Written Communication
28(3) 282–311
© 2011 SAGE Publications
Reprints and permission:
sagepub.com/journalsPermissions.nav
DOI: 10.1177/0741088311410188
http://wcx.sagepub.com
SAGE

The Development of Writing Proficiency as a Function of Grade Level: A Linguistic Analysis

Scott A. Crossley[1], Jennifer L. Weston[2], Susan T. McLain Sullivan[3], and Danielle S. McNamara[2]

Abstract

In this study, a corpus of essays stratified by level (9th grade, 11th grade, and college freshman) are analyzed computationally to discriminate differences between the linguistic features produced in essays by adolescents and young adults. The automated tool Coh-Metrix is used to examine to what degree essays written at various grade levels can be distinguished from one another using a number of linguistic features related to lexical sophistication (i.e., word frequency, word concreteness), syntactic complexity (i.e., the number of modifiers per noun phrase), and cohesion (i.e., word overlap, incidence of connectives). The analysis demonstrates that high school and college writers develop linguistic strategies as a function of grade level. Primarily, these writers produce more sophisticated words and more complex sentence structure as grade level increases. In contrast, these writers produce fewer cohesive features in text as a function of grade level. This analysis supports the notion that linguistic development occurs in the later stages of writing development and that this development is primarily related to producing texts that are less cohesive and more elaborate.

[1]Georgia State University, Atlanta, GA, USA
[2]University of Memphis, Memphis, TN, USA
[3]Oakton High School, Vienna, VA, USA

Corresponding Author:
Scott A. Crossley, Department of Applied Linguistics, Georgia State University, 34 Peachtree St. Suite 1200 One Park Tower Building, Atlanta, GA 30303, USA
Email: sacrossley@gmail.com

Keywords

writing proficiency, cohesion, essay quality, lexical development, syntactic development, computational linguistics, writing assessment

Introduction

Analyzing writing development across grade levels has provided writing researchers with crucial information about how writing skills change as neural, cognitive, and linguistic functions develop (Berninger, Mizokawa, & Bragg, 1991). Learning to write begins with the mastery of producing legible letters and basic spelling (Abbott, Berninger, & Fayol, 2010). Once these skills are attained, young writers work to master basic grammar and sentence structure. During later stages of development, writers begin to focus on text cohesion (McCutchen, 1986; Witte & Faigley, 1981), syntactic structures (Hunt, 1965; McCutchen & Perfetti, 1982), and cognitive strategies such as planning and revising (Abbott et al., 2010; Berninger et al., 1991).

Analyzing writing development as a function of grade level is common in elementary and middle school children (Berninger, Cartwright, Yates, Swanson, & Abbott,, 1994; Hayes & Flower, 1980; O'Donnell, Griffin, & Norris, 1967; Perfetti & McCutchen, 1987) and in high school and college-level students (Freedman & Pringle, 1980; Haswell, 2000; Oliver, 1995). Many researchers focus on the writing development of elementary and middle school children because developmental patterns are strongest at a young age and the opportunity to develop successful interventions are most likely. Research examining the continuing development of writing skills in young adults and adults, although less common, is important for understanding growth in more sophisticated writing strategies that emerge at later stages and how these strategies relate to increasing writing quality.

In this study, like many past studies, we focus on the linguistic factors that develop as writers mature (Berninger et al., 1994; Haswell, 2000; Hayes & Flower, 1980; Perfetti & McCutchen, 1987), but, unlike many previous studies, we focus on writing development in adolescents (9th-grade students, 11th-grade students) and young adults (college freshman). In addition, we do not solely focus on lower levels of linguistic knowledge (i.e., letters, phonemes) but use advanced computational algorithms to analyze deeper level linguistic structure such as lexical sophistication, syntactic complexity, and text cohesion. We also collect human ratings of writing quality for all of the writing samples in our analysis in order to link the development of linguistic features to general writing quality. Thus, our goal in this study is to investigate linguistic development as a function of grade level and writing quality in young adults and adults.

Assessing Writing Skills

Two common approaches for assessing writing skills involve the analysis of neural and linguistic processes important in writing development. These two processes along with cognitive processes are thought to construct the causal mechanisms responsible for writing development constraints (Berninger et al., 1991). The neural level includes skills such as coding of orthographic information, finger movements, and the production of letters. At the linguistic level, word production skills are often automatic, but higher level factors at the word, sentence, and discourse structure levels still constrain the composing process. At the third level, cognitive constraints such as those related to planning, translating, and revising come into play. Neural-level constraints are thought to first develop in Grades 1 through 3, linguistic constraints in Grades 4 through 6, and cognitive constraints in Grades 7 through 9 (Abbott & Berninger, 1993).

Other aspects of writing are also considered to be important when assessing writing skills. For instance, writing skills may be highly constrained by topic knowledge (i.e., the content component), knowledge about how to write (i.e., the discourse component; McCutchen, 1986), or prompt (i.e., the amount of prompt-based information a writer needs or can process; Oliver, 1995). Writing skills may also be constrained by working memory under the assumption that writing skills are strongly related to working memory mechanisms such as storage and processing units for word forms, syntactic processing, phonology, and orthography. Accordingly, expert writers have greater working memory capacity to devote to the writing process. Some theories attribute this capacity to expert writers possessing greater skill and knowledge about language and writing. These working memory mechanisms operate alongside a set of executive functions that allow for the self-government of language (Berninger et al., 2006, 2010; McCutchen, 2000).

Assessment of Writing Quality

A good deal of research has focused on directly assessing writing quality, generally through the use of human raters. Three main approaches are used to assess writing quality: primary trait, analytic, and holistic. Primary trait assessment uses the rhetorical situations (e.g., the purpose, audience, and writing assignment) as the criteria for evaluation. Analytic scoring focuses on individual qualities of a text that correlate to good writing (i.e., content or organization). Last, holistic scoring uses a rater's general impression of a text as an assessment of quality. Holistic scoring has become the default practice for assessing writing quality because it is economical and correlates well with analytic scoring (Huot, 1990).

A common approach to assessing writing quality is through the analysis of linguistic features that characterize proficient writing (e.g., McNamara, Crossley, & McCarthy, 2010; Witte & Faigley, 1981). Much of this research is premised on the notion that the quality of a composition can be partially explained by examining the linguistic structures of a text (Hayes & Flower, 1980; McCutchen & Perfetti, 1982; Perfetti & McCutchen, 1987). Researchers investigating links between linguistic features and writing skills have been less interested in the development of writing skills and more interested in distinguishing which features of writing lead to a higher quality writing sample. This genre of research has tended to focus on the linguistic features of writing because these features are generally the most salient features that can be quantitatively measured. These linguistic features range from simple measures such as the number of words in a text to complex measures such as assessing the levels of intentionality in language through intentional verbs and particles. Assessing linguistic features allows researchers to make links between text properties important in writing quality such as cohesion (Halliday & Hasan, 1976), elaboration (McNamara et al., 2010), abstractness (Hillocks, 2002), sophistication (McNamara et al., 2010), and diversity of ideas (McCarthy & Jarvis, 2010; McNamara et al., 2010; Weston, Crossley, McCarthy, & McNamara, 2011).

Recent developments in computational algorithms have afforded researchers the opportunity to assess large corpora of graded essays to examine overall writing quality. An example of this is a recent study by McNamara et al. (2010) in which they used the computational tool Coh-Metrix (Graesser, McNamara, Louwerse, & Cai, 2004; McNamara & Graesser, in press) to distinguish between high- and low-rated essays written by college freshman using a variety of linguistic features. Their results demonstrated that the three most predictive indices of essay quality were syntactic complexity, lexical diversity, and word frequency. In all cases, higher quality essays contained a greater number of linguistic features associated with text difficulty and linguistic sophistication. So, for instance, higher quality essays were more syntactically complex, contained more lexical diversity, and used more infrequent words. In their study, no linguistic indices related to text cohesion were predictive of essay quality, leading McNamara et al. to conclude that the textual features that characterize good student writing were not aligned with those features that facilitate reading comprehension.

