Title

Linguistic Features of Secondary School Writing: Can Natural Language Processing Shine a Light on Differences by Sex, English Language Status, or Higher Scoring Essays?

Journal

Written Communication, 41(3)

ISSN

0741-0883

Authors

Tate, Tamara P
Kim, Young-Suk Grace
Collins, Penelope
et al.

Publication Date

2024-07-01

DOI

10.1177/07410883241242093

Linguistic Features of Secondary School Writing: Can NLP shine a light on differences by gender, English language status, or higher scoring essays?

Abstract

This article provides three major contributions to the literature: we provide granular information on the development of student writing across secondary school; we replicate the MacArthur et al. (2019) model of NLP writing features that predict quality with a younger group of students; and, because the student population at this level is more heterogeneous than college students, we are able to look at the differences for students across language status. We sought to find the average levels of text length, cohesion, connectives, syntactic complexity, and word-level complexity in this sample across Grades 7-12, by gender, by English learner status, and for essays scoring above and below the median holistic score. We also looked at whether text length, cohesion, connectives, syntactic complexity, and word-level complexity predict writing quality (holistic score) for secondary school writers using the MacArthur et al. (2019) model. Looking across grades, we find that the text length of younger students increased as they aged, but the model otherwise was fairly stable. Gender did not seem to affect the model in meaningful ways beyond the increased fluency of women writers. We found text length and word level differences, but not holistic quality score differences, between initially designated and redesignated bilingual students compared to their English-only peers. Finally, we found that the model works better with our higher scoring essays and was less effective explaining the lower scoring essays.

Keywords

Writing, Linguistic features, Natural language processing, secondary school, literacy

Introduction

Students' difficulties in writing adversely affects their progression through school and ultimately impact their access to, persistence through, and performance in postsecondary education and careers (Brandt, 2014). Nonetheless, secondary students struggle to write well, with large percentages of students scoring poorly on the most recently published results from the National Assessment of Educational Progress (NAEP). Performance is particularly poor for English learners (e.g., average 8th grade English learners scored 108--below basic--compared to 152 for non-English learners) and Hispanic students (average of 136, compared to 158 for White students; National Center for Educational Statistics, nd). Basic scores (120-172) reflect students who are able to address the tasks appropriately and mostly accomplish their communicative purposes, compared to Proficient scores for students who clearly accomplish their communicative purposes. Reports of significant learning delays in reading and math during the pandemic (CREDO, 2020; Dorn et al., 2020; Kogan & Lavertu, 2021; Kuhfeld et al., 2020) are likely to be true of writing, which is historically less of an instructional priority than reading and math, and thus we expect the trend to worsen.

However, teachers struggle to teach students' writing. Writing is an extremely complicated task and most teachers have insufficient training in writing pedagogy, in both pre-service courses and

after they begin teaching. They struggle to find time in the curriculum, are burdened by low-self efficacy for teaching writing, and have limited research-based, efficient and effective curricula to improve student writing in middle and high school. Indeed, teachers--and researchers--have trouble even defining the components of quality writing, let alone instruction. In our effort to create a writing intervention for history teachers

The goal of this paper is to generate insight into what good writing looks like and how it develops over middle and high school in order to develop effective, research-based interventions to train teachers what skills to prioritize and when. For this purpose, we closely examine linguistic features–cohesion, use of connectives, syntactic complexity, and word-level complexity–that are likely to predict the writing quality of the argumentative essays of secondary school students. We chose these features based on current research literature and the model of MacArthur, Jennings, and Philippakos (2019), which was successfully used in connection with basic-level undergraduate students. Our analysis was conducted on a corpora of source-based argument writing gathered prior to a well-established literacy intervention for secondary students, the [masked for review] ("[Blinded]"). This article provides three major contributions to the literature: we provide granular information on the development of student writing across secondary school; we replicate the MacArthur et al. (2019) model of NLP writing features that predict quality with a younger group of students; and, because the student population at this level is more heterogeneous than college students, we are able to look at the differences for students across language status.

Background

Effective written compositions convey the author's intended message clearly and precisely for the given goal and audience. Therefore, language skills are important to the quality of written composition, and this is reflected in the classical and more recent theoretical models of writing (e.g., Berninger et al., 2002; Flower & Hayes, 1981; Kim & Graham, 2022). The contributions of linguistic knowledge to written compositions have been shown in studies where language skills such as vocabulary, grammatical (morphosyntactic and syntactic knowledge) and discourse oral language were examined as predictors of writing quality (Coker, 2006; Kim et al., 2015; Kim & Schatschneider, 2017). Another approach to examining the role of linguistic knowledge in written compositions is an examination of linguistic features in written comprehension. Studies in this line of work found that essays containing more infrequent words were considered to be of higher quality by expert raters (Crossley, Weston, et al., 2011; McNamara et al., 2010). In addition, more advanced writers use fewer common, concrete words (Crossley, Weston, et al., 2011), and fewer short words (Haswell, 2000). Ultimately, more sophisticated words are evidence of higher overall text quality (Crossley, 2020). Text cohesion is also related to writing quality. As young as kindergarten children begin to use cohesive devices, including referential pronouns and connectives (Kim et al., 2011) and continue to develop their proficiency with them until about the 8th grade (McCutchen & Perfetti, 1982; McCutchen, 1986), after which they reduce their reliance on explicit cohesive devices to create coherence (Crossley et al., 2011). Furthermore, syntactic skills are related to writing quality, with more advanced writers showing more refined sentence generation skills than beginning writers (McCutchen et al., 1994) with the development of syntactic complexity developing from 1st grade through college (Haswell, 2000) and students creating fewer run-on sentences and sentence fragments as they develop as writers

over time (Berninger, 2011). However, traditional measures such as mean T-unit length have reported inconsistencies across studies that the CohMetrix measures are intended to reduce (Crossley, 2020).

Differences in writing quality by gender (Graham et al., 2017; Reilly et al., 2019; Steiss et al., 2022) and language proficiency (Graham & Perin, 2007; NCES, 2012; Author et al., 2017) are well documented, and we were motivated to better understand the underlying differences in students' writing in order to inform instruction to close the proficiency gap. For example, only 1% of English learners at both grades 8 and 12 scored proficient or above in the National Assessment of Educational Progress in writing (NCES, 2012).

Coh-Metrix (McNamara & Graesser, 2012) is a natural language processing tool that provides a wide range of indices of linguistic and discourse representations of texts. It provides 108 different indices categorized into 11 groups, ranging from descriptive information such as the number of words in the text, to complex indices, such as lexical diversity and complexity (McNamara et al., 2014). Numerous studies have used Coh-Metrix to identify specific measures of writing that describe the developmental progression of writing and, particularly, the difference between high quality and low-quality writing. For example, McNamara et al. (2010) studied essays by college freshmen to distinguish the difference between those rated by human scorers as high and as low. They found the three most predictive indices of essay quality were syntactic complexity (measured by the number of words before the main verb), lexical diversity (measured by the Measure of Textual Lexical Diversity), and word frequency (measured by Celex). None of the 26 measures of cohesion showed differences between ratings. We replicate their process of splitting the corpus into higher and lower-scoring essays to allow us to understand the differences between essays of higher and lower quality as part of our first research question.

