of these variables were included with ndwz and ndwerz in the ANOVA. Because these two variables met the normality assumption, they were not included in the Kruskal-Wallis Test. Results from the ANOVA show that both ndwz and ndwerz are significant with $F(3,119) = 3.02$ $p = .033$ for ndwz and $F(3,119) = 3.30$ $p = .023$ for ndwerz. Tukey HSD post-hoc analyses show significant pairwise differences, p<.05, for ndwz between levels 1-4 and 2-4. For the ndwerz variables, significant pairwise differences are observed between levels 1-3 and 1-4. These differences within the two variables, however, do not meet the Bonferroni adjustment. Adjusted R squared values are small, .047 for ndwz and .054 for

ndwerz, supporting the use of the Bonferroni adjustment as the criterion for selecting dependent variables.

Syntactic Complexity

Descriptive statistics and tests of normality for the 23 measures of syntactic complexity are given in **Table 4**. In addition, the results of the measures analyzed using the Kruskal-Wallis Test are given in the table. Values are shown for the results that support an adjusted significant difference between levels.

Table 4. Descriptive statistics and nonparametric between group differences of syntactic complexity measures

Measure	Mean	SD	Min	Max	Std. Skewness	Std. Kurtosis	Kruskal-Wallis Test
W	200.67	77.46	51.00	405.00	1.85	.05	-
S	11.64	6.71	3.00	43.00	6.08	8.85	-
VP	29.95	11.96	8.00	65.00	2.63	1.00	25.06
C	23.44	10.01	7.00	56.00	3.60	1.88	19.23
T	13.27	6.56	4.00	40.00	4.80	4.86	13.90
DC	9.43	5.10	.00	24.00	2.88	.52	15.14
CT	6.07	3.16	.00	16.00	3.64	3.09	17.55
CP	5.65	3.64	.00	18.00	4.55	2.31	17.88
CN	22.05	9.66	3.00	48.00	1.92	.29	-
MLS	20.36	9.60	8.67	61.67	6.14	7.19	-
MLT	16.44	5.32	8.67	30.83	3.11	-.23	-
MLC	8.79	1.69	5.54	13.13	1.85	-.19	-
C/S	2.31	.97	1.11	5.67	5.30	3.69	-
VP/T	2.43	.73	1.45	4.67	3.91	1.38	-
C/T	1.87	.51	1.13	3.63	3.76	1.63	-
DC/C	.40	.14	.00	.69	-.80	.07	-
DC/T	.80	.46	.00	2.25	3.73	1.54	-
T/S	1.21	.27	.93	2.50	8.71	12.02	-
CT/T	.49	.20	.00	1.00	.60	-.18	-
CP/T	.49	.38	.00	2.50	10.16	19.45	-
CP/C	.26	.16	.00	.88	5.45	4.71	-
CN/T	1.83	.80	.50	4.17	3.05	.73	-
CN/C	.96	.31	.40	1.75	2.80	.47	-

The Kruskal-Wallis Test supports significant adjusted difference among the frequency based measures of VP, C, T, DC, CT, and CP. Using a standard of 50 that I observed from the previous analysis, none of these values were included in the ANOVA. Running a one-way ANOVA with the remaining variables of W and CN, both of these show significant between group differences, $F(3, 82) = 15.31$ $p < .001$ for W and $F(3, 82) = 19.57$ $p < .001$ for CN. Because both variables met the HOV assumption, I used the Tukey HSD for post-hoc tests. Significant pairwise differences are observed with W for all levels except 3-4. For CN, significant pairwise differences are observed for 1-3, 1-4, 2-3, and 2-4. Effect sizes for W are a partial eta squared value of .359 and an adjusted R squared value of .336. Effects sizes for CN are a partial eta squared value of .417 and an adjusted R squared value of .396. Effect size values for W are considered moderate while values for CN are considered large.