Similar findings have been reported in research concentrating on second-language (L2) writing. Crossley and McNamara (in press), for instance, reported that the linguistic indices that distinguished high-quality L2 essays from low-quality L2 essays were also related to linguistic sophistication and not text cohesion, with high-quality L2 essays containing more lexical diversity, more infrequent words, less meaningful words, less familiar words, and less aspect repetition.

Crossley and McNamara (2010) hypothesized that the coherence of a text may not reside in the incidence of cohesive devices but more likely in their absence. To explore this notion, they examined expert raters' assessments of analytic features of a text (e.g., structure, continuity, strength of thesis, text coherence, grammar). These fine-grained, analytic scores were used to predict the experts' holistic scores for the text using a regression analysis. Crossley and McNamara reported that the expert raters' analytic scores on the coherence of the essay were among the most predictive elements of overall essay quality. However, interestingly, these analytic scores of coherence did not correlate to computational indices of cohesion. Thus, for expert raters, text coherence is an important property of overall quality, but text coherence is not likely the result of the presence or absence of cohesive devices such as word overlap, semantic overlap, and connectives. Such a finding supports the notion that *cohesion* refers to the explicit cues in a text, while *coherence* refers to the understanding that the reader derives from the text, which may be more or less coherent depending on a number of factors, such as prior knowledge and reading skill (McNamara, Kintsch, Butler-Songer, & Kintsch, 1996; O'Reilly & McNamara, 2007).

Writing Development

Research investigating writing development has mostly focused on development as a function of increasing grade level (Akiguet & Piolat, 1996; Berninger et al., 1994; Freedman & Pringle, 1980; Haswell, 2000; Hayes & Flower, 1980; Loban, 1976; Perfetti & McCutchen, 1987). The majority of these studies focus on elementary-aged children and their development of local-level variables such as word reading, spelling, handwriting, punctuation, grammar, copying, letter production, and finger movement (Abbott et al., 2010; Berninger et al., 1992, 2010; Durst, 1984; Grabowski, Weinzierl, & Schmitt, 2010). Some research also focuses on global writing factors such as the ability of elementary school children to write introductions and conclusions (Durst, 1984). These studies generally support the notion that children's neural development is an important factor in learning to write that directly corresponds with the ability to process words in working memory (Berninger et al., 2010).

However, studies have shown that as early as the second-grade writers start to develop past neural-level skills and start to work toward creating more cohesive writing samples through the use of linguistic connections such as referential pronouns and connectives (King & Rentel, 1979). The conventional developments of items related to local coherence emerge around 10 years of age (Akiguet & Piolat, 1996) and continue to develop until around the eighth grade (McCutchen & Perfetti, 1982). The developmental trends demonstrate that essays written by eighth-grade participants contain more local coherence (i.e., connectives) between sentences than essays written by sixth-grade participants (McCutchen, 1986). At the high school and college level, developments in writing cohesion seem to taper off, with Freedman and Pringle (1980) reporting that essays written by graduating high school students and 3rd-year college students do not demonstrate differences in textual unity, organization, development, or coherence. However, such findings should be tempered by the notion that expert writers in the workforce may not heavily depend on text cohesion. In fact, Haswell (1986) found that expert writers in the workforce used fewer referential links (pronouns, demonstratives, and comparatives), lexical overlap, and logical connectors than did college undergraduate writers. This difference in the use of cohesive devices does not lead to expert writing samples that are rated as lower quality but rather to writing samples rated substantially higher. Such findings have been interpreted as expert writers using "unstated structures of expectation to avoid . . . explicit ties that add words without adding new information" (Haswell, 1989, p. 311).

At a later stage in writing, generally with older children who are able to write more coherent texts, we also see the development of more complex syntactic constructions (McCutchen & Perfetti, 1982). However, syntactic development begins quite early with the development of syntactic sophistication noted in children throughout the school years (Loban, 1976). Most studies examining syntactic development generally rely on T-unit (i.e., the main clauses plus additional embedded and subordinating clauses in a sentence) length, demonstrating that T-unit length best predicts the syntactic developments of children from 1st to 12th grades as well as that of skilled adult writers (Hunt, 1965; O'Donnell et al., 1967; Stewart, 1978). Developments in syntactic processing continue to be an important factor in writing proficiency until college and likely afterward (Berninger et al., 2010; Stewart, 1978). For instance, Haswell (2000) found that college juniors tend to write longer sentences with longer clauses than college freshmen.

Students also use more sophisticated words as a function of grade level. For instance, Freedman and Pringle (1980) found no differences in textual unity, organization, development, or coherence between graduating high school student and 3rd-year college student essays, whereas word choice distinguished between the two groups, with 3rd-year college student essays characterized by greater lexical sophistication. Likewise, Haswell (2000) reported that junior-level college students wrote essays containing longer words (a proxy for lexical sophistication) than did freshman-level college students.

Overall, linguistic and cognitive constraints are assumed to influence the writing process at most stages of writing development (Abbott & Berninger, 1993), whereas neural developmental constraints fade by the end of childhood. Linguistically, research demonstrates that development generally occurs first at the word level, then at levels of cohesion, and finally at the syntactic level. Such linguistic changes potentially mirror changes in rhetorical styles as writers move from more chronological and descriptive writing at a young age toward more interpretive and analytical forms of writing in high school (Durst, 1984) and toward more abstract writing in college (Freedman & Pringle, 1980).

Method

This study examines linguistic differences in the writing samples of adolescents and young adult students at different grade levels. We predict that writing styles change in predictable patterns as writers develop proficiency. As McNamara et al. (2010) demonstrated, writers begin to use more sophisticated linguistic devices such as more complex syntactic structures, a greater variety or words, and less frequent words as their proficiency increases. We support our prediction by conducting a variety of computational analyses using the computational tool Coh-Metrix (Graesser et al., 2004) on a corpus of scored essays collected from three different groups of learners (9th-grade writers, 11th-grade writers, and college freshman writers). To link increasing writing proficiency with increasing grade level, we examine whether significant differences in essay quality are observed between the grade levels. We then use the model reported by McNamara et al. to predict the variance in the essay scores to assess the model's extendibility and validity. Last, we examine linguistic differences among the essays as a function of grade level to investigate whether linguistic features other than those reported by McNamara et al. are important indicators of essay quality across grade levels.

Corpus

We collected argumentative essays written by students at three different education levels: 9th grade, 11th grade, and college freshman. All of the essays were written in response to prompts used in the Scholastic Achievement Test (SAT) writing section. The essays at each level were part of the course requirements in a writing class. The ninth-grade writers and college freshman were assigned one of two prompts. The 11th-grade writers all wrote on the same prompt. There was no overlap in prompts within the groups. The prompts were general knowledge prompts that did not require domain knowledge and were meant to induce a variety of ideas. Students were allowed 25 minutes to write the essay. We collected 62 essays from 62 9th-grade writers, 70 essays from 70 11th-grade writers, and 70 essays from 70 college freshmen. We collected the essays from three different geographic areas. The ninth grade essays were collected from a suburban school district in upstate New York. The 11th-grade essays were collected from a suburban school in Washington, DC. The college freshman essays were collected from students attending Mississippi State University. Our corpus differs from the corpus analyzed in McNamara et al. (2010) in that the essays were timed, the prompts were general knowledge, and students of varying grade levels wrote the essays. Descriptive statistics for the corpus are presented in Table 1.

