Which Linguistic Features Predict Quality of Argumentative Writing for College Basic Writers, and How Do those Features Change with Instruction?

Charles A. MacArthur

Amanda Jennings

University of Delaware

Zoi A. Philippakos

University of North Carolina Charlotte

Accepted for publication in *Reading and Writing: An International Journal* in April 2018.

Doi: 10.1007/s11145-018-9853-6

The research reported here was supported by the Institute of Education Sciences, U.S. Department of Education, through Grant R305A160242 to University of Delaware. The opinions expressed are those of the authors and do not represent views of the Institute or the U.S. Department of Education.

Abstract

The study developed a model of linguistic constructs to predict writing quality for college basic writers and analyzed how those constructs changed following instruction. Analysis used a corpus of argumentative essays from a quasi-experimental, instructional study with 252 students (MacArthur, Philippakos, & Ianetta, 2015) that found large effects (ES = 1.22) on quality of argumentative writing. Coh-Metrix (McNamara, Graesser, McCarthy, & Cai, 2014) was used to analyze the essays for lexical and syntactic complexity and cohesion. Structural equation modeling found that referential cohesion ($p < .001$) and lexical complexity ($p < .01$) positively predicted quality on posttest essays while syntactic complexity ($p < .001$) was negatively related to quality. Length explained 30% of variance in quality; the full model explained 48.7%. Confirmatory factor analysis was used to impute factor scores for pretest and posttest essays. Analysis of covariance using these factors found that the treatment group wrote posttest essays with greater lexical complexity ($p < .01$) and referential cohesion ($p < .01$) and less use of connectives ($p < .05$) than a business-as-usual control group.

Descriptors: writing; quality; linguistic features; natural language processing; basic writers; formative assessment

Which Linguistic Features Predict Quality of Argumentative Writing for College Basic Writers, and How Do those Features Change with Instruction?

Despite the importance of writing achievement for academic success and for career advancement, large numbers of students graduate from high school without having developed proficiency in writing. The most recent National Assessment of Educational Progress (NAEP) in 2011 (NCES, National Center for Educational Statistics, 2012) found that only 27% of students in the final year of high school performed at or above a proficient level in writing; similar results were reported in 2007 and 2002 (Salahu-Din, Persky, & Miller, 2008). The results also showed dramatic differences by ethnicity and parental education. In grade 12, 27% of white students, but only 9% of black and 12% of Hispanic students scored proficient or better. The problem persists in college. Community colleges offer an opportunity for low income and minority students to attend college, but open access means that many students are under prepared for college writing. In community colleges, according to the National Center for Educational Statistics (NCES, 2013), 40% or more of students are required to take developmental (remedial) courses in writing, reading, and/or math; the numbers are higher for minority students (54% of African-Americans, and 45% of Hispanics). Developmental courses offer underprepared students an opportunity to attend college, but only a minority of students who take such courses complete a degree or certificate program (Bailey, Jeong, & Cho, 2010; Bremer et al., 2013).

Research on the writing performance of students in developmental writing courses is limited. In a review of literature on literacy skills of underprepared college students, Perin (2013) found only five studies that described writing skills. A recent study (Perin, Lauterman, Raufman, & Kalamkarian, 2017) analyzed persuasive essays and summaries of persuasive texts written by 211 students in developmental writing courses. On average, essays were under 200 words and were rated 2.6 on a 7-point holistic quality scale; only half of the sentences were rated as functional to an argument. The summaries only included half (53% on average) of the main ideas in the source article. Only one study (Perin & Lauterman, 2016) has conducted linguistic analysis of writing by students in developmental writing courses (more information below).

The overall goal of the current study was to contribute to greater understanding of the linguistic skills of college basic writers and how they develop over time with instruction. The study used a corpus of argumentative essays from a quasi-experimental study (MacArthur, Philippakos, & Ianetta, 2015) of an instructional program based on strategy instruction, which found large effects (ES = 1.2) on quality of argumentative writing. The study used Coh-Metrix (CM, McNamara, Graesser, McCarthy, & Cai, 2014), an open-access natural language processing (NLP) tool, to analyze writing for theoretically important linguistic constructs. We developed a model of linguistic constructs to predict writing quality on the posttest and then analyzed how those constructs changed over time in response to instruction in treatment and control classes.

Research on Linguistic Analysis to Describe Writing Development and Predict Quality

Early research. Research on linguistic development in writing has found changes with age and expertise in syntactic complexity, vocabulary sophistication, and cohesion. In a seminal article on the development of syntactic complexity in writing, Hunt (1964) designed a new measure, the terminable unit, or t-unit, and used it to document increases in syntactic complexity across grades four, eight, and twelve. In a longitudinal study, Loban (1976) analyzed the oral and written language of high, middle, and low achieving students from grades 1 to 12, documenting

increases in t-unit length with age as well as differences by achievement level. However, another syntactic measure of subordinate clauses increased until grade 8 and then leveled off; Loban (1976) explained that better writers use other methods of subordination, such as participial and prepositional phrases. A review of research (Hudson, 2009) on the linguistic features associated with writing maturity found that both age and writing quality were correlated with syntactic measures of t-unit length and subordination, and with lexical measures of diversity and sophistication.

Turning to research on college students, Witte and Faigley (1981) took linguistic analysis beyond the sentence boundaries of syntactic complexity to look at cohesion across sentences, using Halliday and Hasan's (1976) theoretical framework for categorizing cohesive ties. From a sample of 90 argumentative essays by first-year college students that had been rated for overall quality, they selected five highly rated and five low rated essays. They found greater density of cohesive ties in higher quality essays. Haswell (2000), in a longitudinal study of college students' persuasive essays, found gains across two years in holistic quality, essay length, syntactic complexity (sentence and clause length), free modifiers, and word length.

Automated linguistic analysis. With the development of automated analysis of linguistic features, research on the features of written language has increased. Automated analysis uses natural language processing (NLP) tools to analyze lexical, syntactic, cohesive, and semantic features of text. The most substantial body of research on automated analysis of writing has focused on systems for automated essay scoring (AES) (Shermis, 2014; Shermis & Burstein, 2103). AES uses NLP and machine learning to detect and compute text features that are associated with quality ratings assigned by humans. Typically, these features are combined in a regression-based algorithm to maximize prediction of human essay ratings (Dikli, 2006). In on-demand assessment situations, AES systems have demonstrated interrater reliability with human ratings approximately equal to agreement between two human raters (Shermis, 2014).

Of more relevance to the current study, automated NLP tools have also been used to describe linguistic features of writing that predict overall quality and linguistic features that change with development. McNamara, Crossley, and colleagues (Crossley, Kyle, & McNamara, 2015; Crossley, Weston, Sullivan, & McNamara, 2011; McNamara, Crossley, and McCarthy, 2010; McNamara, Crossley, & Roscoe, 2013) have conducted several studies using Coh-Metrix to analyze writing samples from college students. One study (McNamara et al., 2010) analyzed argumentative essays of 120 first year college students for lexical and syntactic complexity and cohesion. The study found that quality, rated by humans, was predicted by two measures of lexical complexity (lexical diversity, word frequency) and one syntactic measure (number of words before the main verb), accounting for 22% of variance in quality. However, despite including 26 measures of cohesion (indices of referential cohesion and connectives), none of the cohesion measures contributed significantly to the prediction.