The formula based measures of of MLS, MLT, C/S, VP/T, C/T, DC/T, T/S, CP/T, CP/C, CN/T, and CN/C were analyzed using the Kruskal-Wallis Test, but none of the variables met the adjusted alpha of .01. The remaining variables of MLC, DC/C, and CT/T were analyzed using a one-way ANOVA due to them meeting the normality assumption. I observe similar results from the Kruskal-Wallis Test with none of the variables meeting the previously stated .01 alpha level.

Essay Complexity

Descriptive statistics and tests of normality for the Ideas and Density measures of essay complexity are given in **Table 5**. In addition, the results of the measures analyzed using the Kruskal-Wallis Test are given in **Table 6**. Values are shown for the results that support an adjusted significant difference between levels.

Table 5. Descriptive statistics of essay complexity measures

Measure	Mean	SD	Min	Max	Std. Skewness	Std. Kurtosis	Kruskal-Wallis Test
Ideas	105.43	45.28	25.00	229.00	3.00	.157	61.48
Density	.51	.03	.41	.61	1.57	.62	-

Table 6. Nonparametric group differences and effect sizes of essay complexity measures

Measure	Kruskal-Wallis Test	Partial eta squared	Adjusted R^2
Idea	**61.48**	**.465**	**.451**
Density	**-**	**.105**	**.105**

Because the result of the Kruskal-Wallis Test for Ideas is significant, Ideas and Density were both analyzed using the one-way ANOVA. Results show a significant difference for both variables, $F(3,119) = 34.46$ $p<.001$ for Ideas and $F(3,119) = 4.64$ $p< = .004$ for

Density. I ran the Games-Howell post-hoc test for Ideas and Tukey HSD for Density. Results from the Games-Howell post-hoc test show significant pairwise differences for all four levels. Results from the Tukey HSD for Density, on the other hand, show significant pairwise differences for 1-3 and 1-4. Effects sizes for Ideas are large, a partial eta squared of .465 and an adjusted R squared of .451. Effect sizes for Density are small, a partial eta squared of .105 and an adjusted R squared of .082.

Discussion

In order to discuss the significance of this study's results to the version I rubric, I will first briefly summarize the results. From the lexical complexity measures, I recommend word tokens (wordtokens) and lexical tokens (lextokens) for consideration. I included five measures in the final ANOVA calculation, but these two had the greatest post-hoc differences. The post-hoc results, however, were not significant for every possible comparison. An additional reason for recommending word tokens and lexical tokens is that they complement each other in terms of significant pairwise comparisons. Post-hoc results from wordtokens support significant differences for all pairs but 3-4, which lextokens do support as a significant difference. From the syntactic complexity measures, I recommend words (W) and Complex Nominals (CN) for consideration. Because the words measure is similar to word tokens in lexical complexity, this result serves as a replication of the one previously selected. This redundancy can serve as support that the two tools are counting words in a similar way. This evidence would be similar to running two statistical programs that differ in their rounding rules to check on a result's validity. Finally, I recommend Ideas for consideration from the essay complexity measure.

The results from CPIDR suggest that raters are able to distinguish test takers' performances by the amount of unique information given in their responses. Information organized into clusters of ideas, however, is not supported from the results The results from Idea Density do not support a strong relationship between complex sentences and the length of essay responses. I can also interpret from this result that Textual Knowledge was not successfully operationalized in this study. In other words, raters are better informed by the amount of different verbs, adjectives, adverbs, conjunctions, and prepositions used rather than the connection of sentences within a passage. Although a good essay is considered to consist of simple and complex sentences, the results from idea density show that the length of the essays may not be enough to produce this desired balance. One possible reason could be the time constraint placed on the writing placement test. A limitation of using Ideas alone is that

it does not give information on what types of ideas are contributing to the significant differences. The lexical and syntactic measures could help me to also understand what differences in grammatical knowledge could contribute to the significant sentence complexity results.