Coh-Metrix has also been used to explore writing development. In one example, Crossley et al. (2011) examined 9th grade, 11th grade, and college freshman argumentative essays along the dimensions of lexical sophistication, syntactic complexity, text structure (a combination of word length, number of paragraphs, and number of sentences), and cohesion. Ultimately, they determined that students produce more sophisticated words and use more complex sentence structure as grade level increases, but use fewer cohesive features. Not surprisingly, the strongest predictor of grade level was the number of words in a text. As noted by MacArthur (2019, p. 1557), "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." Length is consistently the strongest predictor of essay quality for secondary and college students (Crossley et al., 2011; McNamara et al., 2013; MacArthur, 2019; Powers, 2005). Syntactic complexity is more complicated, with indications at the college level sometimes showing positive correlations with quality and other times showing no significant relationship (Crossely & McNamara et al., 2014; Crossley et al., 2011; McNamara et al., 2010). The correlation between cohesion and quality has been even more mixed (see McNamara et al., 2010; Crossley et al, 2011; Perin & Lauterbach, 2016; McNamara et al., 2013).

The current study

We investigated linguistic features–cohesion, use of connectives, syntactic complexity, and word-level complexity– in the argumentative essays of secondary school students and their relations to writing quality, using data from students in Grades 7 to 12. Despite the growing literature on the relations of linguistic features identified by natural language processing to writing quality for college and adult writers, there has been insufficient empirical study at the secondary level to shed light on the features of high-quality writing. Therefore, we sought to answer the following research questions:

1. What are the average levels of text length, cohesion, connectives, syntactic complexity, and word-level complexity in this sample across Grades 7-12, by gender, by English learner status, and for essays scoring above and below the median holistic score?
2. How do text length, cohesion, connectives, syntactic complexity, and word-level complexity predict writing quality (holistic score) for secondary school writers?

We hypothesized that the average text length and word-level complexity would increase as the grade level increased and in the higher-scoring essays compared to the lower scoring essays. We had no clear expectations for syntactic complexity because we had seen in other work that immature writers frequently use run-on sentences which are complex but ineffective.

Method

Data source

This is a secondary data analysis of texts from one school district in the southwest United States, with 74% of its students receiving free and reduced-price meals, and 80% Hispanic or Latino, 9% White, 4% Asian, and 3% African-American students. Approximately 61% of the students only spoke English, 6% were classified as initial fluent English proficient (IFEP), 16% as English learners, and 17% as reclassified fluent English proficient (RFEP). Part of the [Blinded] project reported more fully in Author et al. (2019), the texts are from students in English Language Arts classes in grades 7 through 12.

The prompts for these texts asked students to read a single text, a nonfiction newspaper article, and then write about the author's message (i.e., present a theme statement), analyze the author's use of figurative language, and discuss the author's purpose. This type of essay is an argument of interpretive analysis (Smith, Wilhelm, & Fredricksen, 2012) or literary judgment (Hillocks, 2011). All 1067 texts in our sample were gathered prior to the [Blinded] intervention treatment. The analytic sample comprised the following demographics: 53% men; 67% Hispanic, 18% White, 2% Asian, 2% African American; and 58% English only, 6% English learners, 9% IFEP, and 26% RFEP.

For this study, the handwritten student texts were transcribed by a third party service, which was asked to use a standard spelling correction tool and insert any omitted periods to sentence endings where appropriate to enable natural language processing of the texts. Correction of spelling errors at pretest in the MacArthur et al. (2019) sample was more aggressive than those employed with our sample, with researchers correcting homonyms, apostrophes, abbreviations, capitalization as well as general spelling errors.

Of the full sample of 1067 papers, a subset (n = 174) was scored by trained raters overseen by XXX (see Author, et al., 2019 for the sampling method) on a prompt-agnostic rubric for

evaluation purposes, the Analytic Writing Continuum for Literary Analysis (AWC-LA) developed by the National Writing Project and shown to be a valid and reliable measure of student writing (Bang, 2013). Each handwritten paper was given a holistic rating as well as ratings on each of four attributes: content, structure, sentence fluency, and conventions. Raters agreed within a single score point for 90% of papers on the holistic score (Author et al., 2019). These scores are used in our analyses as the quality of writing.

Linguistic indices

The transcribed written texts were processed by researchers using the Coh-Metrix (Graesser, et al., 2004) natural language processing tool. Because studies show that length is highly predictive of quality, that element was controlled for in the model of lexical and syntactic complexity and cohesion. Indices were selected to represent four constructs based on theoretical considerations and prior research: word-level complexity, syntactic complexity, and two types of cohesion--referential cohesion and connectives.

Ultimately, we looked to MacArthur, Jennings, & Philippakos' (2019) study of basic college writers (n = 252) as the model for our work. Basic college writers were close to our high school students' developmental level; they used a corpus of persuasive essays, a genre closely aligned with our corpus; and their analytical method was compelling. They eliminated indices that were highly correlated with length in order to avoid confounding interpretations.

Word-level complexity measures from Coh-Metrix were selected from the categories of Lexical Diversity (the number of unique words compared to total words), Word Information (the frequency with which words are used and indices of age of acquisition of the words, concreteness, and imageability), and Descriptive measures related to words. Syntactic complexity indices were selected from the categories of Syntactic Complexity (e.g., length of nominal phrases and similarity of syntactic structure across sentences), Syntactic Pattern Density (relative incidence of types of phrases and word forms like noun and verb phrases), and Descriptives related to sentences. Referential cohesion came from the Coh-Metrix category of the same name, referring to links between words and across sentences that help with sense-making by readers, and Latent Semantic Analysis which considers semantically related words ("house" and "home"). Finally, connectives came from the category with the same name and describes words that make temporal ("then"), additive ("in addition"), contrastive ("on the other hand"), and other connections within a text. MacArthur et al. (2019) retained variables that had a correlation of less than r = 0.20 with essay length, correlation with other indices in the same construct were less than r = 0.90 (to avoid collinearity) and at least 0.30 (so they were related to the same underlying construct).

Data analysis

Descriptive data on means and correlations were used to determine the average levels of text length, cohesion, connectives, syntactic complexity, and word-level complexity across grades 7-12, by gender, by English learner status, and by holistic score. We addressed the first research question, whether there were systematic differences across the writing dimensions as a function of grade, gender, English proficiency, and writing performance using a series of multivariate

analyses (MANOVAs). We used Pillai's trace statistic as it gives robust results with unbalanced samples with nonnormal and heterogeneous variance (Ateş et al., 2019). We used the Bonferroni adjustment to protect against experiment-error. To address the second research question, the contributions of text length, cohesion, connectives, syntactic complexity, and word-level complexity to writing quality, we calculated structural equation models of the four latent constructs and their contributions to the essay quality (holistic score). MPlus version 8 (Muthén & Muthén, 2017) was used with maximum likelihood estimation. In order to match the MacArthur et al. (2019) analysis, we transformed the SYNSTRUTt and WRDFRQa variables into negative values. SYNLE and DESSLD were transformed due to their distributional properties by taking the log of each.