We separated the corpus into two sets: a training set ($n = 135$) and a testing set ($n = 67$) based on a 67/33 split. The training set was used to select the linguistic variables for the initial statistical analysis. The test set was used to calculate the classification accuracy of the selected variables in an independent corpus (Witten & Frank, 2005). Such a method allows us to predict accurately the performance of our model on an independent corpus.

Essay Evaluation

Expert raters scored the essays in our corpus to assess writing quality. Two raters with at least 4 years of experience teaching freshman composition courses at a large university scored each of the 202 essays in the corpus. The raters evaluated the essays based on a standardized rubric commonly used in assessing SAT essays. The rubric (see appendix) was used to holistically assess the quality of the essays and had a minimum score of 1 and a maximum score of 6. Raters were informed that the distance between each score was equal. Accordingly, a score of 5 is as far above a score of 4 as a score of 2 is above a score of 1. The raters were first trained to use the rubric with 20 similar essays taken from another corpus. A Pearson correlation for each

Table 1. Corpus Description

Grade level	*N*	Prompt
Ninth	34	In many circumstances, optimism—the expectation that one's ideas and plans will always turn out for the best—is unwarranted. In these situations, what is needed is not an upbeat view but a realistic one. There are times when people need to take a tough-minded view of the possibilities of success, give up, and invest their energies elsewhere rather than find reasons to continue to pursue the original project or idea.
Ninth	28	Many persons believe that to move up the ladder of success and achievement, they must forget the past, repress it, and relinquish it. However, others have just the opposite view. They see old memories as a chance to reckon with the past and integrate past and present.
Eleventh	70	Many people believe that the choices we make, rather than our abilities and talents, show who we truly are.
Freshman	38	Having many admirers is one way to become a celebrity, but it is not the way to become a hero. Heroes are self-made. Yet in our daily lives, we see no difference between "celebrities" and "heroes." For this reason, we deprive ourselves of real role models. We should admire heroes—people who are famous because they are great—but not celebrities—people who simply seem great because they are famous.
Freshman	32	We value uniqueness and originality, but it seems that everywhere we turn, we are surrounded by ideas and things that are copies or even copies of copies. Writers, artists, and musicians seek new ideas for paintings, books, songs, and movies, but many sadly realize, "It's been done." The same is true for scientists, scholars, and businesspeople. Everyone wants to create something new, but, at best, we can hope only to repeat or imitate what has already been done.

essay evaluation was conducted between the raters. Once the raters reached a correlation of $r = .70$ (which was significant at $p < .001$), the ratings were considered reliable. After training, each rater scored all the essays in the corpus. The final interrater reliability for all essays in the corpus was $r > .70$. We used the mean score between the raters as the holistic value for the quality of each essay.

Coh-Metrix

To analyze the linguistic qualities of the essays, we used the computational tool, Coh-Metrix. Coh-Metrix is an automated text analysis tool that provides a large array of linguistic indices important in text cohesion, text sophistication, and text readability (Graesser et al., 2004; McNamara & Graesser, in press). We selected indices from Coh-Metrix with theoretical and empirical links to essay quality and writing proficiency. We organized these indices into broad measures that reflected general linguistic constructs. These linguistic constructs included cohesion, lexical sophistication, syntactic complexity, and text structure. Our cohesion measures included causality, incidence of connectives, incidence of logical operators, lexical overlap, semantic coreferentiality, and anaphoric reference. Our measures of lexical sophistication included word hypernymy, word polysemy, lexical diversity, word frequency, and word information indices (e.g., word concreteness, familiarity, meaningfulness, and imageability). Our measures of syntactic complexity included syntactic similarity and phrase structure complexity. Our text structure measures included indices such as word length, number of paragraphs, and number of sentences. Each of these measures and their related Coh-Metrix indices are discussed below in greater detail.

Causal cohesion. Causal cohesion is measured in Coh-Metrix by calculating the ratio of causal verbs (e.g., *break, kick, cause*) to causal particles (*because, hence, so, therefore*). The causal verb count is based on the number of main causal verbs identified through WordNet (Fellbaum, 1998; Miller, Beckwith, Fellbaum, Gross, & Miller, 1990). Causal cohesion is relevant to texts that depend on causal relationships between events and actions. Causal cohesion also helps to create relationships between sentences or clauses (Pearson, 1974-1975).

Connectives. Coh-Metrix assesses connectives on two dimensions. The first dimension contrasts positive versus negative connectives, whereas the second dimension is associated with particular classes of cohesion identified by Halliday and Hasan (1976) and Louwerse (2001). These connectives are associated with positive additive (*also, moreover*), negative additive (*however, but*), positive temporal (*after, before*), negative temporal (*until*), and causal (*because, so*) measures. Connectives play an important role in the creation of cohesive links between ideas and clauses (Crismore, Markkanen, & Steffensen, 1993; Longo, 1994) and provide clues about text organization (van de Kopple, 1985).

Logical operators. The logical operators measured in Coh-Metrix include variants of *or*, *and*, *not*, and *if-then* combinations, all of which have been shown to relate directly to the density and abstractness of a text and correlate to higher demands on working memory (Costerman & Fayol, 1997).

Lexical overlap. Coh-Metrix considers four forms of lexical overlap between sentences: noun overlap, argument overlap, stem overlap, and content word overlap. Noun overlap measures how often a common noun of the same form is shared between two sentences. Argument overlap measures how often two sentences share nouns with common stems (including pronouns), while stem overlap measures how often a noun in one sentence shares a common stem with other word types in another sentence (not including pronouns). Content word overlap refers to how often content words are shared between sentences (including pronouns). Lexical overlap has been shown to aid in text comprehension (Douglas, 1981; Kintsch & van Dijk, 1978; Rashotte & Torgesen, 1985).

Semantic coreferentiality. Coh-Metrix measures semantic coreferentiality using Latent Semantic Analysis (LSA), a mathematical and statistical technique for representing deeper world knowledge based on large corpora of texts. LSA uses a general form of factor analysis to condense a large corpus of texts down to 300-500 dimensions. These dimensions represent how often a word occurs within a document (defined at the sentence level, the paragraph level, or in larger sections of texts) and each word, sentence, or text is represented by a weighted vector. The relationships between the vectors form the basis for representing semantic similarity between words, an important indicator of cohesion (Landauer, McNamara, Dennis, & Kintsch, 2007).

Anaphoric reference. Coh-Metrix measures anaphoric links between sentences by comparing pronouns to previous references. For instance, if the current sentence contains a pronoun, Coh-Metrix calculates if previous sentences contain likely noun and pronoun references that agree in number (singular/plural), gender (male/female), and person (human/nonhuman) with the given pronoun. Anaphoric references are important indicators of text cohesion (Halliday & Hasan, 1976).

Polysemy. Coh-Metrix measures word polysemy (the number of senses a word has) through the WordNet computational, lexical database (Fellbaum, 1998). In WordNet, more than 170,000 English nouns, verbs, adjectives, and adverbs are organized in lexical networks based on connections between related lexical concepts. Polysemy relationships in WordNet are based on synsets (i.e., groups of related lexical items), which are used to represent similar concepts but distinguish between synonyms and word senses (Miller et al., 1990). These synsets allow for the differentiation of senses and provide a basis for examining the number of senses associated with a word. Coh-Metrix reports the mean WordNet polysemy values for all content words in a text. Word polysemy is indicative of text ambiguity because the more senses a word contains relates to the potential for a greater number of lexical interpretations.