Results from both lexical and syntactic complexity support that the amount of word tokens used in each response is meaningful for assigning placement. Word tokens differ from word types in that the count for the former is of every word while the latter is a count of every word's function. Words tokens and idea density could be difficult to distinguish between each other, more words tokens would appear to mean more idea density. The difference between the two is that idea density takes into account how many function words were used in the response. The amount of function words used in a passage will have a different impact on the idea density than count of word tokens. The length of the passage can be given a separate category from the kind of sentences within the passage.

An additional significant variable identified from the syntactic complexity analysis is Complex Nominals (CN). Because the formulas using CN were not identified as supporting a significant difference between the four IEP levels, only its presence in a text supports higher level placement. A limitation with using this variable for grammatical knowledge is that the post-hoc results only support significant difference between 1-3, 1-4, 2-3, and 2-4. A possible solution for using this information could be in the form of a decision tree (Fulcher, Davidson, & Kemp, 2011). By having raters address the amount of complex nominals in a passage toward the beginning, the information could be used to divide the possible placement decision in half, either 1-2 or 3-4.

Finally, one of the recommendations from lexical complexity, word tokens, has been discussed and supported through other results. The other recommendation that has not been discussed so far is lexical tokens. This variable offers a different way for raters to analyze the passages that would contribute additional meaningful information for placement decisions. Lexical tokens differ from word tokens by counting the amount of different words used in a passage. The results from this measure show that test takers placed in higher levels will use a larger variety of words in their responses. A limitation with using this measure by itself to measure grammatical knowledge is that it does not support a significant 2-3 pairwise comparison. This could be addressed by using a decision tree design similar to the one proposed for complex nominals.

After this study was conducted, the results were presented to instructors at the IEP, who created version II of the rubric, **Appendix D**. The language within the rubric

was reduced and descriptions were revised to reflect many of the findings in this study. The amount of writing was added to the top of the rubric to more quickly measure the amount of words written by test takers. Although the results from the CPIDR did not successfully operationalize Textual Knowledge**,** Organization and Unity were retained by IEP instructors. The wording within the descriptor, however, was reduced to focus on the connections between sentences.

Instructors revised Grammar and Sentence Structure to focus on the complexity of sentences, supported by the CN results. Finally, they revised vocabulary to better match the results from ideas and get raters to notice the amount of different words produced by test takers. The next step in the writing performance test is to begin developing the decision trees I proposed for Complex Nominals and Lexical Tokens. Because the goal of the test is to give raters enough information to make a decision within a limited frame of time, I believe decision trees will speed up the holistic rating. Having these refined rubric descriptors should also allow raters to find information at times of disagreement.

A limitation of this study in regards to the IEP is that the results only address one half of the language knowledge model previously discussed (Bachman, 1990; Bachman & Palmer, 1996). The results taken together only account for the second level organizational knowledge portion. Sociolinguistic knowledge and functional knowledge, which make up pragmatic knowledge, are still described in the rubric as Content and Organization but not measured by any of the programs in this study. Measuring essay content within a corpus can be challenging because of issues operationalizing pragmatic knowledge, the selection of linguistic resources to function in a target context or accomplish a task. Scoring essay content with the IEP rubric requires raters to determine whether the answer given by a test taker answers the prompt. In addition, content also measures whether test takers have developed a clear connection of ideas that support their answer. Corpus analysis tools would need to be programmed to understand the essay prompt and measure the extent that the response answers the questions, a more subjective decision.

Torgersen, Gabrielatos, Hoffman, & Fox (2011) show that a major issue with measuring pragmatic knowledge within a corpus is that it requires manually tagging the features of interest. Because students coming into the IEP in this study typically have no prior experience in an English speaking country, raters and I have only observed different pragmatic markers used in the most advanced students, level 4. Examples of these pragmatic markers would be transition phrases such as "on the other hand" or use of modals like "should" to more closely align with academic writing. A

possible way to resolve this is through a priori methods (Weir, 2005). Measurable variables can be proposed that are theoretically supported and can be found within the domain of academic English. The scores given by raters for these would then be compared by level to determine if raters are able to successfully use the descriptors to differentiate test takers.