To illuminate the difference in higher and lower quality essays, as part of our heterogeneity analysis for research question 1, we replicated the McNamara et al. (2010) process for analyzing higher quality essays by splitting the corpus at the median holistic score (in our case, essays of 3 and above, in the McNamara data 3.1 and above, also on a 1-6 point scale in a sample of college freshmen). The McNamara model variables, however, were different and we continued to use our model.

Results

First, we first show correlations among quality, length, and the linguistic indices (Table 1).

Table 1. Correlations.

	Cohesion			Connectives			Syntax			Word Level			Holistic Score	Word Count
	CRFSO1	CRFSOa	LSASS1	CNCLogic	CNCADC	CNCAdd	SYNSTRUTt	SYNLE	DESSLd	WRDFRQa	WRDAOAc	DESWLsy		
CRFSO1	–0													
CRFSOa	0.872	–0												
LSASS1	0.647	0.565	–1.											
CNCLogic	0.054	0.027	0.080	–1.										
CNCADC	-0.018	0.027	-0.020	0.457	–1.									
CNCAdd	-0.106	-0.040	-0.048	0.311	0.310	–1.0								
SYNSTRUTt	0.140	0.159	0.031	0.073	0.143	0.106	–1.							
SYNLE	0.304	0.370	0.139	-0.074	-0.022	-0.139	0.201	–1.						
DESSLd	0.217	0.233	0.145	0.074	0.090	0.128	0.418	0.267	–1.					
WRDFRQa	-0.075	-0.129	0.051	-0.131	-0.084	0.029	-0.092	-0.013	-0.077	–1.				
WRDAOAc	-0.075	-0.136	-0.024	0.050	0.020	0.098	0.019	-0.120	-0.025	0.125	–1.			
DESWLsy	-0.019	-0.026	-0.043	-0.252	-0.165	-0.102	-0.042	0.095	-0.030	0.580	0.235	–1.		
Holistic Score	0.099	0.025	-0.042	0.042	0.190	0.086	0.121	0.066	0.093	-0.092	0.131	0.127	–1.	
Word Count	-0.073	-0.129	-0.104	0.125	0.185	0.104	0.109	0.025	0.210	-0.021	0.117	0.052	0.644	–1.

Research Question 1.

Our first research question sought to describe the quality, length and linguistic characteristics of student writing across grades 7 to 12, with particular attention to individual differences based on gender and English learner status. We also sought to determine how stronger and weaker essays differ along these dimensions systematically.

We first examined writing performance across the secondary school grades. Table 2 summarizes the means and standard deviations of the linguistic indices, quality and length of the essays, both in aggregate and by grade level.

Table 2. Means by grade level.

		All		Grade 7		Grade 8		Grade 9		Grade 10		Grade 11		Grade 12	
		N = 1067		*N* = 220		*N* = 193		*N* = 162		*N* = 189		*N* =189		*N* = 114	
	Variable	Mean	SD	Mean	SD	Mean	SD	Mean	SD	Mean	SD	Mean	SD	Mean	SD
Cohesion	CRFSO1	0.44	0.26	0.48	0.28	0.46	0.28	0.44	0.26	0.40	0.25	0.44	0.28	0.42	0.21
	CRFSOa	0.39	0.24	0.45	0.25	0.39	0.24	0.41	0.25	0.33	0.23	0.41	0.27	0.36	0.183
	LSASS1	0.16	0.09	0.19	0.01	0.19	0.01	0.15	0.01	0.16	0.00	0.15	0.01	0.14	0.00
Connectives	CNCLogic	37.37	19.49	36.39	17.05	38.27	27.54	37.68	17.43	37.77	19.50	36.75	16.09	37.34	16.45
	CNCADC	16.20	11.34	13.35	10.63	13.68	12.71	18.65	11.95	14.68	10.80	17.14	9.70	20.96	10.37
	CNCAdd	44.86	17.77	41.96	17.89	44.33	15.79	51.41	16.58	44.55	19.90	39.94	19.54	47.43	15.70
Syntax	SYNSTRUTt	0.09	0.03	0.09	0.04	0.09	0.04	0.09	0.04	0.09	0.02	0.09	0.03	0.08	0.03
	SYNLE	4.12	2.22	3.80	1.89	4.35	2.35	3.92	1.92	4.10	2.21	4.06	2.41	4.59	2.65
	DESSLD	10.67	7.53	10.19	5.21	13.78	12.42	10.01	7.15	8.30	3.11	9.93	3.78	11.18	7.42
Word Level	WRDFRQa	3.13	0.16	3.17	0.14	3.05	0.24	3.14	0.16	3.16	0.09	3.17	0.08	3.12	0.12
	WRDAOAc	313.44	38.24	294.13	29.31	314.92	61.07	312.76	29.53	322.03	25.02	316.97	35.04	324.58	22.60
	DESWLsy	1.37	0.08	1.32	0.01	1.40	0.01	1.36	0.01	1.36	0.00	1.39	0.00	1.38	0.01
Holistic Score	Holistic Score	2.63	1.25	1.94	0.97	2.41	1.21	2.48	1.09	2.65	1.29	3.23	1.19	3.34	1.28
Word Count	Word Count	226.60	126.01	172.78	85.89	220.03	120.71	224.56	132.15	228.72	127.26	232.22	105.43	301.22	160.20

A pair of analyses of variance (ANOVA) revealed significant grade-level differences in writing length, $F(5, 1061) = 16.37$, $p < .0001$ and quality, $F(5, 169) = 5.95$, $p < .0001$. Bonferroni post-hoc comparisons revealed differences in writing quality for older and younger students. More specifically, 7th grade students' essays received lower holistic scores than students in 11th and 12th grades, and 12th grade students also outperformed students in 8th grade. There were similar patterns in essay length, as students in 11th and 12th grades wrote significantly longer essays than students in 7th through 9th grades. Students in 12th grade wrote longer essays than 10th grade students, but not those in 11th grade.