Hypernymy. Coh-Metrix also uses WordNet to report word hypernymy (i.e., word specificity). In WordNet, each word is located on a hierarchical scale allowing for the measurement of the number of subordinate words below and superordinate words above the target word. Thus, *entity*, as a possible hypernym for the noun *chair*, would be assigned the number 1. All other possible hyponyms of entity as it relates to the concept of a chair (e.g., *object*, *furniture*, *seat*, *chair*, *camp chair*, *folding chair*) would receive higher values. Similar values are assigned for verbs (e.g., *hightail*, *run*, *travel*). As a result, a lower value reflects an overall use of less specific words, while a higher value reflects an overall use of more specific words.

Lexical diversity. Lexical diversity indices generally measure the number of types (i.e., unique words occurring in the text) by tokens (i.e., all instances of words), forming an index that ranges from 0 to 1, where a higher number indicates greater diversity. Traditional indices of lexical diversity are highly correlated with text length and are not reliable across a corpus of texts where the token counts differ markedly. To address this problem, a wide range of more sophisticated approaches to measuring lexical diversity have been developed. Those reported by Coh-Metrix include MTLD (McCarthy & Jarvis, 2010) and D (Malvern, Richards, Chipere, & Duran, 2004). Lexical diversity measures relate to the number of words that a writer knows.

Word frequency. Word frequency indices measure how often particular words occur in the English language. The indices reported by Coh-Metrix are taken from CELEX (Baayen, Piepenbrock, & Gulikers, 1995), a 17.9 million-word corpus. Coh-Metrix reports frequency values for all words in the text and for content words only. Coh-Metrix also reports frequency values taken from the entire CELEX corpus (including both written and spoken texts) and for the spoken subset corpus contained in CELEX, which consists of 1.3 million spoken tokens. The Coh-Metrix indices report a value for all the word tokens in the text except those not contained in the CELEX database. If a word in a text is not included in the CELEX corpus, it is not computed in the Coh-Metrix indices. The frequency indices are calculated using a mean logarithm (to the base of 10). Word frequency is an important indicator of lexical knowledge.

Word information measures. Coh-Metrix calculates word information indices from the MRC Psycholinguistic Database (Wilson, 1988). These indices include concreteness, familiarity, imageability, and meaningfulness. Words that reference an object, material, or person generally receive a higher concreteness score than an abstract word. More familiar words are more readily recognized but not necessarily more frequent (compare *eat* to *while*). A highly imageable word such as *cat* evokes images easily and is thus scored more highly than a word such as *however*, which produces a mental image with difficulty. Words with high meaningfulness scores are highly associated with other words (e.g., *people*), whereas a low meaningfulness score indicates that the word is weakly associated with other words. All of these indices are important indicators of word knowledge.

Syntactic complexity. Syntactic complexity is measured by Coh-Metrix by calculating the mean number of words before the main verb, the mean number of high-level constituents (defined as sentences and embedded sentence constituents) per word, and the average number of modifiers per noun phrase. Sentences with difficult syntactic constructions include the use of embedded constituents and are often structurally dense, syntactically ambiguous, or ungrammatical (Graesser et al., 2004). As a consequence, they are more difficult to process and comprehend (Perfetti, Landi, & Oakhill, 2005).

Syntactic similarity. Coh-Metrix assesses syntactic similarity by measuring the uniformity and consistency of the syntactic constructions in the text. This index not only looks at syntactic similarity at the phrasal level but also takes account of the parts of speech involved. More uniform syntactic constructions result in less complex syntax that is easier for the reader to process (Crossley, Greenfield, & McNamara, 2008).

Basic text measures. Coh-Metrix reports a variety of basic indices related to the text. These include indices of text length, number of paragraphs in the text, and average number of sentences per paragraph. Text length and structure (i.e., number of paragraphs) are important indices of essay quality.

Statistical Analyses

We used analysis of variance (ANOVA) to investigate if there were significant differences between the grade levels in terms of the human scores of writing quality. We then examined the strength of the regression model reported in McNamara et al. (2010) for predicting essay quality in our corpus. Last, we conducted ANOVAs and a discriminant function analysis (DFA) to examine which linguistic indices were the best predictors of essay grade level for the essays in our current corpus. For this analysis, we used the significance values from the ANOVA to select the variables to be used in the DFA. Only those variables that demonstrated significant differences between the three groups (9th-grade writers, 11th-grade writers, and college freshmen) were selected. We also used ANOVAs to examine the effects of the prompts on the production of linguistic features to ensure that differences among grade levels were a function of the grade level and not the prompt. To check for multicollinearity, we conducted correlations and tolerance tests on the selected variables. If the variables did not exhibit prompt effect differences and collinearity, they were then included in the DFA. We used the DFA to predict which variables were most predictive of grade level. In order to allow for stable interpretation of the DFA by avoiding overfitting, we ensured that there were at least 15 times more cases (the essays in the training set) than variables (the selected Coh-Metrix indices). A 15 to 1 ratio allows for the interpretation of the discriminant function coefficients as predictors for each variable's individual contribution to the discriminant function (Field, 2005).

Results

Essay Quality

To assess differences in the quality of the essays among the grade levels, we conducted an ANOVA to compare the human scores of essay quality on the entire corpus. The ANOVA showed significant differences in human assessments of essay quality among the levels of writers, $F(2, 201) = 80.056$, $p < .001$, $\eta^2 = .446$. Essays written by 9th-grade participants received the lowest score ($M = 1.653$; $SD = 0.766$) followed by 11th-grade essays ($M = 2.979$; $SD = 1.133$) and essays written by college freshmen ($M = 3.757$, $SD = 0.924$). To ensure that prompt differences did not affect holistic scoring, we also conducted *t* tests between the holistic scores for each essay at the 9th grade and freshman level based on the essay prompt. These analyses demonstrated no significant differences between the holistic scores of the essays as a function of prompt. Overall, these analyses support the hypothesis that writing quality increases as a function of grade level.

Regression Model

We used the following regression model as reported by McNamara et al. (2010) to predict the amount of variance in the essay scores:

$$\begin{aligned}15.81 &+ (\text{Number of words before main verb value} \times .289)\\ &+ (\text{Measure of text and lexical diversity value} \times .002)\\ &+ (\text{CELEX frequency for all words value} \times -4.442)\end{aligned}$$

The model yielded $r = .442$ and $r^2 = .196$. This finding demonstrates that the McNamara et al. model explains 20% of the variance found in the human scores of essay quality for our corpus. This finding is similar to the findings of McNamara et al.'s study in which the model explained 22% of the variance in the *test* set of essays.

Variable Selection for Discriminant Function Analysis

We conducted ANOVAs to examine which indices from the Coh-Metrix measures (e.g., lexical overlap, word frequency, syntactic complexity) were the best predictors of essay grade level. We selected the variables for which there were significant differences as a function of the writers' grade level and demonstrated the highest partial eta square value when comparing differences between the grade levels. In order to avoid issues of overfitting (see above), we were limited to selecting only nine indices (one predictor per 15 cases).

The measures that reported indices with the largest effect values were basic *text measures* (number of words and number of paragraphs), *word information* (word familiarity and word concreteness), *word hypernymy* (average hypernymy all words), *word frequency* (CELEX frequency for all words), *connectives* (incidence of positive logical connectives), *lexical diversity* (*D*), *syntactic complexity* (average number of modifiers per noun phrase), *anaphoric resolution* (weighted anaphor back 10 sentences), *lexical overlap* (content word overlap), and *word polysemy* (average polysemy).