Implications

One implication from the results of this study concerns the use of corpora to identify quantifiable aspects of language knowledge. Previous research using corpora has looked at identifying patterns through qualitative data analysis methods (Dryer, 2013). This wealth of information, however, can be analyzed as well through quantitative data analysis methods. The change in data analysis method pushes the question to be framed within a hypothesized construct that is quantifiable and present through either sample or latent trait statistics. Because researchers have criticized the amount of information provided by a holistic rubric (Hamp-Lyons, 1991 Weigle, 2002), an additional implication from this study is a method that identifies linguistic features present but different across levels. These features would then become descriptors that contribute to holistic decisions but provide more targeted information about performances within each level.

An implication of this study for language testing is that it suggests a step towards transforming a paper rubric to a computerized one. Language programs and testing companies are becoming increasingly interested in automated raters and feedback systems (Xi, 2010). ETS currently uses the E-rater® system to assist with grading the TOEFL-iBT writing section (Enright & Quinlan, 2010). I envision this as the next step in development of the writing placement test at this IEP. Similar to E-rater®, the automated rating would be a compliment to instructors' scores that has the benefit of speeding up the grading process. After instructors verify the decisions match the ones they would make, they would only need to review borderline decisions. Because E-rater® provides a holistic score from a regression equation, the resources exist here to use the same data analysis method.

One contribution I can make to the body of research in Applied Linguistics is the challenge of using T-units to measure grammatical complexity. Biber, Gray, and Poonpon (2011) argue that the use of T-units to measure grammatical complexity does not fully account for academic writing development. Results from the syntactic complexity show no significant between group differences for formulas that used T-

units. The difference between these results and the ones reported by Lu (2010) is that the sample in this study is more diverse in terms of educational background. Participants used in this study received no instruction from the IEP before taking the test. The wide range of levels within the IEP program are one example of a desired trajectory for academic writing development that is not shown through the use of T-units.

In addition to validating the placement test, future studies could look at further understanding the complexity measures. Different genres or participant samples could be looked at to test the discriminative properties of the three complexity measures. Many of the measures used in this study are transformations and violate the independence assumption for using multivariate measures. The independence assumption could be met for using multivariate measures in a future study by limiting the amount of measures included in the analysis. Using the results from this study, I could select the measures that showed significant between group differences and meet the necessary assumptions. Gathering additional information on which complexity measures are more discriminating of group differences could help in the identification of a construct model.

Acknowledgments

I would like to thank Dr. Xiaofei Lu for his guidance and feedback during his course that this idea came from as a final project. I would also like to thank the reviewers for their feedback and patience with me during this process. I am grateful for the opportunity to continue developing my research skills and contribute to the field of language testing and Applied Linguistics.

References

Anthony, L. (2014). AntConc (Version 3.4.3) [Computer Software]. Tokyo, Japan: Waseda University. Available from http://www.laurenceanthony.net/

Bachman, L. F. (1990). *Fundamental considerations in language testing.* Oxford: Oxford University Press.

Bachman, L. F. (2007). What is the construct? The dialectic of abilities and contexts in defining constructs. In J. Fox, M. Wesche, D. Bayliss, L. Cheng, C. E. Turner, & C. Doe (Eds.). *Language Testing Reconsidered* (pp. 41-71). Ottawa, Ontario: University of Ottawa Press.

Bachman, L., & Palmer, A. (1996). *Language testing in practice: Designing and developing useful language tests.* Oxford: Oxford University Press.

Banerjee, J., & Yan, X. (2015). *Keeping up with the times: Triangulating multiple data sources to inform revisions to a writing rubric.* Paper presented at the annual American Association for Applied Linguistics, Toronto, ON.

Barkaoui, K. (2007). Rating scale impact on EFL essay marking: A mixed-method study. *Assessing Writing, 12*, 86-107.