We then examined grade-level patterns in the different linguistic features. We calculated a series of multivariate analysis of variance (MANOVA) on the three measures of cohesion (CRFSO1, CRFSOa and LSASS1), use of connectives (CNCLogic, CNCADC, and CNCAdd), syntax (SYNSTRUTt, SYNLE, and DESSLd) and word use (WRDFRQa, WRDAOAc, and DESWLsy). We found significant grade-level differences for cohesion, $F(15, 3183) = 2.18$, $p = .005$, the use of connectives, $F(15, 3183) = 3.29$, $p < .0001$, syntax, $F(15, 3183) = 2.20$, $p = .005$, and word use, $F(15, 3183) = 7.44$, $p < .0001$. When considering cohesion, a subsequent series of ANOVAs using the Bonferroni adjustment for multiple comparisons revealed significant grade-level differences only for LSASS1, $F(5, 1061) = 4.63$, $p = .0004$. Tukey's post-hoc tests showed that students in 7th and 8th grades had higher LSASS1 scores than students in 12th grade. When considering connectives, we found significant grade-level differences for CNCLogic, $F(5, 1061) = 2.84$, $p = .015$, and CNCAdd, $F(5, 1061) = 3.78$, $p = .002$, but not for CNCADC, $F(5, 1051) = 1.50$, $p = .19$, with students in 7th grade used logical connectives (CNCLogic) to a greater degree than those in 11th and 12th grades. When considering syntax, we found significant grade-level differences for SYNSTRUTt, $F(5, 1051) = 3.42$, $p = .004$, and DESSLD, $F(5,1051) = 2.63$, $p = .02$. Students in 12th grade used a wider range of syntactic structures than students in 7th and 8th grades, as indicated by lower SYNSTRUTt scores. Finally, we found significant grade-level differences for each of the word-level variables (WRDFRQa: $F(5,1061) = 4.88$, $p = .0002$; WRDAOAc: $F(5,1061) = 11.05$, $p < .0001$; DESWLsy: $F(5,1061) = 12.58$, $p < .0001$). Students in 7th grade used higher frequency words (WRDFRQa) than students in grades 8, 9, 10, and 12, and words with younger age of acquisition scores (WRDAOAc). Bonferroni post-hoc comparisons indicated that 8th grade students used lower frequency words (WRDFRQa) than students in grades 7, 10 and 11. Students in 7th grade used shorter words with younger age of acquisition scores (DESWLsy and WRDAOAc) than students in all other grade levels. No other grade-level effects were significant.

Student performance as a function of gender is summarized in Table 3. First, girls wrote longer papers, $F(1, 1065) = 34.80$, $p < .0001$, that received higher holistic scores, $F(1, 173) = 4.70$, $p = 0.03$. We next calculated a series of MANOVAs to examine variations in the linguistic features as a function of gender. Boys' writing showed higher cohesion scores, $F(3, 1063) = 5.35$, $p = .001$. A subsequent series of ANOVAs using the Bonferroni adjustment for multiple comparisons revealed that boys had higher cohesion scores for CRFSO1, $F(1, 1065) = 10.17$, $p = .001$, CRFSOa, $F(1, 1065) = 13.94$, $p = .0002$, and LSASS1, $F(1,1065) = 10.03$, $p = .002$. While girls showed greater overall use of connectives, $F(3, 1063) = 3.89$, $p = .009$, these differences were limited to the use of adversative connectives (CNCADC), $F(1, 1065) = 11.09$, $p = .0001$. No other gender differences were significant.

Table 3. Means by gender.

		Boys		Girls	
	Variable	Mean	SD	Mean	SD
Cohesion	**CRFSO1**	0.44	0.26	0.39	0.22
	CRFSOa	0.40	0.23	0.35	0.20
	LSASS1	0.16	0.09	0.14	0.07
Connectives	CNCLogic	36.05	18.91	37.73	16.59
	CNCADC	15.14	11.23	17.32	9.98
	CNCAdd	42.54	19.44	42.90	17.23
Syntax	SYNSTRUTt	0.09	0.04	0.09	0.03
	SYNLE	3.98	1.96	3.88	1.86
	DESSLd	9.64	6.04	10.03	5.80
Word Level	WRDFRQa	3.15	0.13	3.15	0.13
	WRDAOac	314.97	37.66	313.67	29.77
	DESWLsy	1.37	0.09	1.38	0.09
Holistic Score	**Holistic Score**	2.43	1.17	2.83	1.29
Word Count	**Word Count**	218.50	116.11	262.81	129.24

Table 4 presents the mean scores as a function of English language learner status. A pair of ANOVAs revealed significant differences in essay length, $F(3, 1063) = 7.94$, $p < 0.0001$, and holistic scores, $F(3, 171) = 14.72$, $p < 0.0001$. Bonferroni post-hoc tests revealed that EO and IFEP students wrote longer papers than RFEP and EL students. However, students who were proficient in English (EO, IFEP and RFEP) received higher holistic writing scores than EO students. The sole MANOVA to reveal significant differences based on English proficiency was at the word level (WRDFRQa, WRDAOAc, and DESWLsy), $F(9, 3189) = 7.57$, $p < .0001$. A subsequent series of ANOVAs using the Bonferroni adjustment for multiple comparisons revealed significant differences for WRDFRQa, $F(3,1063) = 9.80$, $p = .0001$, and WRDAOAc, $F(3,1063) = 4.25$, $p = .005$. Bonferroni post-hoc comparisons indicated that EL students used higher frequency words (WRDFRQa) than students who were proficient in English (EO, IFEP and RFEP). IFEP students' writing contained more advanced vocabulary (higher WRDAOAc scores) than both EL and EO students. No other differences were found.

Table 4. Means by English language learner status.

		English Learners		English Only		IFEP		RFEP	
	Variable	Mean	SD	Mean	SD	Mean	SD	Mean	SD
Cohesion	CRFSO1	0.40	0.26	0.43	0.24	0.40	0.22	0.40	0.25
	CRFSOa	0.37	0.23	0.38	0.22	0.36	0.20	0.36	0.22
	LSASS1	0.16	0.20	0.15	0.09	0.14	0.06	0.15	0.08
Connectives	CNCLogic	35.97	23.04	36.13	17.58	36.38	16.31	38.87	17.53
	CNCADC	14.19	11.74	16.00	10.60	27.55	9.78	16.54	10.99
	CNCADD	44.06	19.31	41.50	18.08	43.01	17.45	45.02	19.15
Syntax	SYNSTRUTt	0.10	0.04	0.09	0.03	0.09	0.03	0.10	0.04
	SYNLE	4.14	2.35	3.98	1.88	3.92	1.88	3.76	1.85
	DESSLD	11.05	7.72	9.85	5.92	9.74	5.70	9.49	5.49
Word Level	**WRDFRQa**	3.07	0.21	3.16	0.13	3.14	0.10	3.15	0.11
	WRDAOAc	305.11	47.48	313.46	34.12	323.32	29.90	315.38	30.99
	DESWLsy	1.37	0.11	1.38	0.09	1.39	0.09	1.37	0.09
Holistic Score	**Holistic Score**	1.61	0.82	3.04	1.36	3.09	1.03	2.66	1.02
Word Count	**Word Count**	198.37	128.28	247.59	129.54	270.64	121.32	219.68	106.80

Note: Students identified as English learners, English only speakers, initially fluent in English, or redesignated as fluent in English.