A later study (McNamara et al., 2013) with a larger sample of 313 first year college students, first predicted writing quality using the three measures from the earlier study (McNamara et al., 2010), but it only predicted 6% of variance in quality. However, a regression model using the full set of linguistic measures from the earlier study predicted 39% of variance from six linguistic measures. Length was the strongest predictor; other predictors included lexical complexity (word frequency and two measures of word abstractness), and two measures of cohesion. Of the two cohesion measures, overlap of words across sentences was negatively related to quality, while the ratio of given to new information was positively related to quality.

A subsequent study of students in grade 9 through college (Crossley et al., 2015) used exploratory factor analysis to reduce over 200 linguistic indices from Coh-Metrix and two other NLP programs to nine factors, which they then labeled conceptually. Factors of length, lexical complexity, and global cohesion (i.e., links across introduction, body, and conclusion) explained 40% of variance in overall quality. Length was the strongest predictor, and global cohesion was the weakest; similar to Crossley et al. (2011), factors for local cohesive links across sentences were negatively correlated with quality. This study (Crossley et al., 2015) is also of interest methodologically for its use of factor analysis to reduce the large number of available linguistic indices to a manageable and meaningful set of measures.

A cross-sectional comparison of argumentative essays written by students in grades 9 and 11 and the first year of college (N=202) (Crossley et al., 2011) found increases by grade in quality (human ratings), length (words and paragraphs), syntactic complexity (modifiers per noun), and vocabulary complexity (diversity, word frequency, concreteness, and polysemy), but decreases in cohesion measures (word overlap, logical connectives).

Perin and Lauterbach (2016) attempted to replicate the results of McNamara et al. (2010) with argumentative essays from a sample of basic college writers in developmental writing classes. The three linguistic measures from the earlier study did not predict quality significantly ($R^2 = .01$). However, when they included all the measures and followed the methodology of the earlier study, they did find that quality was negatively predicted by single measures of lexical diversity and cohesion. Unfortunately, the measure of lexical diversity, type-token ratio, is known to be strongly correlated with length (McCarthy & Jarvis, 2010), so the finding may reflect primarily a positive correlation of length and quality. As in McNamara et al. (2013) and Crossley et al. (2011), the referential cohesion measure was negatively related to quality. The study (Perin & Lauterbach, 2016) also analyzed students' summaries, finding that quality of summary was positively correlated with lexical complexity.

Overall, the research has found consistent correlations of quality with length and with lexical diversity and complexity, but variable correlations of quality with syntactic complexity and cohesion. In the studies reviewed above, length is consistently the strongest predictor of essay quality (Haswell, 2000; Crossley et al. 2011; 2015; McNamara et al., 2013). Although syntactic complexity is generally correlated with quality with younger students (Hudson, 2009), research with college students is varied. Three studies (Crossley et al., 2011; Haswell, 2000; McNamara et al., 2010) found positive correlations of quality with various measures of syntactic complexity, but three more recent studies (Crossley et al., 2015; McNamara et al., 2013; Perin & Lauterbach, 2016) found no significant relationships. For cohesion, an early study with college students (Witte & Faigley, 1981) found positive correlations with quality, but more recent studies have varied, finding no relationship with quality (McNamara et al., 2010), negative relationships (Crossley et al., 2011; Perin & Lauterbach, 2016), or mixed relationships depending on the measure (Crossley et al., 2015; McNamara et al., 2013). Further research is needed to explore how linguistic features vary based on specific linguistic measures, genres of writing, and the writing proficiency and cultural and linguistic characteristics of students.

Change after instruction. It would also be useful to understand how the linguistic features that predict quality change over time with instruction. Such understanding might guide the design of instruction or the development of better formative assessments. Formative assessment requires measures that are both sensitive to change over relatively short periods of time and also predictive of change in quality (Chapelle, Cotos, & Lee, 2015). A set of studies

addressed this question using a common corpus of writing produced over the course of a semester by second-language (L2) students in writing courses at a college in the United States (Connor-Linton & Polio, 2014); essays had been scored by human raters for overall quality, which improved over time. Polio and Shea (2014) found no change over time in several measures of linguistic errors despite the improvements in quality. Crossley and McNamara (2014) used 11 measures of syntactic complexity from Coh-Metrix, finding six that changed over time and three that predicted quality, but only one that both predicted quality and changed over time. Bulte and Housen (2104) found seven syntactic measures that changed over time including six that also predicted quality. Generally, the studies found increases in syntactic complexity over time for this small group of L2 college students, but no clear pattern of how linguistic changes might account for the gains in quality. To our knowledge, no prior research with native-English-speaking college students has investigated the question of how the linguistic features that change over time with development and instruction correspond to the features that predict quality.

Methodological issues. One of the challenges of using linguistic analysis to predict quality is the very large number of available linguistic indices. Coh-Metrix (McNamara et al., 2014) includes over 100 linguistic indices. One of the studies above (Crossley et al., 2015) addressed this problem by using factor analysis to reduce the large number of available linguistic indices to a manageable and meaningful set of measures. Even with human-scoring, the number of linguistic measures can be very large. The measures used by Haswell (2000) were based on factor analysis of a large set of specific linguistic indices. One other study (Wilson, Roscoe, & Yusra, 2017) with middle school students used confirmatory factor analysis (CFA) and structural equation modeling (SEM) to evaluate whether a model of linguistic factors would predict performance on two statewide writing tests.

Another challenge is posed by the strong correlations between length and quality in writing; as noted above, length was the strongest correlate of quality in all the studies reviewed. Many of the individual linguistic indices are also correlated with length, though some indices have been developed specifically to avoid correlations with length. For example, the MTLD measure of lexical diversity avoids the negative correlation with length of older measures of type-token ratio (McCarthy & Jarvis, 2010). Analyses need to control for this problem; otherwise, observed correlations with quality may be due to common correlations with length, confounding interpretation.

The Current Study

The purpose of the current study was to develop a model of linguistic constructs to predict quality and then to investigate change in those constructs over time in response to instruction. The analysis focused on basic college writers, a large population of struggling writers that is culturally and linguistically diverse.