Biber, D., Gray, B., & Poopon, K. (2011). Should we use characteristics of conversation to measure grammatical complexity in L2 writing development? *TESOL Quarterly, 45*(1), 5-35.

Brown, J. D. (2005). *Testing in language programs.* New York: Prentice McGraw Hill.

Brown, J.D. (2012). *Developing, using, and analyzing rubrics in language assessment with case studies in Asian and Pacific languages.* Honolulu, HI: National Foreign Language Resource Center University of Hawai'i.

Brown, J. D., & Hudson, T. (2002). *Criterion referenced language testing.* Cambridge: Cambridge University Press.

Charney, D. (1984). The validity of using holistic scoring to evaluate writing: A critical overview. *Research in the teaching of English, 10*(1), 65-81.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences.* (2nd ed.). Hillsdale, NJ: Lawrence Erlbaum Associates.

Cooper, T. C. (1976). Measuring written syntactic patterns of second language learners of German. *The Journal of Educational Research, 69* (5), 176–183.

Covington, M. A. (2012). Computerized Propositional Idea Density Rater. (Version 5.2). [Software]. Available from http://www.ai.uga.edu/caspr/

Davies, M. (2008). *The Corpus of Contemporary American English: 450 million words, 1990-present.* Available online at http://corpus.byu.edu/coca.

Dryer, D. B. (2013). Scaling writing ability: A corpus-driven inquiry. *Written Communication, 30*(1), 3-35.

Elvevaag, B., Wynn, R., & Covington, M. A. (2011). Case report: Meaningful confusions and confusing meanings in communication with schizophrenia. *Psychiatry Research, 186*, 461-464.

English Language Institute (2006). *English Placement Test (EPT) Examiner's Manual.* Ann Arbor, MI: University of Michigan.

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

Flowerdew, J. (2009). Corpora in language teaching. In M. H. Long, & C. J. Doughty (Eds.). *The handbook of language teaching* (pp. 327-350). Oxford: Wiley-Blackwell.

Fulcher, G., Davidson, F., & Kemp, J. (2011). Effective rating scale development for speaking tests: Performance decision trees. *Language Testing, 28*(1), 5-29.

Green, A. (2012). Placement testing. In Coombe, C., Davidson, P., O'Sullivan, B., & Stoynoff, S. (Eds.). *The Cambridge guide to second language assessment* (pp. 164-170). Cambridge: Cambridge University Press.

Hamp-Lyons, L. (1991). *Assessing second language writing in academic contexts.* Norwood, NJ: Ablex Publishing Corp.

Hamp-Lyons, L. (2007). Worrying about rating. *Assessing Writing, 12*, 1-9.

Harrington, S. (1998). New visions of authority in placement test rating. *Writing Program Administration, 22*(1-2), 53-84.

Humphreys, K., Carroll, J., & Minnen, G. (2003). Morphological Processing of English [Software]. Available from https://drive.google.com/folderview?id=0B9NJPETf6dB7elAyeTl6aVl0aWc&usp=drive-web

Hunt, K. W. (1970). Do sentences in the second language grow like those in the first? *TESOL Quarterly, 4* (3), 195–202.

Hyland, K. (2010). Constructing proximity: Relating to readers in popular and professional science. *Journal of English for academic purposes, 9*, 116-127.

Klein, D., & Manning, C. D. (2002). Fast exact inference with a factored model for Natural Language Parsing. *Advances in Neural Information Processing Systems, 15*, 3-10.

Larson-Hall, J. (2010). *A guide to doing statistics in second language research using SPSS.* New York: Routledge.

Leech, G., Rayson, P., & Wilson, A. (2001). *Word frequencies in written and spoken English: Based on the British National Corpus.* London: Longman.

Lu, X. (2010). Automatic analysis of syntactic complexity in second language writing. *International Journal of Corpus Linguistics, 15*(4), 474-496.