Finally, we considered how the linguistic features varied for papers scoring above and below the median holistic score (please see Table 6 for the descriptives). High-scoring essays were longer, $F(1,72) = 64.36$, $p<.00001$, than low-scoring essays. Similarly, MANOVA found that high-scoring essays showed greater use of cohesive devices, $F(3,170) = 7.14$, $p = .00015$, with subsequent ANOVAs revealing that high-scoring essays had higher CNCADC, $F(1,172) = 20.38$, $p = .00001$, and CNCADD scores, $F(1,172) = 5.31$, $p = .022$. Further, although the MANOVA testing word-level features was significant, $F(3,170) = 2.77$, $p = .043$, high and low-scoring essays did not differ significantly on any of the individual word-level variables (WRDFRQa, WRDAOAc, DESLsy). No other effects were significant.

Table 6. Mean values (standard variance in parentheses) for above and below median scoring essays.

		High Scoring Essays	Low Scoring Essays	ANOVA Prob>F
Cohesion	CRFSO1	0.44 (0.25)	0.45 (0.28)	0.254
	CRFSOa	0.39 (0.22)	0.40 (0.26)	0.343
	LSASS1	0.149 (0.07)	0.174 (0.10)	0.042
Connectives	CNCLogic	40.61 (17.76)	35.22 (20.01)	0.098
	CNCADC	19.88 (10.98)	12.66 (10.11)	0.000
	CNCAdd	47.99 (16.72)	41.96 (17.76)	0.022
Syntax	SYNSTRUTt	0.09 (0.03)	0.09 (0.04)	0.129
	SYNLE	3.89 (0.17)	4.38 (2.56)	0.066
	DESSLd	10.23 (5.08)	10.62 (9.05)	0.308
Word Level	WRDFRQa	3.16 (0.10)	3.14 (0.15)	0.523
	WRDAOAc	319.48 (25.50)	312.65 (35.08)	0.331
	DESWLsy	1.377 (0.07)	1.36 (0.09)	0.135
Holistic Score	**Holistic Score**	3.70 (0.80)	1.59 (0.49)	0.000
Word Count	**Word Count**	286.62 (109.35)	161.78 (95.75)	0.000

Research Question 2

The final indices in the constructs for cohesion, connectives, syntactic complexity, and word-level complexity are described in Table 7, below.

Table 7. Coh-Metrix indices in the study.

Construct	**Coh-Metrix Variable**	**Description**

Cohesion	CRFSO1	Stem overlap, adjacent sentences, binary, mean
	CRFSOa	Stem overlap, all sentence, binary, mean
	LSASS1	LSA overlap, adjacent sentences, mean
Connectives	CNCLogic	Logical connectives incidence
	CNCADC	Adversative and contrastive connectives incidence
	CNCAdd	Additive connectives incidence
Syntax	SYNSTRUTt[1]	Sentence syntax similarity, all combinations, across paragraphs, mean
	SYNLE	Left embeddedness, words before main verb, mean
	DESSLd	Sentence length, number of words, standard deviation
Word Level	WRDFRQa[1]	CELEX Log frequency for all words, mean
	WRDAOAc	Age of acquisition for content words, mean
	DESWLsy	Word length, number of syllables, mean
Text length	DESWC	Word count

[1]Index is negatively weighted in the construct, to match MacArthur et al. (2019).

Our SEM model (on MPlus) with the four latent variables, word count, and holistic scores of quality showed that connectives and syntactic complexity did not significantly predict quality, but word count, referential cohesion and particularly lexical complexity did predict quality for our text set. Our model fit was acceptable and similar to that of the MacArthur et al. (2019) SEM. Our model had a Chi-squared of 670.948 (df=69), CFI of 0.876, SRMR of 0.073, RMSEA of 0.083, and TLI of 0.837.

We then compared our entire sample and the essays with holistic scores, which we separated into the high and low scoring essays to determine whether the differences in quality are reflected in our model.

Table 8. Entire sample compared to randomly selected subset with holistic scores, separately indicating higher (3 and above) and lower (below 3) scoring essays.

	All			High			Low		
Variable	Coef.		SE	Coef.		SE	Coef.		SE
Word Count	0.644	***	0.028	0.562	***	0.047	0.261	***	0.066
Cohesion	0.171	***	0.045	0.030		0.067	0.390	***	0.093
Connectives	0.063		0.054	-0.138		0.080	0.036		0.105
Syntax	-0.056		0.057	0.131		0.084	-0.153		0.103
Word Level	0.099	*	0.043	0.183	**	0.064	0.054		0.071
Cohesion									
CRFSO1	0.988	***	0.008	0.983	***	0.009	0.982	***	0.022
CRFSOa	0.883	***	0.001	0.891	***	0.010	0.870	***	0.027
LSASS1	0.653	***	0.017	0.665	***	0.018	0.611	***	0.049
Connectives									
CNCLogic	0.691	***	0.03	0.735	***	0.031	0.363	***	0.096
CNCADC	0.661	***	0.03	0.668	***	0.030	0.030	**	0.095
CNCAdd	0.459	***	0.03	0.454	***	0.031	0.747	***	0.135
Syntax									
SYNSTRUTt	0.570	***	0.032	0.568	***	0.035	0.587	***	0.077
SYNLE	0.402	***	0.032	0.364	***	0.035	0.572	***	0.077
DESSLD	0.712	***	0.035	0.738	***	0.039	0.628	***	0.077
Word Level									
WRDFRQa	0.580	***	0.019	0.587	***	0.020	0.578	***	0.050
WRDOAc	0.235	***	0.027	0.229	***	0.029	0.271	***	0.069
DESWLsy	1.000		0.000	1.000		0.000	1.000		0.000
Covariances									
Cohesion x Connect	-0.009		0.035	0.028		0.037	-0.351	***	0.099
Cohesion x Syntax	0.344	***	0.033	0.339	***	0.036	0.383	***	0.087
Cohesion x Word	-0.021		0.028	-0.006		0.031	-0.088		0.076
Connect x Syntax	0.176	***	0.043	0.188	***	0.045	0.070		0.122
Connect x Word	-0.300	***	0.032	-0.328	***	0.033	-0.103		0.097
Syntax x Word	-0.014		0.035	-0.023		0.038	0.016		0.095

Discussion

Variable scores by grade suggest a developmental progression with respect to the holistic score and text length (word count; Figure 2). We found significant grade-level differences in both, specifically between the high and low end of the grade levels (e.g.7th and 12th grade). This finding is consistent with widespread findings of student improvement in writing quality over time and increased fluency expected as students automate some of the processes of writing (see discussion in Graham, 2019). The statistically significant differences in cohesion were isolated to LSASS1, with 7th and 8th graders having higher scores than 12th graders. LSASS1 measures how conceptually similar each sentence is to the next. Younger students are more likely to repeat

themselves, while older students often have more background knowledge and reasoning ability, which would lead to more variety in the sentences. Grade level differences in connectives were found for logical and additive connectives (CNCLogic, CNCAdd), with students in 7th grade using them more than older students, but not adversative and contrastive connectives (CNCADC). Logical and additive connectives are less complex and require lower levels of thinking than contrastive connectives and younger students rely on simple connectives, e.g., "and," to string together their thoughts. This is consistent with the research showing that use of contrastive connectives occurs later developmentally (Spenader, 2017). With respect to syntax, we saw that students in 12th grade used a wider range of syntactic structures than younger students, which is consistent with their increased familiarity with text and text structures. Finally, students in 7th grade used higher frequency words and words acquired earlier than older students, while older students were able to access more unique and complicated vocabulary.