Using a corpus of pretest and posttest argumentative essays from an experimental study that found strong effects of treatment on quality afforded the opportunity to compare changes for the treatment and control groups on linguistic features that affected posttest quality. The experimental curriculum (MacArthur et al., 2015) was based on self-regulated strategy development (SRSD, Graham, 2006; Harris & Graham, 2009) with a strong emphasis on peer review and self-evaluation (MacArthur, 2016). SRSD has demonstrated strong effects on writing quality for elementary and secondary students (Graham, McKeown, Kuihara, & Harris, 2012;

Graham & Perin, 2007). Students learned systematic strategies for planning, drafting, evaluating, and revising that were based on the purposes and text structure elements of various genres (Englert, Raphael, Anderson, Anthony, & Stevens, 1991). For arguments, students generated reasons on both sides of an issue and used a graphic organizer to select reasons and evidence to support a position along with counterarguments and rebuttals. Students followed this organizer to draft paragraphs with clear topic sentences. Evaluation criteria asked whether reasons were clearly connected to the position and supported by evidence, and whether rebuttals directly addressed opposing reasons. Instruction also included efforts to develop self-regulation strategies for goal setting, task management, progress monitoring, and reflection. Grammar was addressed only in the context of editing. Treatment classes used the curriculum for a semester while control classes continued with business as usual.

Research questions:

1. Can theoretically meaningful latent constructs for lexical complexity, syntactic complexity, and cohesion be used to predict human ratings of quality for college basic writers?
2. Which linguistic constructs change over time as a result of instruction, and are those changes different for an experimental curriculum based on SRSD and typical instruction?

Analysis of linguistic features was conducted using Coh-Metrix (McNamara et al., 2014), using indices of lexical diversity and sophistication, syntactic complexity, and cohesion. At the lexical level, based on prior research, we expected vocabulary diversity and sophistication to increase over time and to contribute to quality. At the syntactic level, length and complexity of syntactic units were expected to increase with improvement in writing. Although prior research with college students has found mixed results, we expected a sample of basic writers would show increased syntactic complexity consistent with findings for less mature students. Given the mixed results for cohesion and quality both with typical college students and basic writing students, our expectations for cohesion were tentative; however, we did expect that the focus of the treatment curriculum on clear organization would lead to increases in cohesive devices.

Structural equation modeling was used to model latent constructs for lexical complexity, syntactic complexity, referential cohesion, and connectives and to predict quality on the posttest essays. Confirmatory factor analysis was used to impute factor scores for pretest and posttest essays, and analysis of covariance was used to investigate the effects of instruction on linguistic changes.

Method

Data Source

As noted earlier, the study used a corpus of pretest and posttest essays from a quasi-experimental study with college basic writers that found large effects of instruction on writing quality (MacArthur et al., 2015). Pretest and posttest essays were available from 252 students in 19 developmental writing classes in two four-year universities in the mid-Atlantic region of the United States, 115 students in 9 classes in the treatment group, and 137 in 10 classes in the control. Of this sample of 252, 54% were female, 52% were White (35% Black, 4% Asian, 6% Latino, and 6% other); 10% of the participants were not native speakers of English though all were considered to have adequate English to participate in an English writing course. There were no significant demographic differences between treatment and control groups ($p > .3$). In the

study (MacArthur et al., 2015), using hierarchical linear modeling with students nested in classes, significant positive effects were found for overall quality of writing on a persuasive essay (ES = 1.22) and for length (ES = .71), but not for grammar.

Human Scoring of Essays

In the first week of class and at the end of the course, students wrote persuasive essays on controversial topics. Persuasive writing was a primary focus in both treatment and control classes as is common in college writing (Wolfe, 2011). At each time, students had a choice of three topics (different at each time) that had been piloted tested in prior research. Prior to quality scoring, spelling errors, but not grammar errors, were corrected to avoid undue bias in raters' judgments of quality (Graham, Hebert, & Harris, 2015). Pretest and posttest essays were mixed, and two graduate students, unaware of the purpose of the study, scored them for overall quality using a 7-point rubric. The holistic rubric directed raters to form an overall judgment of quality based on criteria for ideas or content, organization, word choice, sentence fluency, and errors in grammar and usage. Raters were trained using anchor papers and essays from prior studies (MacArthur & Philippakos, 2013), and both raters scored all 504 papers. Interrater reliability was adequate with a correlation of .82; exact agreement was 52% and agreement within one point was 92%.

Analysis Tool: Coh-Metrix

Analysis of linguistic features was conducted using Coh-Metrix (CM 3.0, McNamara et al., 2014), an open-access program that brings together a range of linguistic analysis tools for syntactic parsing, analysis of lexical characteristics and diversity, latent semantic analysis, and other components. Studies have validated the measures of lexical diversity (McCarthy & Jarvis, 2010) and cohesion and latent semantic analysis (McNamara, Louwerse, McCarthy, & Graesser, 2010). Originally developed to assess text difficulty, a study (Dufty, Graesser, Louwerse, & McNamara, 2006) found that adding CM cohesion measures to traditional measures of text difficulty based on sentence and word length improved prediction of grade level of K-12 textbooks. Research on its application in analysis of writing quality and development was discussed in the introduction.

Selection of Linguistic Indices

CM includes over 100 indices organized by linguistic constructs. Indices were selected to represent four constructs based on theoretical considerations and prior research: lexical complexity, syntactic complexity, and two types of cohesion - referential cohesion and connectives.

Indices for referential cohesion were selected from categories of Referential Cohesion and Latent Semantic Analysis. Referential cohesion refers to links between words across sentences. Such references to words in prior sentences are linguistic cues that can help readers make connections among propositions and ideas (Halliday & Hasan, 1976). The indices in the CM category Referential Cohesion tap overlap in words across sentences, ranging from exact repetition of words to repetition of words with common roots. The indices in Latent Semantic Analysis (LSA) extend the set of links to semantically related words; for example, 'home' in one sentence is semantically related to 'house' and 'furniture' in a later sentence.

Indices for connectives were selected from CM category of Connectives. Connectives are words that make temporal, causal, additive, and other types of connections within a text.

Indices for syntactic complexity were selected from categories of Syntactic Complexity, Syntactic Pattern Density, and Descriptives related to sentences. Syntactic complexity includes indices related to length of nominal phrases and similarity of syntactic structures across sentences. CM does not include traditional syntactic indices such as t-unit length and clauses per t-unit, but Descriptives includes indices related to sentence length. Syntactic Pattern Density includes indices of the relative incidence of types of phrases and word forms, such as noun and verb phrases, gerunds, and passive verb forms.

Indices for lexical complexity were selected from CM categories of Lexical Diversity, Word Information, and Descriptives related to words. Lexical diversity indices measure the number of unique words in comparison to total words. Word Information indices tap the frequency with which words are used as well as several ratings of words such as age of acquisition, concreteness, and imagability, all expected to affect readability.

The numbers of indices in these categories were reduced based on the following criteria: 1) Indices fit the construct based on theory or prior research, for example, indices of word length from the Descriptives category for lexical complexity. 2) Correlation with essay length was less than $r = .2$. As argued in the introduction to this paper, because of the well established correlation between length and quality, it was important to include only indices that were independent of length. 3) Correlation with other indices in the same construct was less than $r = .9$ to avoid problems of co-linearity. 4) Correlation with other indices in the construct was at least $r = .3$. Selection procedures led to the identification of three indices for each construct as shown in Table 1.

[Insert Table 1 approximately here.]