Lu, X. (2012). The relationship of lexical richness to the quality of ESL learners' oral narratives. *The Modern Language Journal, 96*, 190-208.

McNamara, T. (1996). *Measuring second language performance.* New York: Longman.

McNamara, D. S., Louwerse, M. M., Cai, Z., & Graesser, A. (2013). Coh-Metrix version 3.0. Retrieved [4/1/15] from http://cohmetrix.com.

Nicholson, C. (2009). *Proceedings of the IEEE Souteastcon '09.* Piscataway, NJ: IEEE.

Norris, J. M., Brown, J. D., Hudson, T., & Yoshioka, J. (1998). *Designing second language performance assessments.* Honolulu, HI: University of Hawaii Press.

Templin, M. C. (1957). *Certain language skills in children: Their development and interrelationships.* Minneapolis, MN: University of Minnesota Press.

Torgersen, E. N., Gabrielatos, C., Hoffman, S., & Fox, S. (2011). A corpus-based study of pragmatic markers in London English. *Corpus Linguistics and Linguistic Theory, 7*(1), 93-118.

Toutanova, K., Klein, D., Manning, C., & Singer, Y. (2003). *Conference proceeding of HLT-NAACL '03*. Boston: Association for Computational Linguistics.

Weigle, S. C. (2002). *Assessing Writing.* Cambridge: Cambridge University Press.

Weir, C. J. (2005). *Language Testing and Validation.* New York: Palgrave Macmillan.

Xi, X. (2010). Automated scoring and feedback systems: Where are we and where are we heading? *Language Testing, 27*(3), 291-300.

Appendix A

Writing Placement Test Rubric Version 1 (Emphasis in Original)

	Level 1	Level 2	Level 3	Level 4
Characterized by several of the following:				
Content and Organization	• main idea difficult to identify • loose and/or repetitious development • few sentences logically connected • introduction or conclusion missing or vague • whole paragraphs could be said in a single sentence • subtopics are unclear, missing, or unrelated to topic; if present, subtopics are unsupported • organization limited to sentence level	• main idea may be difficult to identify • demonstrates some recognizable development from beginning to end • sentences not always logically connected • introduction states main idea but with little or no development; conclusion minimal or missing • multiple ideas are expressed, but not sufficiently developed or succinctly expressed • subtopics are present but may be underdeveloped; support is present but weak • organization resembles large paragraph(s) or several small, underdeveloped paragraphs	• main idea easy to identify • introduction and conclusion clearly stated with some development • sentences and paragraphs logically connected • subtopics are present and logically divided; support is logical but may not be thoroughly developed • organization beginning to resemble an essay • fluency apparent and somewhat natural-structure may not resemble traditional essay organization	• main idea easy to identify • introduction shows development and is related to a specific main idea; introduction is interesting and may have a hook; conclusion offers insight • sentences and paragraphs clearly and logically connected • topic insightfully analyzed into subtopics which are clearly supported logically • organization resembles a well-developed essay
Grammar and Sentence Structure	• basic sentence structures with frequent errors • simple tense use • frequent errors with basic grammar	• simple sentences are mostly clear, but few examples of compound and complex sentences • simple and compound sentence structures with some errors • mostly simple tense use • may contain some attempt/evidence of more than simple tense use • some errors with basic grammar	• simple, compound, and complex sentence structures with some errors • varied tense use • some errors with advanced grammar	• sentence structures show clarity and variety with few errors • advanced tense use • few errors with advanced grammar
Vocabulary	• very limited range with frequent errors in word choice and form • repetitive, basic, general word use	• narrow range with few errors in word choice and form • mostly basic content words, may contain a handful of specific, less frequent content words	• wider range with some errors in word choice and form • beginning awareness of native speaker collocation patterns	• sophisticated range with few errors in word choice and form • awareness of native speaker collocation patterns

Appendix B

Measures of Lexical Complexity Adapted from Lu (2012)