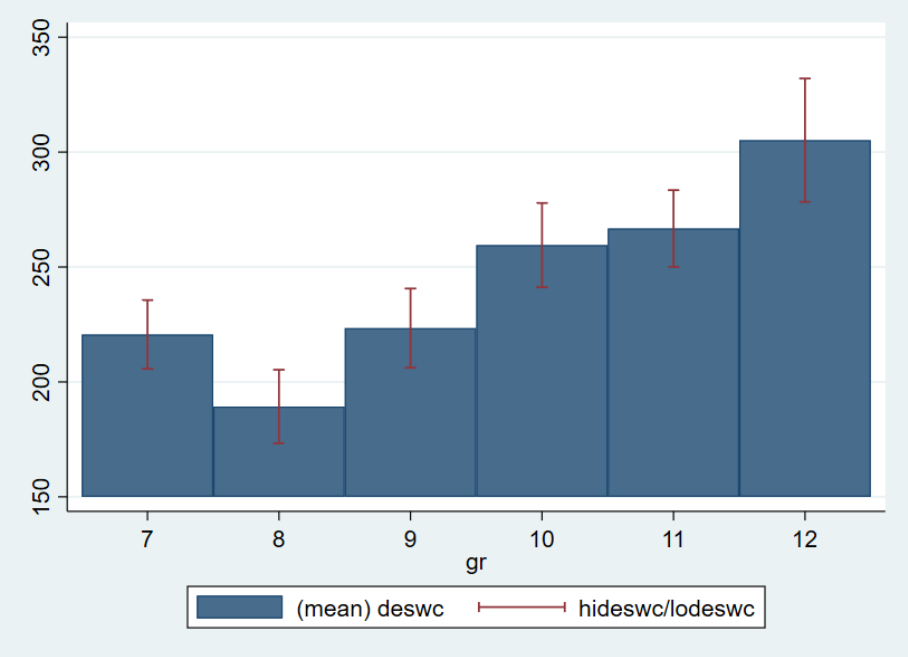

Figure 2. Word count by grade level.

We consider how our secondary school writers compare to basic college writers to analyze developmental differences. Fluency, or word count, is similar to the basic college writers. Cohesion, the overlap in content words between sentences (McNamara, 2014), contributed more to college writing quality (0.29) than to secondary writing quality (0.17). This is likely due to the lower ability of younger students to maintain a clear and consistent underlying claim in their writing. Secondary students relied more on connectives (0.06) than college students (-0.03, posttest) to achieve quality. The use of connectives may have made our writers' thinking more explicit and easier to follow -- both for themselves and for their readers. Our connectives latent variable included logical ("and"), adversative/ contrastive ("although"), and additive ("moreover") words to help create cohesive links between ideas and provide clues about text organization (McNamara et al., 2014).

Syntactic complexity had a negative relation to overall quality scores in our sample, just as it did in the basic college writers' sample (-0.31 for basic college writers and -0.06 for secondary

writers). The decreased negative impact of syntactic complexity for our writers may be attributable to the fact that it played a larger role in differentiating between higher and lower quality texts in the older writers (e.g., many of the secondary writers had run on sentences, so it did not provide as much discrimination). For our younger writers, the latent variable for syntactic complexity was less attributable to the mean number of words before the main verb (SYNLE, 0.50 and 0.60, for college writers pre and posttest respectively, and 0.40 for secondary writers). The uniformity and consistency of the syntactic constructions in the text across paragraphs was similar for your younger writers, as was the standard deviation of the mean length of sentences (a larger number indicates more variation in the essay sentence length).

Finally, the word level latent construct contributed less to our students' text quality than the basic college writers' sample (0.15 college posttest, 0.10 secondary). The composite measures underlying the latent construct also differed with respect to the indicator for the age when content words become part of a person's vocabulary, with a higher score indicating words learned later in life. Our model had a loading of 0.24 compared to 0.67 for the older students.

Comparison of means by gender found a significant difference on the holistic quality rating and a large difference in text length (word count of 218.5 for boys and 262.81 for girls). This finding is consistent with findings in other studies (Graham et al., 2017; Reilly et al., 2019; Steiss et al., 2022). Smaller, but statistically significant, differences are found for the cohesion variables (with boys scoring higher on all three components). The higher cohesion score parallels our finding that younger students had higher cohesion scores (although only in one component). Girls showed greater use of connectives, but only the adversative and contrastive ones -- again a parallel with our findings by grade level suggesting that these connectives, in particular, are indicative of more mature writers. No other gender differences were significant. Ultimately, in our model the difference in quality of essays was primarily due to the increased word count of women writers. This is consistent with the literature base that suggests women writers' performance reflects increased fluency (Reilly et al, 2018) and provides additional information that the performance increase does not reflect increased cohesion, word or syntax-level complexity, or use of connectives. Instead, the increased writing fluency may reflect unmeasured constructs such as oral fluency (Kim et al., 2014), attitudes (Pajares & Valiante, 1999), or transcription skills (McCutchen, 1996).

We also looked at mean differences across four levels of English language status: English learners, English only, initially fluent in English (IFEP), and redesignated fluent in English (RFEP). Not surprisingly, students writing in a language that they had been fluent in from the beginning of their education (English only and IFEP) wrote longer texts than the English learners or RFEP students. English learners had a mean word count of 198 compared to English only students of 248 and initially fluent students of 271. Redesignated students had a mean word count of 220, which was significantly different from their English only and initially fluent peers. Interestingly, however, this pattern changed slightly for the quality score, with only English learners having significantly lower scores and RFEP joining the fluent grouping with higher scores. The only significant differences in the component scores was at the word levels, with English learners using higher frequency (i.e., less sophisticated) than their proficient peers and IFEP students' containing more advanced vocabulary. While the former suggests still-developing English vocabulary, the latter might suggest a bilingual advantage consistent with Collier &

Thomson's work finding RFEP students outperforming English only students starting in middle schools (Thomas & Collier, 1997; Thomas & Collier, 2002).