For referential cohesion, overlap of nouns, arguments, and content words were omitted for the broader indices of stem overlap and LSA overlap. Stem overlap includes words with the same root. LSA overlap includes semantically related words, e.g., home and house. These indices are proportional to the number of sentences. Indices for stem overlap in adjacent sentences and across all sentences were used. Of the LSA measures, LSA given/new and LSA across paragraphs were omitted as correlated with length, and LSA across adjacent sentences was co-linear with LSA across sentences within paragraphs. Thus, LSA across adjacent sentences was used.

Connectives contribute to cohesion by providing explicit cues to connections among ideas. Indices expressing logical (e.g., ‘because’), additive (e.g., ‘or’), and adversative (e.g., ‘although’) connections were selected. The index of total connectives was omitted as mathematically related to the separate indices.

For syntactic complexity, three indices were intercorrelated—number of words before the main verb in a sentence, sentence length standard deviation, and syntactic similarity of sentences; this last measure of similarity was negatively correlated with the other two indices. None of these syntactic indices were correlated with length. Length of noun phrases, though theoretically relevant, was correlated with words before the main verb but not the other syntactic measures. Three other indices of ‘minimal edit distance’ were not correlated with other syntax measures, but were correlated with the referential cohesion indices, so we did not include them.

For lexical complexity, CM indices of lexical diversity were all correlated with length except for Measure of Textual Lexical Diversity (MTLD), and MTLD was correlated with word

frequency but not other lexical indices. Three inter-correlated indices were chosen: word length (in syllables), word frequency (logarithm), and age of acquisition.

Data Analysis Procedures

In addressing the study's research questions, two strands of data analysis were employed. The first research question was addressed with a structural equation model (SEM), and the second research question was addressed using imputed factor scores generated from a confirmatory factor analysis (CFA) in an analysis of covariance.

The first research question was addressed using SEM. The model predicted essay quality based on essay length and the four-hypothesized latent linguistic factors: referential cohesion, connectives, syntactic complexity, and lexical diversity. Due to small sample size (N=252), CFI and SRMR were used to determine model fit as RMSEA and TLI tend to over-reject models with small sample sizes (Hu & Bentler, 1999). Under these a priori criteria, models are determined to have adequate fit with CFI $\geq$ .90 and SRMR $\leq$ .07.

The second research question was addressed using a two-step process. In the first step, CFA was used to generate four pre-instruction and four post-instruction latent linguistic factors: pre- and post-referential cohesion, pre- and post-connectives, pre- and post-syntactic complexity, and pre- and post-lexical diversity. In the CFA, strong longitudinal factorial invariance between the pre- and post-instruction pairs was imposed to ensure that the same constructs are measured at both time points and thus are comparable (Widaman, Ferrer, & Conger, 2010). When imposing factorial invariance in a model, it is important to compare model fit of the restricted (factor invariant) model with the fit of the unrestricted (freely loading) model to ensure that imposing factorial invariance does not result in poorer fit. The likelihood ratio chi square difference test was used to test for differences in model fit between the restricted model and the unrestricted model. Due to small sample size (N=252), CFI and SRMR were used to determine model fit as RMSEA and TLI tend to over-reject models with small sample sizes (Hu & Bentler, 1999). Based on our a priori criteria, models are determined to have adequate fit with CFI $\geq$ .90 and SRMR $\leq$ .07. In the second step, both between-group and within-person comparisons were made. Analysis of Covariance (ANCOVA) was used to identify whether post-instruction factor scores differed by treatment condition while controlling for pre-instruction factor scores. The Benjamini-Hochberg procedure was conducted to adjust for multiple comparisons, and adjusted p-values are reported. Paired samples t-tests were used to determine if there were significant differences in pre- and post-instruction factor scores for treatment and control conditions. Estimated marginal means and effect sizes for each comparison are reported for ANCOVA results. Effect-sizes (*ES*) were calculated using partial eta squared for ANCOVAs and r for paired samples t-tests where .1, .3, and .5 indicate small, medium, and large effect sizes respectively (Cohen, 1988).

Results

Linguistic Features as Predictors of Human Quality

Correlations among quality, length, and the linguistic indices for the posttest are shown in Table 2. The strongest correlate of quality is length, with a correlation of .557. The correlations among the three indices in each construct are in bold; all are significant at $p < .001$.

[Insert Table 2 approximately here.]

The SEM modeled four latent post-instruction linguistic factors and essay length as

predictors of human ratings of essay quality. The model is presented in Figure 1. The model had adequate fit based on set a priori criteria (χ^2 = 216.28, *df* = 68, CFI = .90, SRMR = .07). Three linguistic features plus length were significantly associated with quality. Length was significant and positively associated with quality (*B* = .564, *p* < .001). Referential cohesion and lexical complexity were significant and positively associated with quality (*B* = .286, *p* < .001 and *B* = .149, *p* < .01 respectively). Syntactic complexity was significant and negatively associated with quality (*B* = -.310, *p* < .001). Connectives was not significantly associated with quality. Length alone explained 30.0% of the variance in in human-rated quality on the posttest. The complete model increased the variance explained to 48.7%.

[Insert Figure 1 approximately here.]

Changes in Linguistic Features across Time for Treatment and Control Groups

The CFA modeled four pre- post-instruction pairs of latent linguistic factors for a total of eight factors. Each linguistic factor was predicted by three measured variables. The linguistic factors and corresponding measure variables are reported in Table 1. Likelihood ratio chi square test of differences showed insignificant differences in model fit between the restricted (strong factorial invariance) and unrestricted (freely loading) models (χ^2_{diff} = 8.181, df_{diff} = 20, *p* > .99). As there was no statistical difference in model fit, strong factorial invariance was imposed for each pre- post-instruction pair to ensure comparability of the pre- and post-factor scores. Tables 3 and 4 present factor loadings and correlations among factors respectively. The model had adequate fit based on a priori criteria (χ^2 = 485.83, *df* = 244, CFI = .90, SRMR = .06). Factor scores were generated for each of the eight factors. These factor scores were then used in four ANCOVAs, one for each pre- post-instruction pair of latent linguistic factors. All four analyses met the assumption of homogeneity of variance of slopes (Field, 2013). Table 5 presents means for all pre- post-instruction factor pairs.

[Insert Tables 3, 4, and 5 approximately here.]

For referential cohesion, there was a significant difference between the estimated marginal means by treatment condition. The estimated marginal mean for referential cohesion for students in the treatment condition was higher than that of the students in the control condition (*F*(1, 250) = 9.192, *p* = .006, *ES* =.036). Additionally, there was a significant increase in students pre- and post-instruction referential cohesion scores for the treatment group (*t* = 3.19, *p* = .002, *ES* = .29). There was no significant increase for the control group (*t* = .047, *p* = .639, *ES* = .04).

For connectives, there was a significant difference between the estimated marginal means by treatment condition. The estimated marginal mean for connectives for students in the treatment condition was lower than that of the students in the control condition (*F*(1, 250) = 5.287, *p* = .027, *ES* =.021). Additionally, there was a significant decrease in students pre- and post-instruction connectives scores for the treatment group (*t* = -2.46, *p* = .015, *ES* = .22). There was small but insignificant increase for the control group (*t* = .741, *p* = .460, *ES* = .06).