Code	Description	Formula
T	Word Types – Counts a word by the function it serves in the sentences, e.g., noun, adjective, verb, etc	Frequency
N	Word Tokens – Counts every word in a text	Frequency
sentences	Sentences – Counts every sentences, marked by a period in the text	Frequency
swordtypes	Sophisticated Word Types – A word and its function are counted as sophisticated if it is not on the British National Corpus's (BNC) list of 2,000 most frequent words (Leech, Rayson, & Wilson, 2001)	Frequency
lextypes	Lexical Types – Counts a word each time it is used in a different function but only once	Frequency
slextypes	Sophisticated Lexical Types – Counts a word and its function if it is not in the BNC's list of most frequent words	Frequency
swordtokens	Sophisticated Word Tokens – Counts every word that is not in the BNC's list of most frequent words	Frequency
lextokens	Lexical Tokens – Counts every word in the text but only once	Frequency
slextokens	Sophisticated Lexical Tokens – Counts every word that is not in the BNC's list of most frequent words but only once	Frequency
ndw	Number of Different Words – Counts the number of word types in the utterance or text (Templin, 1957)	Frequency
vs1	Verb Sophistication I – A ratio of the number of sophisticated verbs, those not in the BNC's list of most frequent words, and the total number of verbs in the text	T_{sverb}/N_{verb}
vs2	Verb Sophistication II – The sophisticated verb count is squared to place more value on an individual selecting to use a less salient word	T^2_{sverb}/N_{verb}
cvs1	Corrected Verb Sophistication I – Reduces the impact of the total number of verbs in the text, a variation of placing more emphasis on less salient verbs	$T_{sverb}/\sqrt{2N_{verb}}$
ndwz	Number of Different Words counted only from the first 50 words in the text	μ_T
ndwerz	Number of Different Words counted from a chunk of the text 50 words long randomly selected	μ_T
ndwesz	Number of Different Words counted from the text divided into 50 words sequences	μ_T
ttr	Type-Token Ratio – The number of word types in the text divided by the number of word tokens	T/N
msttr	Mean Segmental Type-Token Ratio 50 Word Sequence – The average of type-token ratio values calculated from the text divided into chunks of 50 words	μ_{TTR}
cttr	Corrected Type-Token Ratio – The number of word tokens is	$T/\sqrt{2N}$

	adjusted to place more value on the number of word types in the text	
rttr	Root Type-Token Ratio – A similar transformation to cttr but the amount of word tokens is not artificially increased to guard against scale shrinkage near 0	$T/\sqrt{N}$
uber	Uber Index – A transformation of ttr that converts word tokens and types to log values in order to keep them on the same scale and flips the formula to focus on the amount written rather than the types used in the text	$Log^2N/Log(N/T)$
svv1	Squared Verb Variation I – A ratio of all verb types used, squared, and the number of verb tokens used in a text	T^2_{verb}/N_{verb}
cvv1	Corrected Verb Variation I – A transformation of svv1 that reduces the impact of verb tokens rather than increasing the impact of verb types	$T_{verb}/\sqrt{2N_{verb}}$
advv	Adjective Variation – A ratio of all adjective types and adjective tokens used in a text	T_{adj}/N_{lex}
modv	Modifier Variation – A ratio of the modifiers, adjective types and adverb types, and the lexical tokens in a text	$(T_{adj} + T_{adv})/N_{lex}$

Appendix C

Measures of Syntactic Complexity Adapted from Lu (2010)