Our results are slightly different from those found in Steiss et al. (2022) human analytic factors, in which language proficiency had a significant effect on performance in each of the evidence use, ideas/structure, and language use factors. Looking at the variables underlying their language use factors, we believe that many of the constructs in our model would be related to that factor. Thus, our model may be breaking down the components of that factor into even more discrete categories and discerning areas where language proficiency has the most impact. Given the text length differences, it is even more impressive that the initially fluent and redesignated bilingual students achieved holistic quality scores comparable to their English-only peers.

Looking at the means for above and below median essays showed a statistically significant difference in word count, with high scoring essays averaging 287 words and low scoring essays averaging 162 words. High-scoring essays also had higher cohesion, specifically both additive and adversative and contrastive connectives. Like McNamara et al. (2010), we do not find a difference between high and low scoring essays on the cohesion measures but, unlike them, we did find a difference in the use of connectives in our younger writers. One of the challenges of our NLP measure is that it does not differentiate between the different types of connectives; advanced writers use more sophisticated connectives, while beginning writers are likely to depend on simple ones such as "first," "second," and "next."

Despite some of the differences discussed above, we see that overall the model performed well with our secondary student text sample. When looking at our correlations among quality, length and linguistic indices, we find that the variables for this text set generally conform to the suggested parameters used by MacArthur et al. (2019) in the basic college writers study. First, correlation with essay length was less than $r = 0.20$ for all of our variables, with the slight exception of DESSLd, our sentence length measure which had $r = 0.21$. Next, correlation with other indices in the same construct was between $r = 0.90$ (to avoid collinearity) and $r = 0.30$. Our cohesion and connectives met these criteria, but our syntactic measure was slightly below the desired $r = 0.30$ or above with respect to the correlation between SYNLE and SYNSTRUTt ($r = 0.20$) and SYNLE and DESSLD ($r = 0.27$), as was our word level variable, with the correlation of WRDAOAc and DESWLsy $r = 0.24$ and WRDFRQa and WRDAOAc $r = 0.13$. Correlation between word count and quality were high as expected, $r = 0.64$, slightly higher than for basic college writers ($r = 0.56$), which supports our understanding that text length is a more significant predictor in younger, less developed writers who struggle with production. Given the fact that this was a time-limited task, it is not surprising that text length predicts 41% of the variance in quality score.

As seen in prior research, there continues to be a correlation of quality with length, which is even stronger in our younger writers than the college writers at posttest. Before less mature writers can successfully employ the constructs in our model, they must first be fluent enough to produce sufficient text. The one area this was not true was in the use of connectives, which we hypothesize our younger writers used to support text organization and clarity. Syntactic complexity was a negative indicator for both sets of students, suggesting that basic college

writers struggled with ineffective complex sentence structure and run-on sentences just like our younger students.

We believe that these analyses provide several contributions: we use the MacArthur et al. (2019) model with younger students and our larger sample allows us to consider variables such as grade level, gender, English language status, and higher versus lower scoring essays. Looking across grades we see that the text length of younger students increased as they aged, but the model otherwise was fairly stable. Gender did not seem to affect the model in meaningful ways beyond the increased fluency of women writers. We saw text length and word level differences, but not holistic quality scores differences, between initially designated and redesignated bilingual students compared to their English-only peers. Finally, we see that the model works better with our higher scoring essays and is less effective explaining the lower scoring essays.

These findings suggest that a deeper analysis of the writing of emerging bilingual students would be valuable to ensure that the writing curriculum not only meets student needs but takes advantage of student strength. We also think an important next step is to understand the impact different genres, prompts, and task attributes (e.g., timed versus untimed) have in the model and underlying variables. We also note a concern with using Coh-Metrix variables in this way: the variables do not differentiate appropriate, high quality use of connectives, syntactic complexity, and lexical complexity from poor uses. They only indicate the frequency of such usage. While the failure to differentiate between appropriate uses of complex words, for example, or complex sentences versus run-on sentences may be less concerning when analyzing the writing of competent adult writers, it becomes more problematic when trying to understand the meaning of variables on writing quality for younger writers.

References

Applebee, A. N. (1990). Learning to Write in Our Nation's Schools: Instruction and Achievement in 1988 at Grades 4, 8, and 12. Report No. 19-W-02. National Assessment of Educational Progress, Educational Testing Service, Rosedale Rd., Princeton, NJ 08541-0001.

Ateş, C., Kaymaz, Ö., Kale, H. E., & Tekindal, M. A. (2019). Comparison of Test Statistics of Nonnormal and Unbalanced Samples for Multivariate Analysis of Variance in terms of Type-I Error Rates. *Computational and Mathematical Methods in Medicine*, *2019*, 2173638. https://doi.org/10.1155/2019/2173638

Bang, H. J. (2013). Reliability of national writing project's analytic writing continuum assessment system. *Journal of Writing Assessment, 6*(1), 13–24.

Berninger, V. W., Abbott, R. D., Abbott, S. P., Graham, S., & Richards, T. (2002). Writing and reading: Connections between language by hand and language by eye. *Journal of Learning Disabilities, 35*(1), 39-56.

Berninger, V. W., Nagy, W., & Beers, S. (2011). Child writers' construction and reconstruction of single sentences and construction of multi-sentence texts: contributions of syntax and transcription to translation. *Reading and Writing, 24*(2), 151–182. https://doi.org/10.1007/s11145-010-9262-y

Brandt, D. (2014). The rise of writing: Redefining mass literacy. Cambridge University Press.

Center for Research on Education Outcomes (CREDO, 2020). Estimates of learning loss in the 2019-2020 school year. https://credo.stanford.edu/sites/g/files/sbiybj6481/f/short_brief_on_learning_loss_final_v.3.pdf

Coker, D. (2006). Impact of first-grade factors on the growth and outcomes of urban schoolchildren's primary-grade writing. *Journal of Educational Psychology, 98*(3), 471.

Crossley, S. A. (2020). Linguistic features in writing quality and development: An overview. *Journal of Writing Research, 11*(3), 415-443.

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 (*4), 66-79. https://doi.org/10.1016/j.jslw.2014.09.006

Crossley, S. A., Weston, J., McLain-Sullivan, S. T., & McNamara, D. S. (2011). The development of writing proficiency as a function of grade level: A linguistic analysis. *Written Communication, 28* (3), 282-311. doi: 10.1177/0741088311410188

Dorn, E., Hancock, B., Sarakatsannis, J., & Viruleg, E. (2020, December 8). COVID-19 and learning loss—disparities grow and students need help. McKinsey & Company.

https://www.mckinsey.com/industries/public-and-social-sector/our-insights/covid-19-andlearning-loss-disparities-grow-and-students-need-help.

Flower, L., & Hayes, J. R. (1981). A cognitive process theory of writing. *College Composition and Communication, 32*(4), 365-387.