For syntactic complexity, there was no significant difference between the estimated marginal means by treatment condition (*F*(1, 250) = 0, *p* =.984). Additionally, there was a marginally significant increase in students pre- and post-instruction syntactic complexity scores for the treatment group (*t* = -1.95, p = .053, *ES* = .18). There was no significant increase for the control group (*t* = -1.319, *p* = .189, *ES* = .11).

For lexical complexity, there was a significant difference between the estimated marginal means by treatment condition. The estimated marginal mean for lexical complexity for students in the treatment condition was higher than that of the students in the control condition ($F(1, 250) = 11.175$, $p = .004$, $ES = .043$). Additionally, there was a marginally significant increase in students pre- and post-instruction lexical complexity scores for the treatment group ($t = 1.88$, $p = .063$, $ES = .17$). There was a small but insignificant decease for the control group ($t = 1.328$, $p = .186$, $ES = .11$).

Discussion

The purpose of this study was to develop a model of linguistic constructs that predicted writing quality for college basic writers and then to analyze how those same constructs changed over time in response to instruction. The first research question asked which linguistic features predicted quality on the posttest. As found in other studies (e.g., Crossley et al. 2011; 2015; Haswell, 2000; McNamara et al., 2013), length was the strongest predictor of quality on the posttest ($r = .56$). However, it is important to note that length was even more highly correlated with quality on the pretest ($r = .70$) (MacArthur et al., 2015). The average length of papers in the current study increased from about 250 words on the pretest to 650 on the posttest. In other analyses using the same data set (MacArthur & Wilson, 2016), it was found that an automated essay scoring system correlated more highly with human raters on the pretest ($r = .69$) than on the posttest ($r = .55$). The well-known correlation between length and quality presents a potential problem for formative assessment when students have the time, motivation, and skill to write longer papers. It is important to find linguistic factors that predict quality independent of length.

The current study did find a set of linguistic features that predicted quality beyond the effects of length. Length alone predicted 30% of variance in quality on the posttest. The linguistic features increased the variance accounted for to 48.7%. Referential cohesion and lexical complexity were positively associated with writing quality, and syntactic complexity was negatively associated with quality.

The second research question was about the effect of instruction on linguistic features. The adjusted posttest scores were significantly higher for the treatment group than the control group for referential cohesion and lexical complexity, but significantly lower for use of connectives, and not significantly different for syntactic complexity. Consistent with these contrasts, on the change from pretest to posttest, the treatment group scored significantly higher on the posttest for referential cohesion and lower on connectives. No significant changes from pretest to posttest were found for control students on any of the factors. Overall, of the three factors that predicted posttest quality, the treatment group increased more than controls on two (referential cohesion and lexical complexity), and no difference was found on the third (syntactic complexity).

Prior research predicting quality from linguistic features has consistently found that lexical complexity is a significant predictor (Crossley, et al., 2011; 2015; Haswell, 2000; McNamara et al., 2010), and our findings are consistent with those findings. However, our study found a significant, and unexpected, negative relationship between quality and syntactic complexity. A long research tradition prior to automated analysis has generally found increases in syntactic complexity with development (Haswell, 2000; Hudson, 2009; Hunt, 1964). Some research using NLP systems has found positive relationships between syntactic complexity and essay quality (Crossley et al., 2011; McNamara et al., 2010, while other studies have found no

significant relationship (McNamara et al., 2013; Perin & Lauterbach, 2016). We are not aware of any prior studies that found a significant negative relationship.

Also, the current study found a significant positive relationship between referential cohesion and quality, whereas prior research has reported inconsistent results on cohesion. Early research with manual coding (Witte & Faigley, 1981) found a positive association of density of cohesive ties with quality in essays by first-year college students. In contrast, more recent studies with first-year college and high school students using NLP have reported no relationship of cohesion and quality (McNamara et al., 2010), a positive relationship with one measure of cohesion (McNamara et al., 2013), or a negative relationship with local referential cohesion but a positive relationship with global cohesion (McNamara et al., 2015). Perin and Lauterbach's (2016) study with basic college writers found a negative correlation between referential cohesion and quality for essays.

Many factors might explain different findings of this study compared to prior research, and among prior studies themselves. One important factor is the population studied. Basic writers are by definition not proficient writers; as a group, they have difficulties in all aspects of writing from grammar and fluency of production to organization and content (Perin 2013; Perin, Lauterbach, Raufman, & Kalamkarian, 2017). However, there has been only one other study of using automated linguistic analysis with basic writers (Perin & Lauterbach, 2016), and, contrary to the current study, it found no relationship of quality with syntactic complexity and a negative association with cohesion.

Instructional Interpretations and Implications

Another possible explanation is that the highly effective experimental instruction provided to the treatment group accounts for the findings. Quality was predicted for the posttest with the intention of finding linguistic features that might help to explain the positive effects of instruction. The planning and revising strategies that students learned in treatment classes focused on generating ideas and organizing text to meet the purposes of varied genres. The instructional focus on clear organization and coherent connections across text may account for the increased use of cohesive ties in their writing and the fact that those cohesive ties were used effectively in ways that contributed to essay quality.

The curriculum was focused on struggling writers with a belief that they needed to focus on writing coherent essays that clearly convey their purposes and ideas. The negative correlations between cohesion and quality in some prior studies (Crossley et al., 2011; 2015) may indicate that more proficient college writers use more sophisticated techniques for establishing coherence. An interesting parallel is found in research on the effects of cohesion on reading comprehension. Readers with low knowledge of the content benefit from texts with more cohesive ties, whereas readers with high topic knowledge actually better comprehend texts with fewer cohesive ties (McNamara, Kintsch, Songer, & Kintsch, 1996). The theoretical explanation is that knowledgeable readers do not need the extra links and benefit from making more inferences about the content. Similarly, perhaps less proficient writers need to use more cohesive ties to make their writing clear.

The lack of change in syntactic complexity for both groups may be partially explained by the fact that neither treatment nor control instruction placed a major emphasis on grammar (MacArthur et al., 2015). In treatment classes, grammar was addressed as part of editing, with instructors providing brief lessons on problems in student writing. Observations of control

classes found that all instructors provided some grammar instruction, but it was not a major focus in any of the classes. However, the fairly strong negative association between syntactic complexity and quality is difficult to explain. Apparently, simpler syntax with shorter sentences was associated with higher quality. Perhaps, the struggles of basic writers with syntax mean that attempts to write more complex sentences lead to problems of clarity. The findings may have instructional implications, suggesting the need for more effective instruction in sentence construction.

Linguistic complexity was also positively impacted by the curriculum and predicted writing quality. The curriculum did not focus directly on vocabulary or use of more complex terminology. However, it did include instruction in generating ideas and evaluating whether those ideas were clearly expressed in drafts, which may have helped students to refine and expand their vocabulary along with their ideas.