Prompt Effect

We conducted ANOVAs between prompts at the ninth-grade and college freshmen levels for all significant variables in order to ensure that differences among the grade levels were not the results of prompt. For the ninth-grade prompts, two indices (*average hypernymy all words* and *weighted anaphor back 10 sentences*) demonstrated significant differences between the ninth-grade prompts. For the college freshman essays, two indices (*average hypernymy all words* and *word familiarity*) demonstrated significant differences between the college freshman prompts. To ensure that prompt-based differences for these variables did not influence our discriminant analysis, these three indices were excluded from the analysis.

Collinearity

To check for collinearity between the selected indices, we conducted Pearson correlations and tolerance checks. Assessing multicollinearity is important because variables that are highly correlated make interpretation of the findings difficult because redundant variable mask the contribution of individual variables (Brace, Kemp, & Snelgar, 2006; Tabachnick & Fidell, 2001). None of the remaining variables was highly correlated ($r > .70$). The tolerance checks demonstrated that all Variance Inflation Factors (VIF) values and tolerance levels were about 1, indicating that the model data did not suffer from multicollinearity (Field, 2005). Descriptive and ANOVA statistics for each of the final variables selected for the DFA analysis are presented in Table 2.

Pairwise Comparisons

A series of pairwise comparisons were conducted to examine specific differences among essays written by participants in the 9th grade, 11th grade, and college for each selected Coh-Metrix index. Those results are reported below.

Number of words. All grade levels demonstrated significant differences from one another in terms of the number of words in the text. The ninth-grade essays contained the fewest number of words while essays written by college freshman contained the most words.

Word frequency. Ninth-grade essays showed significant differences from college freshman essays in terms of the word frequency values of the text. No significant differences were found between 11th-grade essays and 9th-grade essays or 11th-grade essays and college freshman essays. The ninth-grade essays contained the most frequent words while college freshman essays contained the least frequent words.

Positive logical connectives. Ninth-grade essays demonstrated significant differences with 11th-grade and college freshman essays in terms of positive logical connective (e.g., *and*, *also*, *then*, *in sum*, *next*), but no differences were found between 11th-grade essays and college freshman essays. Ninth-grade essays contained significantly more logical positive connectives than both 11th-grade and college freshman essays.

Word concreteness. All grade levels showed significant differences from one another in terms of word concreteness values. The ninth-grade essays contained the least concrete words while essays written by college freshmen contained the most concrete words.

Lexical diversity (D). All grade levels demonstrated significant differences from one another in terms of the variety of words used in the text. The ninth-grade essays contained the least variety of words while essays written by college freshmen contained the greatest variety of words.

Number of modifiers per word phrase. All grade levels demonstrated significant differences from one another in terms of the average number of modifiers per noun phrase. The ninth-grade essays contained the fewest number of

Table 2. Analysis of Variance (ANOVA) Results for Selected Linguistic Indices Among Grade Levels: Means, Standard Deviations, *f* Value, *p* Value, and hp2 partial eta squared

	9th grade	11th grade	College freshman	*f* value	*p* value	hp2
Number of words	178.614 (83.339)	287.745 (92.623)	384.205 (117.251)	47.844	<.001	0.420
Word frequency—all words	3.239 (0.087)	3.207 (0.104)	3.124 (0.083)	18.254	<.001	0.217
Incidence of positive logical connectives	45.267 (21.582)	25.912 (12.299)	30.113 (13.136)	17.844	<.001	0.213
Word concreteness—content words	327.539 (17.930)	340.119 (18.820)	350.712 (18.148)	17.653	<.001	0.211
Lexical diversity *D*	55.114 (21.690)	73.234 (25.927)	86.318 (27.179)	17.207	<.001	0.207
Modifiers per noun phrase	0.583 (0.160)	0.690 (0.173)	0.764 (0.132)	14.922	<.001	0.184
Content word overlap	0.166 (0.071)	0.148 (0.046)	0.107 (0.045)	13.025	<.001	0.165
Word polysemy	4.562 (0.612)	4.402 (0.547)	4.074 (0.386)	9.891	<.001	0.130
Number of paragraphs	2.318 (1.459)	3.809 (1.227)	4.909 (4.724)	8.768	<.001	0.117

modifiers per noun phrase while essays written by college freshmen contained the most modifiers per noun phrase.

Content word overlap. All grade levels demonstrated significant differences from one another in terms of the content word overlap. The ninth-grade essays contained the most content word overlap while essays written by college freshmen contained the least amount of content word overlap.

Word polysemy. Ninth-grade essays showed significant differences from college freshman essays in word polysemy values of the text. No significant differences were found between 11th-grade essays and 9th-grade essays or 11th-grade essays and college freshman essays. The ninth-grade essays contained the most polysemous words while college freshman essays contained the least polysemous words.

Number of paragraphs. Ninth-grade essays showed significant differences from 11th-grade and college freshman essays in the average number of paragraphs per essay. No significant differences were found between 11th-grade essays and college freshman essays. The ninth-grade essays contained the fewest paragraphs per essay while essays written by college freshmen contained the most paragraphs.

Accuracy of Model

To assess the accuracy of these linguistic indices to distinguish among grade levels, we conducted a discriminant function analysis. A discriminant function analysis is a statistical procedure that is able to predict group membership (the grade level of the essays) using a series of independent variables (the selected Coh-Metrix variables). The training set is used to generate a discriminant function. The discriminant function acts as the algorithm that predicts group membership. Later, the discriminant function analysis model from the training set is used to predict group membership of the essays in the test set. If the results of the discriminant analysis are statistically significant, then the findings support the predictions of the analysis (that linguistic differences exist between essays of different grade levels and that those differences can be used to classify the essays based on the grade level). We report the findings using an estimation of the accuracy of the analysis, which is made by plotting the correspondence between the groupings in the testing and training sets and the predictions made by the discriminant analysis.

The results demonstrate that the discriminant analysis correctly allocated 95 of the 135 essays in the training set ($df = 4$, $n = 135$; $\chi^2 = 95.325$, $p < .001$) for an accuracy of 70.4% (chance for this analysis is 33%). The reported

Table 3. Classification Results for Grade-Level Analyses

Predicted membership			9th grade	11th grade	College freshman	Total
Training set	Count	9th grade	34	9	1	44
		11th grade	7	29	11	47
		College freshman	1	11	32	44
	Percentage	9th grade	77.3	20.5	2.3	100
		11th grade	14.9	61.7	23.4	100
		College freshman	2.3	25	72.7	100
Test set	Count	9th grade	17	1	0	18
		11th grade	4	15	4	23
		College freshman	2	5	19	26
	Percentage	9th grade	94.4	5.6	0	100
		11th grade	17.4	65.2	17.4	100
		College freshman	7.7	19.2	73.1	100

weighted kappa between the DFA classification and the actual classification is .644, indicating a substantial agreement. For the test set, the discriminant analysis correctly allocated 51 of the 67 essays ($df = 4$, $n = 67$; $\chi^2 = 58.104$, $p < .001$) for an accuracy of 76.1% (chance for this analysis is also 33%). The reported weighted kappa between the DFA classification and the actual classification is .712, indicating a substantial agreement. Results for the analysis are presented in Table 3.

As is also common, we report these results in terms of recall and precision. Recall scores are computed by tallying the number of hits over the number of hits + misses. Precision is the number of correct predictions divided by the number of incorrect predictions. This distinction is important because if an algorithm predicted everything to be a member of a single group, it would score 100% in terms of recall but could only do so by claiming members of the other group. If this happened, then the algorithm would score low in terms of precision. By reporting both values, we can better understand the accuracy of the model. The accuracy of the model for predicting the grade level of the essays can be found in Table 4. The combined accuracy of the model (F1) for the training set was .707. The combined accuracy for the test set was .759. The results provide strong evidence that the linguistic features of essays can be used to classify essays in terms of the likely grade level of the writer.