Code	Description	Formula
W	Words - Counts every word in a text	Frequency
S	Sentences - Counts every sentences, marked by a period in the text	Frequency
VP	Verb Phrase – Counts a structure consisting of a verb and its complements	Frequency
C	Clauses – counts a structure consisting of a subject and finite verb	Frequency
T	T-unit – counts a structure consisting of a main clause plus a subordinate clause or nonclausal structure (Hunt, 1970)	Frequency
DC	Dependent Clause – counts a finite adjective, adverb, or nominal clause	Frequency
CT	Complex T-unit – counts a t-unit that consists of a dependent clause	Frequency
CP	Coordinate Phrase – Counts adjective, adverb, noun and verb phrases around a coordinating conjunction	Frequency
CN	Complex Nominal – Counts a noun phrase consisting of adjectives, possessives, prepositional phrases, relative clauses, participle, or appositive. Nominal clauses, gerunds, and infinitives in the subject are also counted (Cooper, 1976).	Frequency
MLS	Mean Length of Sentence – a ratio of the number of words and the number of sentences in a text	W/S
MLT	Mean Length of T-unit – a ratio of the number of words and the number of T-units	W/T
MLC	Mean Length of Clause – a ratio of the number of words and the number of clauses	W/C
C/S	Sentence Complexity Ratio – number of clauses divided by the number of sentences	C/S
VP/T	Verb Phrase per T-unit	VP/T
C/T	T-unit Complexity Ratio – number of clauses divided by the number of T-units	C/T
DC/C	Dependent Clause Ratio – number of dependent clauses divided by the total number of clauses	DC/C
DC/T	Dependent Clauses per T-unit	DC/T
T/S	Sentence Coordination Ratio – number of T-units divided by the number of sentences	T/S
CT/T	Complex T-unit Ratio – number of complex T-units divided by the number of T-units	CT/T
CP/T	Coordinate Phrases per T-unit	CP/T
CP/C	Coordinate Phrases per Clause	CP/C
CN/T	Complex Nominals per T-unit	CN/T
CN/C	Complex Nominals per clause	CN/C

Appendix D

Writing Placement Test Rubric Version II

Typical volume/ Summary	Level 1: ≤1 page **Needs work at compound sentence and intra-sentence level**	Level 2: **1-1.5 pages** **Needs work at complex sentence &paragraph level**	Level 3: **1-2 pages** **Needs work at essay level**	Level 4: **1-2 or 2+ pages** **Needs work at longer essay level**
Unity and Organization	• Few sentences logically connected • Organization does not show any paragraph-level structure	• Some sentences logically connected • Inadequate paragraph-level coherence & unity • No or little awareness of essay structure	• Sentences are logically organized and clearly connected within a paragraph • some awareness of appropriate essay structure, but paragraphs may not be organized appropriately	• Sentences and paragraphs clearly and logically connected both within and between paragraphs • awareness of essay structure, including the logical ordering of ideas and clear subtopics
Content and Development	• May attempt to offer examples or explanations, but these are underdeveloped and difficult to follow. • Difficulty expressing ideas at the sentence level; trouble articulating main and supporting ideas in coherent sentences	• Main idea may be difficult to identify, or is present but underdeveloped • Demonstrates some recognizable development from beginning to end • Support and explanations may be present but are unclear or not sufficiently detailed	• Main idea easy to identify • Subtopics are present but not fully developed	• Main idea easy to identify • Ideas are developed in a logical and sophisticated manner; author clearly expresses, explains, and supports the ideas
Grammar and Sentence Structure	• Simple and/or compound sentence structures with frequent errors • Frequent errors with basic grammar • Simple tense use	• Simple and compound sentence structures with some errors • Mostly simple tense use; some attempt at other tenses, but often not successfully	• Simple, compound, and complex sentence structures with some errors • Varied tense use • Some errors with advanced grammar	• Sentence structures show clarity and variety, including advanced grammar, with few errors • Advanced tense use
Vocabulary	• Very limited range of vocabulary with frequent errors in word choice and form • Errors in vocabulary use impede comprehensibility	• Narrow range with few errors in word choice and form • Attempts to use some sophisticated vocabulary but makes frequent word choice or word form errors.	• Mostly accurate word forms and word choice, though general and/or repetitive vocabulary use may characterize the writing • Beginning awareness of collocation patterns	• Effective, specific, and varied vocabulary use • Sophisticated range with few errors in word choice and form • Some awareness of collocation patterns