Another possible source of differing results across studies is methodological differences. Two prior studies have used factor analysis to create latent factors. Crossley et al. (2015) found that factors that they labeled length, lexical complexity, and global cohesion were positively related to quality; however, they also found that referential cohesion across sentences was negatively predictive of quality, contrary to our study. Wilson et al. (2017), with middle school students, used theoretically based confirmatory factor analysis and found latent factors for lexical complexity, syntactic complexity, and referential cohesion. As in the current study, the cohesion factor predicted quality.

Another methodological issue that may explain some of the varying findings is that the current study deliberately selected indices that were not correlated with length to avoid any confounds, given the strong correlation between length and quality. If length is not adequately controlled in an analysis, some relationships between linguistic indices and quality might be the result of confounds resulting from joint correlations with length.

Implications for Assessment

The findings about the impact of instruction and the connections to quality have implications for formative assessment. The study did find linguistic factors based on NLP measures that changed as a result of instruction and that contributed to quality. Sensitivity to change over time and prediction of quality ratings are basic requirements for valid formative assessments (Chapelle et al., 2015). In addition, the constructs included in the current study are theoretically important and could be interpreted by instructors and students. Thus, the findings increase confidence that NLP measures can be used to evaluate improvement over time and support instructional recommendations.

Future Research

This study is a beginning analysis of the linguistic features of the writing of college students placed in developmental writing courses. Much further research is needed to confirm and extend the present findings. Studies are needed that include a wider range of students and writing proficiency and a wider range of genres. We think that automated linguistic analysis of writing has potential to contribute to improved formative assessment and to improved understanding of potential targets for instruction.

References

Bailey, T., Jeong, D. W., & Cho, S.-W. (2010). Referral, enrollment, and completion in developmental education sequences in community colleges. *Economics of Education Review, 29*, 255-270. doi: 10.1016/j.econedurev.2009.09.002

Bremer, C. D., Center, B. A., Opsal, C. L., Medhanie, A., Jang, Y. J., & Geise, A. C. (2013). Outcome trajectories of developmental students in community colleges. *Community College Review, 41*, 154-175. doi: 10.1177/0091552113484963

Bulté, B., & Housen, A. (2014). Conceptualizing and measuring short-term changes in L2 writing complexity. *Journal of Second Language Writing, 26*, 42-65. doi: 10.1016/j.jslw.2014.09.005

Chapelle, C. A., Cotos, E., & Lee, J. (2015). Validity arguments for diagnostic assessment using automated writing evaluation. *Language Testing, 32*, 385-405. doi: 10.1177/0265532214565386

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). New York: Academic Press.

Connor-Linton, J., & Polio, C. (2014). Comparing perspectives on L2 writing: Multiple analyses of a common corpus. *Journal of Second Language Writing, 26*, 1-9. doi:10.1016/j.jslw.2014.09.002

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

Crossley, S. A., Kyle, K., & McNamara, D. S. (2015). To aggregate or not? Linguistic features in automatic essay scoring and feedback systems. *Journal of Writing Assessment, 8*. doi:http://journalofwritingassessment.org/article.php?article=80

Crossley, S. A., Weston, J. L., Sullivan, S. T. M., & McNamara, D. S. (2011). The development of writing proficiency as a function of grade level: A linguistic analysis. *Written Communication, 28*, 282-311. doi: 10.1177/0741088311410188

Dikli, S. (2006). An overview of automated scoring of essays. *The Journal of Technology, Learning, and Assessment, 5*(1). doi:

Dufty, D. F., Graesser, A. C., Louwerse, M. M., & mcnamara, D. S. (2006). Assigning grade level to textbooks: Is it just readability? In R. Sun & N. Miyake (Eds.), *Proceedings of the 28th Annual Conference of the Cognitive Science Society* (pp. 1251-1256). Austin, TX: Cognitive Science Society.

Englert, C. S., Raphael, T. E., Anderson, L. M., Anthony, H. M., & Stevens, D. D. (1991). Making writing strategies and self-talk visible: Cognitive strategy instruction in writing in regular and special education classrooms. *American Educational Research Journal, 28*, 337-372. doi: 10.3102/00028312028002337

Field, A. (2013). *Discovering statistics using SPSS Statistics* (4th ed.). Los Angeles: Sage Publications Inc.

Graham, S. (2006). Strategy instruction and the teaching of writing: A meta-analysis. In C. A.

MacArthur, S. Graham, & J. Fitzgerald (Eds.), *Handbook of Writing Research* (pp. 187-207). New York: Guilford.

Graham, S., Hebert, M., & Harris, K. R. (2015). Formative assessment and writing: A meta-analysis. *The Elementary School Journal, 115*, 523-547. doi: 10.1086/681947

Graham, S., McKeown, D., Kiuhara, S., & Harris, K. R. (2012). A meta-analysis of writing instruction for students in the elementary grades. *Journal of Educational Psychology, 104*, 879-896. doi: 10.1037/a0029185

Graham, S., & Perin, D. (2007). A meta-analysis of writing instruction for adolescent students. *Journal of Educational Psychology, 99*, 445-476. doi:10.1037/0022-0663.99.3.445

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

Harris, K. J., & Graham, S. (2009). Self-regulated strategy development in writing: Premises, evolution, and the future. In V. Connelly, A. L. Barnett, J. E. Dockrell, & A. Tolmie (Eds.), *Teaching and learning writing* (*British Journal of Educational Psychology* Monograph Series II: Part 6, pp. 113-135). Leicester, United Kingdom: British Psychological Society.

Haswell, R. H. (2000). Documenting improvement in college writing: A longitudinal approach. *Written Communication, 17*, 307-352. doi: 10.1177/0741088300017003001

Hu, L., & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling*, 6, 1-55. http://dx.doi.org/10.1080/10705519909540118

Hudson, R. (2009). Measuring maturity. In R. Beard, J. Riley, D. Myhill, & M. Nystrand (Eds.), *The Sage Handbook of Writing Development* (pp. 349-362). London, UK: Sage Publishing. doi: 10.4135/9780857021069.n24

Hunt, K. W. (1964). *Differences in grammatical structures written at three grade levels, the structures to be analyzed by transformational methods*. ERIC Document Reproduction Service No. ED003322

Loban, W. D. (1976). Language development: Kindergarten through grade twelve. Research report #18. Urbana, IL: National Council of Teachers of English.

MacArthur, C. A. (2016). Instruction in evaluation and revision. In MacArthur, C. A., Graham, S., & Fitzgerald, J. (Eds.), *Handbook of writing research, 2nd Ed*. (272-287). New York: Guilford.

MacArthur, C. A., & Philippakos, Z. A. (2013). Self-regulated strategy instruction in developmental writing: A design research project. *Community College Review*, 41, 176-195. DOI: http://dx.doi.org/10.1177/0091552113484580

MacArthur, C. A., Philippakos, Z. A., & Ianetta, M. (2015). Self-regulated strategy instruction in college developmental writing. *Journal of Educational Psychology*, *107*, 855-867. doi:10.1037/edu0000011.