Table 4. Precision and Recall Finding (Training and Test Set)

Grade level	Precision	Recall	F1
Training set			
9th grade	0.810	0.773	0.791
11th grade	0.592	0.617	0.604
College freshman	0.727	0.727	0.727
Test set			
9th grade	0.714	0.833	0.769
11th grade	0.682	0.652	0.667
College freshman	0.875	0.808	0.840

Discussion

This study has demonstrated that linguistic differences at the word level, at the syntactic level, and at the level of cohesion distinguish essays written at the 9th grade, 11th grade, and college level. This finding lends support to the notion that writing development at the linguistic level continues into college and that differences for advanced levels of writers are not strictly syntactic (Hunt, 1965; Loban, 1976; O'Donnell et al., 1967) but also word (Freedman & Pringle, 1980; Haswell, 2000) and cohesion based (Akiguet & Piolat, 1996; Haswell, 1986, 2000). Overall, the findings lend support to the McNamara et al. (2010) study, which found that higher quality essays were the result of differences in linguistic sophistication. However, unlike the McNamara et al. study, our study shows that cohesion features also play a prominent role in distinguishing grade levels. However, it is the lack of cohesive devices that characterize the college-level writer as compared with the 9th- and 11th-grade writers. Such a finding provides support for the notion that writers develop linguistic strategies that focus on the sophistication of linguistic features in the text as compared with text cohesiveness. We discuss these findings below as well as provide essay examples to highlight the difference fleshed out in our analyses.

Not surprisingly, our strongest predictor of grade level was the number of words in a text. Longer texts (and, by proxy, those containing the greatest number of paragraphs) afford writers the opportunity to elaborate sufficiently on topics and arguments in their essays and enhance central ideas, all characteristics of proficient writers (NAEP Writing Assessment, Institute of Education Sciences, 2003).

Our strongest lexical predictor was word frequency. Word frequency was one of many lexical-based indices that distinguished among grade level. The other indices included lexical diversity, word concreteness, and word polysemy. In all cases, the lexical elements of higher level writers tended toward greater sophistication (in the case of word frequency and lexical diversity) or toward less textual ambiguity (in the case of word concreteness and polysemy). Our frequency index demonstrated that more advanced writers used less frequent words, while our lexical diversity index showed that more advanced writers produce a greater variety of words. Thus, more advanced writers produce more challenging texts by producing a greater variety of words that are less accessible to the reader (Haberlandt & Graesser, 1985). The concreteness index showed that more advanced writers used more concrete words while the polysemy index supported the notion that advanced writers produced texts with less ambiguous words. Thus, while advanced writers produce texts that are more difficult to process for low-level readers, they also produce texts that are more concrete and less ambiguous.

Our analysis also demonstrated that cohesive devices are important indicators of writing quality. However, more cohesive texts are produced by less skilled writers. These texts are also scored lower using holistic scales. For instance, ninth-grade writers are more likely to produce texts with a higher incidence of positive logical connectives and more content word overlap. These cohesive features should make the text more readable, but, conversely, the text is assessed to be of a lower quality. One likely reason that raters judge more cohesive texts to be of lower quality is because our raters are expert raters and likely high-knowledge readers (McNamara, 2001). High-knowledge readers do not benefit from explicit text connections because their knowledge base allows them to make meaning inferences. Cohesion gaps in texts can be beneficial to comprehension because they induce readers to generate inferences, which are successful for readers with more knowledge about the domain. The generation of such inferences aids memory and learning because they connect directly to the reader's representation of text. It is likely that more advanced writers are higher knowledge readers and thus write in a manner that matches their knowledge base. Thus, although a text may have lower scores on cohesion indices, this may not imply that the text is less coherent. Coherence is in the mind of the reader and, as Crossley and McNamara (2010) point out, text coherence for high-knowledge readers may not be reflected by higher incidences of cohesive devices.

Last, our analysis showed that texts written by more advanced writers contained more syntactically complex structures in the form of the number of modifiers per noun phrase. As writers advance, they produce noun phrases that consist of more words and are, hence, more difficult to process, especially for lower level readers (Just & Carpenter, 1992). Such a finding

Table 5. Linguistic Differences in Example Essays

	Ninth-grade essay	College freshman essay
Number of words	281.000	275.000
Word frequency—all words	3.386	2.937
Incidence of positive logical connectives	85.409	18.182
Word concreteness—content words	304.131	367.416
Lexical diversity (*D*)	29.000	129.000
Modifiers per noun phrase	0.596	0.778
Content word overlap	0.368	0.048
Word polysemy	4.331	3.98
Number of paragraphs	1	5

supports previous research that indicates that syntactic complexity develops from first grade through college (Haswell, 2000; Hunt, 1965; O'Donnell et al., 1967; Stewart, 1978). Also, as syntactic complexity develops, it seems likely that writers begin to use element of the subsentence (i.e., modification and embedding) to implicitly connect ideas. This increased interclausal and interphrasal embedding diminishes the need for explicit connections (i.e., the incidence of positive logical connectives) and would thus lead to the lower cohesion scores reported for the advanced texts in this study.

We have placed two essays in the appendix to help illuminate the differences in essays based on grade level. The first essay was written by a participant in the ninth grade. The second essay was written by a college freshman. The two essays differ in lexical features, syntactic complexity, cohesive devices, and the number of paragraphs. Descriptive statistics for these differences are located in Table 5. Overall, these linguistic differences are telling. The essay written by the ninth-grade student never fully develops ideas and instead focuses solely on a single theme (optimism) and repeats the theme creating more content word overlap as well as less lexical diversity. In addition, the lack of detail in the author's argument forces word choices that are less concrete (e.g., *people*, *better*, *things*) and more ambiguous (e.g., *get*, *chance*, *are*). The prevalence of connectives (e.g., *and* and *because*) in the ninth-grade text also produces a more cohesive text with more explicit connections, unlike the college freshman text. These differences create a text that is less sophisticated with more explicit cohesive devices to guide the reader. From a readability perspective, the ninth-grade text should be easier to comprehend and process (Just & Carpenter, 1980; Rayner & Pollatsek, 1994). However, those cues that make the text more readable also make it simpler and less effective.

Overall, we see that more skilled writers create texts that are more sophisticated and less cohesive. Such a result likely emerges from the effects of working memory and background knowledge (e.g., Kellogg, 2008; McCutchen, 2000; Swanson & Berninger, 1996). Thus, the more skilled a writer is, the greater the working memory capacity the writer has. This capacity along with a writer's background knowledge allows the writer to produce texts containing a greater depth of information and that use more sophisticated language because their working memory assists them in producing more syntactically complex sentences that are filled with a greater diversity of word types.

Conclusion

This study has focused on writing development in adolescents and young adults by taking advantage of sophisticated linguistic tools to provide strong quantitative evidence as to the linguistic differences that emerge between grade levels. We find that even in advanced writers, lexical and syntactic constructions continue to develop. Also, in a similar manner to Haswell (1986), we find that cohesive devices are important indicators of writing development but that fewer cohesive devices are the mark of more mature writers. Developing writers thus exhibit quadratic trends in the production of cohesive devices with elementary- and college-level students producing fewer cohesive devices than junior high school students.

To build on this research, future studies should replicate this study using longitudinal methods of data collection instead of cross-sectional methods of data collection (e.g., Haswell, 2000). Such analyses would help control for writer-specific variables (i.e., demographic differences) that were not possible in our analysis. Future studies would also benefit from a focus on the cognitive factors related to the writing process and how they interact with the production of linguistic features. Extensions from this approach should also consider the development of structural and rhetorical patterns in essay writing. Such studies would provide support for not only the development of linguistic factors as a function of grade level but also makes links between these linguistic factors and the cognitive, structural, and rhetorical factors that may underlie them.