Graesser, A.C., McNamara, D.S., Louwerse, M.M. et al. Coh-Metrix: Analysis of text on cohesion and language. *Behavior Research Methods, Instruments, & Computers 36*, 193–202 (2004). https://doi.org/10.3758/BF03195564

Graham, S. (2019). Changing how writing is taught. *Review of Research in Education, 43*(1), 277–303 doi: 10.3102/0091732X18821125

Graham, S., Harris, K. R., Kiuhara, S. A., & Fishman, E. J. (2017). The relationship among strategic writing behavior, writing motivation, and writing performance with young, developing writers. *The Elementary School Journal, 118*(1), 82-104.

Graham, S., & Perin, D. (2007). *Writing next-effective strategies to improve writing of adolescents in middle and high schools– A report to Carnegie Corporation of New York*. Washington, DC: Alliance for Excellent Education.

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

Hillocks, G. (2011). Commentary on" Research in Secondary English, 1912—2011: Historical Continuities and Discontinuities in the NCTE Imprint". *Research in the Teaching of English, 46*(2), 187-192.

Kim, Y. S., Al Otaiba, S., Folsom, J. S., Greulich, L., & Puranik, C. (2014). Evaluating the dimensionality of first-grade written composition. *Journal of Speech, Language, and Hearing Research* (57)1, 199-211. doi:10.1044/1092-4388(2013/12-0152)

Kim, Y.-S. G., & Graham, S. (2022). Expanding the Direct and Indirect Effects Model of Writing (DIEW): Reading–writing relations, and dynamic relations as a function of measurement/dimensions of written composition. *Journal of Educational Psychology, 114*(2), 215–238. https://doi.org/10.1037/edu0000564

Kim, Y. S. G., Park, C., & Park, Y. (2015). Dimensions of discourse level oral language skills and their relation to reading comprehension and written composition: An exploratory study. *Reading and Writing, 28*(5), 633-654.

Kim, Y. G., & Schatschneider, C. (2017). Expanding the developmental models of writing: A direct and indirect effects model of developmental writing (DIEW). *Journal of Educational Psychology, 109*(1), 35–50. https://doi.org/10.1037/edu0000129

Kim, Y. S., Wagner, R. K., & Foster, E. (2011). Relations among oral reading fluency, silent reading fluency, and reading comprehension: A latent variable study of first-grade readers. *Scientific Studies of Reading, 15*(4), 338-362.

Kogan, V. & Lavertu, S. (2021) How the COVID-19 pandemic affected student learning in Ohio: Analysis of spring 2021 Ohio state tests. Ohio State Univ. Retrieved October 19, 2021, from http://glenn.osu.edu/educational-governance/reports/reports-attributes/210828_KL_OST_Final.pdf

Kuhfeld, M., Soland, J., Tarasawa, B., Johnson, A., Ruzek, E., & Liu, J. (2020). Projecting the potential impacts of COVID-19 school closures on academic achievement. *Educational Researcher,* 49(8), 549- 565. https://doi.org/10.3102/0013189X20965918

MacArthur, C. A., Jennings, A., & Philippakos, Z. A. (2019). Which linguistic features predict quality of argumentative writing for college basic writers, and how do those features change with instruction? *Reading and Writing, 32*(6), 1553-1574.

McCutchen, D. (1986). Domain knowledge and linguistic knowledge in the development of writing ability. *Journal of Memory and Language, 25*, 431-444. doi: 10.1016/0749-596X(86)90036-7

McCutchen, D. (1996). A capacity theory of writing: Working memory in composition. *Educational Psychology Review, 8*(3), 299–325. https://doi.org/10.1007/BF01464076

McCutchen, D., Covill, A., Hoyne, S. H., & Mildes, K. (1994). Individual differences in writing: Implications of translating fluency. *Journal of Educational Psychology, 86*, 256–266. doi:10.1037/0022-0663.86.2.256

McCutchen, D., & Perfetti, C. (1982). Coherence and connectedness in the development of discourse production. *Text, 2*, 113-139.doi: 10.1515/text.1.1982.2.1-3.113

McNamara, D. S., Crossley, S. A., & McCarthy, P. M. (2010). The linguistic features of quality writing. *Written Communication, 27* (1), 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* (2), 499-515. doi: 10.3758/s13428-012-0258-1

McNamara, D. S., & Graesser, A. C. (2012). Coh-Metrix: An automated tool for theoretical and applied natural language processing. In Applied Natural Language Processing: Identification, investigation and resolution (pp. 188-205). IGI Global.

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

Muthén, L.K. and Muthén, B.O. (2017). Mplus. Eighth Version. Los Angeles, CA: Muthén & Muthén

National Center for Education Statistics (NCES, n.d.). NAEP achievement levels. https://nces.ed.gov/nationsreportcard/writing/achieve.aspx

National Center for Education Statistics. (2012). The nation's report card: Writing 2011 (NCES 2012-470). Institute for Education Sciences; U.S. Department of Education. https://nces.ed.gov/nationsreportcard/pubs/main2011/2012470.aspx

Olson, C.B., Matuchniak, T., Chung, H.,Stumpf, R., & Farkas, G. (2017). Reducing achievement gaps in academic writing for Latinos and English learners in Grades 7–12. *Journal of Educational Psychology, 109*(1).

Olson, C. B., Woodworth, K., Arshan, N., Black, R., Chung, H. Q., D'Aoust, C., Dewar, T., Friedrich, L., Godfrey, L., Land, R., Matuchniak, T., Scarcella, R., & Stowell, L. (2019). The Pathway to Academic Success: Scaling Up a Text-Based Analytical Writing Intervention for Latinos and English Learners in Secondary School. *Journal of Educational Psychology, 112*(4), 701.

Pajares, F. & Valiante,G. (1999). Grade level and gender differences in the writing self-beliefs of middle school students. *Contemporary Educational Psychology, 24* (4) (1999), pp. 390-405.

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

Powers, D. E. (2005). Wordiness: A selective review of its influence, and suggestions for investigating its relevance in tests requiring extended written responses. Online Research Memorandum, Educational Testing Service. Available: http://www. ets. org/Media/Research/pdf/RM-04-08. pdf.

Reilly, D., Neumann, D. & Andrews, G. (2019). Gender differences in reading and writing achievement: Evidence from the National Assessment of Educational Progress (NAEP). *American Psychologist, 74* (4).

Smith, M. W., Wilhelm, J. D., & Fredricksen, J. E. (2012). Oh, yeah?!: Putting argument to work both in school and out. Heinemann.

Spenader, J. (2018). Children's comprehension of contrastive connectives. *Journal of Child Language, 45*(3), 610-640. doi:10.1017/S0305000917000423

Steiss, J., Krishnan, J., Kim, Y. S. G., & Olson, C. B. (2022). Dimensions of text-based analytical writing of secondary students. *Assessing Writing, 51,* 100600.

Thomas, W., & Collier, V. (1997). *School effectiveness for language minority students*. Washington, DC: National Clearinghouse for Bilingual Education.

Thomas, W., & Collier, V. (2002). *A national study of school effectiveness for language minority students' long-term academic achievement.* Santa Cruz: Center for Research on Education, Diversity and Excellence (CREDE).