MacArthur, C. A., & Wilson, J. (2016, Dec.). *The reliability and validity of an automated essay scoring program for assessment of the outcomes of instruction.* Paper presented at the annual conference of the Literacy Research Association, Nashville, TN.

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

McNamara, D. S., Crossley, S. A., & McCarthy, P. M. (2010). Linguistic features of writing quality. *Written Communication, 27*, 57-86. doi: 10.1177/0741088309351547

McNamara, D. S., Crossley, S. A., & Roscoe, R. (2013). Natural language processing in an intelligent writing strategy tutoring system. *Behavior Research Methods, 45*, 499-515. doi: 10.3758/s13428-012-0258-1

McNamara, D.S., Graesser, A.C., McCarthy, P., & Cai, Z. (2014). *Automated evaluation of text and discourse with Coh-Metrix*. Cambridge: Cambridge University Press. doi: 10.1017/CBO9780511894664

McNamara, D. S., Kintsch, E., Songer, N. B., & Kintsch, W. (1996). Are good texts always better? Text coherence, background knowledge, and levels of understanding in learning from text. *Cognition and Instruction, 14*, 1-43. doi: 10.1207/s1532690xci1401_1

McNamara, D. S., Louwerse, M. M., McCarthy, P. M., & Graesser, A. C. (2010). Coh-Metrix: Capturing linguistic features of cohesion. *Discourse Processes, 42*, 292-330. doi: 10.1080/01638530902959943

National Center for Education Statistics. (2012). The Nation's Report Card: Writing 2011 (NCES 2012–470). Washington, DC: NCES, Institute of Education Sciences, U.S. Department of Education.

National Center for Education Statistics. (2013). 2011-12 National Postsecondary Student Aid Study (NPSAS:12). Washington, DC: NCES, Institute for Education Sciences, U.S. Department of Education.

Perin, D. (2013). Literacy skills among academically underprepared students. *Community College Review, 41*, 118-136. doi: 10.1177/0091552113484057

Perin, D., & Lauterbach, M. (2016). Assessing text-based writing of low-skilled college students. *International Journal of Artificial Intelligence in Education*, online first, https://doi-org.udel.idm.oclc.org/10.1007/s40593-016-0122-z.

Perin, D., Lauterbach, M., Raufman, J., & Kalamkarian, H. S. (2017). Text-based writing of low-skilled postsecondary students: Relation to comprehension, self-efficacy and teacher judgments. *Reading and Writing: An Interdisciplinary Journal, 30*, 887-915. doi: 10.1007/s11145-016-9706-0

Polio, C., & Shea, M. C. (2014). An investigation into current measures of linguistic accuracy in second language writing research. *Journal of Second Language Writing, 26*, 10-27. doi:10.1016/j.jslw.2014.09.003

Salahu-Din, D., Persky, H., & Miller, J. (2008). *The Nation's Report Card: Writing 2007* (No. (NCES 2008–468)). Washington, DC: National Center for Education Statistics, Institute of Education Sciences, U.S. Department of Education.

Shermis, M. D. (2014). State-of-the-art automated essay scoring: Competition, results, and future directions from a United States demonstration. *Assessing Writing, 20*, 53–76. doi: 10.1016/j.asw.2013.04.001

Shermis, M. D., & Burstein, J. (Eds.). (2013). *Handbook of automated essay evaluation: Current applications and new directions*. New York: Routledge.

Widaman, K. F., Ferrer, E., & Conger, R. D. (2010). Factorial invariance within longitudinal structural equation models: Measuring the same construct across time. *Child Development Perspectives*, *4*(1), 10–18. http://doi.org/10.1111/j.1750-8606.2009.00110.x

Wilson, J., Roscoe, R., & Yusra, A. (2017). Automated formative writing assessment using a levels of language framework. *Assessing Writing, 34*. doi: 10.1016/j.asw.2017.08.002

Witte, S. P., & Faigley, L. (1981). Coherence, cohesion, and writing quality. *College Composition and Communication,, 32,* 189-204. doi: 10.1016/j.asw.2017.08.002

Wolfe, C. R. (2011). Argumentation across the curriculum. *Written Communication, 28*, 193-219. doi: 10.1177/0741088311399236

Table 1

Coh-Metrix Indices Used in the Study

Latent Construct	**CM Indices**	**Description**
Referential Cohesion		
	CRFSO1	stem overlap across adjacent sentences
	CRFSOa	stem overlap across all sentences
	LSASS1	semantic overlap across adjacent sentences
Connectives		
	CNCLOGIC	logical connectives
	CNCADC	adversative/contrastive connectives
	CNCADD	additive connectives
Syntactic Complexity		
	SYNSTRUTt [a]	syntactic similarity across all sentences
	SYNLE	mean number of words before the main verb
	DESSLd	sentence length standard deviation
Lexical Complexity		
	WRDFRQa [a]	word frequency for all words
	WRDAOAc	age of acquisition for content words
	DESWLsy	word length in syllables

[a] Index is negatively weighted in the construct.

Table 2

Correlations among Quality, Length, and Coh-Metrix Indices on the Posttest

	Quality	Length	CRFSO1	CRFSOa	LSASS1	CNCLOG	CNCADC	CNCAD	SYNSTt	SYNLE	DESSLd	DESWL	WRDAO
Quality	1												
Length	***.557******	1											
CRFSO1	.250***	0.092	1										
CRFSOa	.170**	0.004	***.866******	***1***									
LSASS1	.291***	.154*	***.764******	***.706******	1								
CNCLOGIC	-.128*	-0.097	0.007	0.045	-0.032	1							
CNCADC	-0.11	-0.072	-0.05	-0.037	-0.118	***.623******	***1***						
CNCADD	-.174**	-0.087	-0.069	-0.052	-.125*	***.368******	***.434******	1					
SYNSTRTt	.201**	-0.115	-0.106	-0.095	-0.042	-0.103	-.177**	-.147*	1				
SYNLE	0.016	.153*	.346***	.300***	.227***	-0.01	0.066	0.033	***-.331******	***1***			
DESSLd	-.148*	0.027	.130*	.235***	.145*	0.059	0.027	.156*	***-.482******	***.479******	1		
DESWLsy	.182**	-0.037	0.076	0.046	-0.085	-0.105	-0.055	0.076	.172**	0.034	-0.085	1	
WRDAOAc	.215***	0.018	.129*	.145*	-0.001	0.003	-0.038	-0.002	0.108	0.028	-0.035	***.675******	***1***
WRDFRQa	-.267**	-0.014	-.128*	-0.073	-.142*	0.048	0.09	-0.07	-.322**	0.033	.145*	***-.641******	***-.383*****

Note: Numbers in bold are correlations within the constructs in the SEM.