Appendix

Example 1: Essay Written by a Participant From the Ninth Grade

Being realistic is better for people than being optimistic because if one person was optimistic, there is a better chance that they will get their hopes up for nothing because people that are optimistic always thinks the best of something and that is not always what happens when in other words when a person is more realistic they have a better life because they look at all the possible outcomes instead of what they want to happen so they don't get their hopes up as much as the people that are as optimistic because the different ways that they think about of the views that they give to things. It shows that people that are more optimistic get upset from the facts that they get because they are not what they wanted the final answer to be and the people that are realistic don't get upset because they think about all the possible out comes that someone could get in the situation that they are in and not the outlook of what they want to happen in the situation. Most optimistic people are very daring cause they think nothing will happen to them because of the positive outlooks that they are giving to themselves for being optimistic when someone that is more realistic they are more afraid because they look at all the possible outcomes that could occur because they know that not everything that happens in life is going to turn out good and that it is more likely for something bad to happen then for something good to happen in what ever situation that they they might be on. Most optimistic people probably die earlier.

Example 2: Essay Written by a College Freshman

Although it's very uncommon, I believe it is possible to still be somewhat original. It simply takes a little, extra creativity. Scientist are constantly discovering new things, authors, although often conforming to archetypes in their writing, create relatively new stories, and engineers and architects are constantly pushing the limits and creating new and exciting "works of art"!

Many new discoveries are made in the scientific fields every year. Within the past decade scientist have cloned animals, mapped the human genome, and explored regions of space formerly unseen. For scientist, the universe is regularly providing new fields to work in and new ideas to work towards.

These days, many authors write purely original works. While the plots of these stories can easily be compared to some ancient archetypes, their is often something unique about the characters and their settings. I believe this is what makes the story it's own and therefore certainly not a mere copy of a previous work. Books such as Harry Potter and Twilight have defined recent readers interest and fantasies.

Architects and engineers are designing new buildings everyday, and with the use of new technologies, they can often create structures which were previously impossible. Buildings such as an apartment tower which rotates on a central point, massive skyscrapers in Tokyo, and bridges which span farther than any before are redefining our cities and landscapes.

To be original in present times is difficult. With so much history and so many prior ideas, designs, and discoveries, one might consider it nearly impossible to create anything original, but by stopping to look around, one might find an original right in front of them.

Authors' Note

Ideas expressed in this material are those of the authors and do not necessarily reflect the views of the Institute for Education Sciences (IES).

Acknowledgments

The authors would like to thank Brad Campbell, Daniel White, Steve Chrestman, Michael Kardos, Becky Hagenston, LaToya Bogards, Ashley Leonard, and Marty Price for their help in scoring the corpus of essays.

Declaration of Conflicting Interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: This research was supported in part by the Institute for Education Sciences (IES R305A080589 and IES R305G20018-02).

References

Abbott, R., & Berninger, V. W. (1993). Structural equation modeling of relationships among developmental skills and writing skills in primary- and intermediate-grade writers. *Journal of Educational Psychology, 85*, 478-508.

Abbott, R., Berninger, V. W., & Fayol, M. (2010). Longitudinal relationships of levels of language in writing and between writing and reading in grades 1 to 7. *Journal of Educational Psychology, 102*, 281-298.

Akiguet, S., & Piolat, A. (1996). Insertion of connectives by 9- to 11-year-old children in an argumentative text. *Argumentation, 10*, 253-270.

Baayen, R. H., Piepenbrock, R., & Gulikers, L. (1995). *The CELEX lexical database (release 2)* [CD]. Philadelphia, PA: Linguistic Data Consortium, University of Pennsylvania.

Berninger, V., Abbott, R., Swanson, H. L., Lovitt, D., Trivedi, P., Lin, S., . . . Amtmann, D. (2010). Relationship of word- and sentence-level working memory to reading and writing in second, fourth, and sixth grade. *Language, Speech, and Hearing Services in Schools*, *41*, 179-193.

Berninger, V., Abbott, R., Thomson, J., Wagner, R., Swanson, H. L., Wijsman, F., & Raskind, W. (2006). Modeling developmental phonological core deficits within a working-memory architecture in children and adults with developmental dyslexia. *Scientific Studies in Reading, 10*, 165-198.

Berninger, V., Cartwright, A., Yates, C., Swanson, L., & Abbott, R. (1994). Developmental skills related to writing and reading acquisition in the intermediate grades. *Reading and Writing: An Interdisciplinary Journal, 6*, 161-196.

Berninger, V., Mizokawa, D., & Bragg, R. (1991). Theory-based diagnosis and remediation of writing disabilities. *Journal of School Psychology, 29*, 57-79.

Berninger, V., Yates, C., Cartwright, A., Rutberg, J., Remy, E., & Abbott, R. (1992). Lower-level developmental skills in beginning writing. *Reading and Writing: An Interdisciplinary Journal, 4*, 257-280.

Brace, N., Kemp, R., & Snelgar, R. (2006). *SPSS for psychologists*. Mahwah, NJ: Lawrence Erlbaum.

Costerman, J., & Fayol, M. (1997). *Processing interclausal relationships: Studies in production and comprehension of text*. Hillsdale, NJ: Lawrence Erlbaum.

Crismore, A., Markkanen, R., & Steffensen, M. S. (1993). Metadiscourse in persuasive writing: A study of texts written by American and Finnish university students. *Written Communication, 10*, 39-71.

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

Crossley, S. A., & McNamara, D. S. (2010). Cohesion, coherence, and expert evaluations of writing proficiency. In R. Catrambone & S. Ohlsson (Eds.), *Proceedings of the 32nd annual conference of the Cognitive Science Society* (pp. 984-989). Austin, TX: Cognitive Science Society.

Crossley, S. A., & McNamara, D. S. (in press). Predicting second language writing proficiency: The role of cohesion, readability, and lexical difficulty. *Journal of Research in Reading*.

Douglas, D. (1981). An exploratory study of bilingual reading proficiency. In S. Hudelson (Ed.), *Learning to read in different languages* (pp. 33-102). Washington, DC: Center for Applied Linguistics.

Durst, R. K. (1984). The development of analytic writing. In A. N. Applebee (Ed.), *Contexts for learning to write: Studies of secondary school instruction* (pp. 79-102). Norwood, MA: Ablex Press.

Fellbaum, C. (1998). *WordNet: An electronic lexical database*. Cambridge, MA: MIT Press.

Field, A. (2005). *Discovering statistics using SPSS*. London, UK: SAGE.

Freedman, A., & Pringle, I. (1980). Writing in the college years: Some indices of growth. *College Composition and Communication, 31*, 311-324.

Grabowski, J., Weinzierl, C., & Schmitt, M. (2010). Second and fourth graders' copying ability: From graphical to linguistic processing. *Journal of Research in Reading, 33*, 39-53.

Graesser, A. C., McNamara, D. S., Louwerse, M., & Cai, Z. (2004). Coh-Metrix: Analysis of text on cohesion and language. *Behavior Research Methods, Instruments, & Computers, 36*, 193-202.

Haberlandt, K., & Graesser, A. C. (1985). Component processes in text comprehension and some of their interactions. *Journal of Experimental Psychology: General, 114*, 357-374.