Table 3

Factor Loadings for Coh-Metrix Indices

	Pre-Instruction			Post-Instruction		
	λ	S.E.	P-Value	λ	S.E.	P-Value
Referential Cohesion						
CRFSO1	0.96	0.013	.001	0.964	0.013	.001
CRFSOa	0.91	0.015	.001	0.902	0.016	.001
LSASS1	0.726	0.027	.001	0.775	0.025	.001
Connectives						
CNCLOGIC	0.619	0.044	.001	0.699	0.048	.001
CNCADC	0.814	0.054	.001	0.871	0.046	.001
CNCADD	0.49	0.047	.001	0.522	0.047	.001
Syntactic Complexity						
SYNSTRUTt	-0.531	0.052	.001	-0.624	0.049	.001
SYNLE	0.496	0.049	.001	0.597	0.053	.001
DESSLd	0.656	0.056	.001	0.764	0.048	.001
Lexical Complexity						
DESWLsy	0.84	0.04	.001	0.973	0.029	.001
WRDAOAc	0.54	0.038	.001	0.668	0.037	.001
WRDFRQa	-0.582	0.04	.001	-0.667	0.037	.001

Note: Standardized factor loadings are different for pre- and post- factor pairs despite measurement invariance due to standardization.

Table 4

Correlations among Pre- and Post-Instruction Factors

Factor	1	2	3	4	5	6	7
Pre-Referential Cohesion	1						
Post-Referential Cohesion	.321***	1					
Pre-Connectives	-0.048	.06	1				
Post-Connectives	-0.036	-0.031	0.185*	1			
Pre-Syntactic Complexity	.336***	.107	0.23*	.185*	1		
Post-Syntactic Complexity	.007	.285***	0.149	.144	.414***	1	
Pre-Lexical Complexity	-.138	.113	-0.087	.011	-.26**	.048	1
Post-Lexical Complexity	.018	.07	0.031	-0.067	-.093	-.126	.544***

*** $p < .001$; ** $p < .01$; * $p < .05$

Table 5

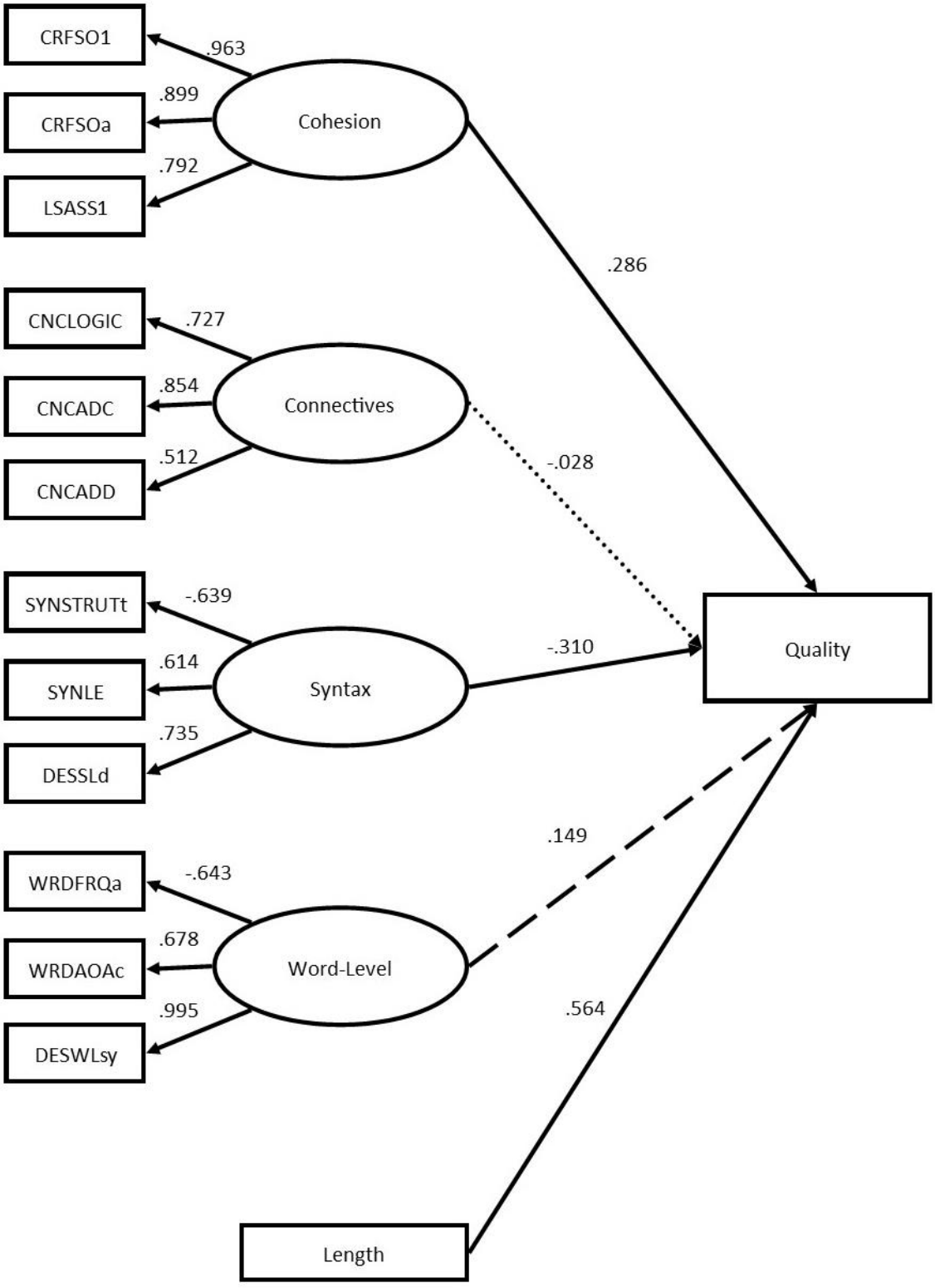
Means for Linguistic Factors by Group and Time

	Treatment (n = 115)			Control (n = 137)		
Factors	Pre M (SD)	Post M (SD)	Adjusted Post M (SE)	Pre M (SD)	Post M (SD)	Adjusted Post M (SE)
Referential cohesion	1.15 (.52)	1.31 †† (.40)	1.31 ** (.40)	1.16 (.43)	1.18 (.34)	1.18 ** (.34)
Connectives	116.9 (36.47)	108.3 † (22.07)	108.0 * (2.32)	112.5 (31.44)	115.0 (27.63)	115.2 * (2.12)
Syntactic complexity	13.89 (4.43)	15.38 [a] (7.22)	15.48 (.53)	14.85 (6.58)	15.58 (4.15)	15.50 (.49)
Linguistic complexity	223.8 (34.0)	229.7 [a] (17.2)	229.9** (1.56)	226.5 (31.33)	223.0 (17.79)	222.8** (1.43)

Note: ANCOVA comparing adjusted posttests: * $p < .05$, ** $p < .01$.

Note: t-tests comparing pretest and posttest within groups: † $p < .05$, †† $p < .01$, [a] $p < .10$

Figure 1. *Structural Equation Model with Post-Instruction Factors as Predictors of Quality*



Note: All parameters estimates are standardized.