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

Haswell, R. H. (1986). *Change in undergraduate and post-graduate writing performance: Quantified findings* (Technical Report). Retrieved from ERIC database. (ED269780)

Haswell, R. H. (1989). Textual research and coherence: Findings, intuition, application. *College English, 51*, 305-319.

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

Hayes, J., & Flower, L. (1980). Identifying the organization of the writing process. In L. W. Gregg & E. R. Steinberg (Eds.), *Cognitive processes in writing* (pp. 3-30). Hillsdale, NJ: Erlbaum.

Hillocks, G. (2002). *The testing nap: How state writing assessments control learning*. New York, NY: Teachers College Press.

Hunt, K. W. (1965). *Grammatical structures written at three grade levels* (NCTE Research Report Number 3). Urbana, IL: National Council of Teachers of English.

Huot, B. (1990). The literature of direct writing assessment: Major concerns and prevailing tends. *Review of Educational Research, 60*, 237-263.

Institute of Education Sciences. (2003). *The nation's report card: Writing 2002* (NCES 2003-529). Retrieved from http://nces.ed.gov/nationsreportcard/pubs/main2002/2003529.asp

Just, M. A., & Carpenter, P. A. (1980). A theory of reading: From eye fixations to comprehension. *Psychological Review, 87*, 329-354.

Just, M. A., & Carpenter, P. A. (1992). A capacity theory of comprehension: Individual differences in working memory. *Psychological Review, 99*, 122-149.

Kellogg, R. T. (2008). Training writing skills: A cognitive developmental perspective. *Journal of Writing Research, 1*, 1-26.

King, M., & Rentel, V. (1979). Toward a theory of early writing development. *Research in the Teaching of English, 13*, 243-253.

Kintsch, W., & van dijk, T. A. (1978). Toward a model of text comprehension and production. *Psychological Review, 85*, 363-394.

Landauer, T., McNamara, D. S., Dennis, S., & Kintsch, W. (Eds.). (2007). *Handbook of latent semantic analysis*. Mahwah, NJ: Erlbaum.

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

Longo, B. (1994). Current research in technical communication: The role of metadiscourse in persuasion. *Technical Communication, 41*, 348-352.

Louwerse, M. M. (2001). An analytic and cognitive parameterization of coherence relations. *Cognitive Linguistics, 12*, 291-315.

Malvern, D. D., Richards, B., Chipere, N., & Duran, P. (2004). *Lexical diversity and language development: Quantification and assessment*. Basingstoke, UK: Palgrave Macmillan.

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.

McCutchen, D. (1986). Domain knowledge and linguistic knowledge in the development of writing ability. *Journal of Memory and Language, 25*, 431-444.

McCutchen, D. (2000). Knowledge, processing, and working memory: Implications for a theory of writing. *Educational Psychologist, 35*, 13-23.

McCutchen, D., & Perfetti, C. (1982). Coherence and connectedness in the development of discourse production. *Text, 2*, 113-139.

McNamara, D. S. (2001). Reading both high-coherence and low-coherence texts: Effects of text sequence and prior knowledge. *Canadian Journal of Experimental Psychology, 55*, 51-62.

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

McNamara, D. S., & Graesser, A. C. (in press). Coh-Metrix. In P. M. McCarthy & C. Boonthum (Eds.), *Applied natural language processing and content analysis: Identification, investigation, and resolution*. Hershey, PA: IGI Global.

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

Miller, G. A., Beckwith, R., Fellbaum, C., Gross, D., & Miller, K. (1990). *Five papers on WordNet*. Princeton, NJ: Princeton University, Cognitive Science Laboratory.

O'Donnell, R. C., Griffin, W. J., & Norris, R. C. (1967). *Syntax of kindergarten and elementary school children: A transformational analysis* (NCTE Research Report Number 8). Urbana, IL: National Council of Teachers of English.

Oliver, E. (1995). The writing quality of seventh, ninth, and eleventh graders, and college freshmen: Does rhetorical specification in writing prompts make a difference? *Research in the Teaching of English, 29*, 422-450.

O'Reilly, T., & McNamara, D. S. (2007). Reversing the reverse cohesion effect: Good texts can be better for strategic, high-knowledge readers. *Discourse Processes, 43*, 121-152.

Pearson, P. D. (1974-1975). The effects of grammatical complexity on children's comprehension, recall, and conception of certain semantic. *Reading Research Quarterly, 10*, 155-192.

Perfetti, C. A., Landi, N., & Oakhill, J. (2005). The acquisition of reading comprehension skill. In M. J. Snowling & C. Hulme (Eds.), *The science of reading: A handbook* (pp. 227-247). Oxford, UK: Blackwell.

Perfetti, C., & McCutchen, D. (1987). Schooled language competence: Linguistic abilities in reading and writing. In S. Rosenberg (Ed.), *Advances in applied psycholinguistics* (pp. 105-141). Cambridge, UK: Cambridge University Press.

Rashotte, C. A., & Torgesen, J. K. (1985). Repeated reading and reading fluency in learning disabled children. *Reading Research Quarterly, 20*, 180-188.

Rayner, K. & Pollatsek, A. (1994). The psychology of reading. Englewood Cliffs, NJ: Prentice Hall.

Stewart, M. F. (1978). Syntactic maturity from high school to university: A first look. *Research in the Teaching of English*, *12*(1), 37-46.

Swanson, H. L., & Berninger, V. W. (1996). Individual differences in children's working memory and writing skill. *Journal of Experimental Child Psychology, 63*, 358-385.

Tabachnick, B. G., & Fidell, L. S. (2001). *Using multivariate statistics* (4th ed.). Needham Heights, MA: Allyn & Bacon.

Van de Kopple, W. (1985). Some exploratory discourse on metadiscourse. *College Composition and Communication, 36*, 82-93.

Weston, J. L., Crossley, S. A., McCarthy, P. M., & McNamara, D. S. (2011). Number of words versus number of ideas: Finding a better predictor of writing quality. *Proceedings of the 24th International Florida Artificial Intelligence Research Society*. Retrieved from http://www.aaai.org/ocs/index.php/FLAIRS/FLAIRS11/paper/download/2618/3182

Wilson, M. D. (1988). The MRC psycholinguistic database: Machine-readable dictionary, version 2. *Behavioural Research Methods, Instruments, and Computers, 201*, 6-11.

Witten, I. H., & Frank, E. (2005). *Data mining: Practical machine learning tools and techniques*. San Francisco, CA: Elsevier.

Witte, S. P., & Faigley, L. (1981). Coherence, cohesion, and writing quality. *College Composition and Communication, 32*, 189-204.

Bios

Scott Crossley is an assistant professor at Georgia State University, Atlanta. His interests include computational linguistics, corpus linguistics, and second-language acquisition. He has published articles in second-language lexical acquisition, multidimensional analysis, discourse processing, speech act classification, cognitive science, and text linguistics.

Jennifer Weston is a graduate student as the University of Memphis, Memphis, Tennessee. She works at the Institute for Intelligent Systems on a variety of projects focused on student writing. Her current interests include development of assessments for student strategy use and the use of these assessments in intelligent tutoring systems.

Susan McLain Sullivan is an English instructor at Oakton High School, Vienna, Virginia. Her instruction includes 7 years of Advanced Placement English Language and Composition and 3 years of Advanced Placement English Literature as well as instruction in pre-AP classes. Her work includes AP prep and SAT prep and advising the OHS literature and arts magazine. She has also taught writing at the University of Memphis.

Danielle McNamara is a professor at the University of Memphis and director of the Institute for Intelligent Systems. Her work involves the theoretical study of cognitive processes as well as the application of cognitive principles to educational practice. Her current research ranges a variety of topics including text comprehension, writing strategies, building tutoring technologies, and developing natural language algorithms.