This is a preprint of a paper that will appear in *Reading Psychology*.

Does One Size Fit All?

Exploring the Contribution of Text Features, Content, and Grade of Use on Comprehension

Heidi Anne Mesmer, Virginia Tech

Elfrieda H. Hiebert, TextProject

James W. Cunningham, University of North Carolina

Madhu Kapania, Virginia Tech

Abstract

Readability systems have once more become prominent in policy and practice because of recommendations in the Common Core State Standards. This study revisited two features of current text analysis (readability) systems: their generalizability to all grade levels and to all content areas. A database that encompassed texts across the grade bands and content areas and included aggregate comprehension performance on the texts was used to: (a) describe how the text features (i.e., word frequency, word length, sentence length) varied at different grade levels and within different subject areas and (b) examine if the prediction of comprehension with the text features was moderated by the grade or content area of the text. Results indicated that texts did having differing levels of various word features along both grade and content lines especially in the area of sentence length. In addition, content and grade moderated the relationship between sentence length and comprehension.

Keywords: Reading, text, readability

Does One Size Fit All?

Exploring the Contribution of Text features, Text content, and Grade of Use on

Today, and for many decades, almost all readability formulas have taken a one-size-fits-all approach (Klare, 1984) meaning that the same formulas are applied to texts across content areas (e.g., science, English) and grade levels of use. Using the same formula with texts from all grades may not be best due to differences in texts at various levels of schooling and readers at various stages of development. Using the same formula with texts from all subject areas may not be best due to differences in content-specific textual demands. For example, an informational text on whales for first graders developing fluency with the 300 most-frequent words and *Finnegan's Wake* (Joyce, 1939) offered as part of a twelfth-grade International Baccalaureate program may differ in kind and not merely in degree.

A hiatus occurred in the use of readability formulas after the publication of *Becoming a Nation of Readers* (Anderson, Hiebert, Scott, & Wilkinson, 1985). However, recommendations on text levels within the Common Core State Standards for the English Language Arts (CCSS) have reversed this pattern (National Governors Association Center for Best Practices (NGA) & Council of Chief State School Officers (CCSSO), 2010a, 2010b). Due to a perceived gap between the complexity of texts in college/career and those in high school, minimum and modal text difficulties across grades have been quantified using a set of readability formulas and increased (Appendix A). At present, there is a heavy use of formulas, particularly the one promoted in Appendix A of the CCSS, and yet, a recent study concluded, as had previous studies, that text features and formulas performed differently and often required adjustments at various grades and text types (Deane, Sheehan, Sabatini, Futagi, & Kostin, 2006; Fry, 1969; Nelson, Perfetti, Liben, & Liben, 2012; Spache, 1953).

Five previously formulas are currently in heavy use (e.g., Lexiles, ATOS, Degrees of Reading Power, Reading Maturity Measure (RMM), Source Rater). All rely on at least two long-recognized text factors (Klare, 1984): a measure of word complexity and a measure of sentence complexity to estimate text difficulty. In almost all, the measure of sentence complexity is mean number of words per sentence. The most common word factors are a measure of word frequency and/or a measure of word length. This study was based on the assumption that testing how well variables in current formulas (specifically, variables of word and sentence complexity) predict student comprehension could shed light on their efficacy across grades and content areas. Specifically, we examined how measures of word complexity (i.e., word frequency, word length) and syntactic complexity (i.e., sentence length) varied across text at different grade levels and within different content areas (e.g. science, social studies). In addition, we examined if the grade and/or content of a text moderated the relationship between these word and sentence-level features and comprehension. As Figure 1 shows, we were curious as to whether grade or content altered the already well-known relationship between text features (i.e., word frequency, word length, sentence length) and comprehension.

Theoretical and Empirical Rationale

Theoretically, the study was grounded in three areas: (a) approaches to predicting text difficulty (see Authors, 2012a; 2014; Klare & Buck, 1954), (b) developmental theory (Chall, 1983), and (c) distinctions among text types (Biber, 1988; Graesser, MacNamara, & Kulikowich, 2011; Karlsson, 2009; Saukkonen, 2007). To answer this study's first inquiry, we compared the word and sentence features of texts at different grades and in different content areas in order to address variation in their text complexity. In our second inquiry, we examined how word and

sentence complexity variables predicted comprehension of texts and if the predictions were moderated by grade or content area.

Approaches to Measuring and Predicting Text Difficulty

Readability formulas are essentially regression equations that use independent predictor variables (e.g. text features) to predict a criterion variable representing text difficulty (e.g., the grade level assigned to a text, a teacher's rating of a text, the students' reading of a text). The predictor variables are text complexity variables derived from the features of texts that can be enumerated and fashioned into continuous variables (e.g., word length in letters, word frequency). Criterion variables have included student reading comprehension performance, previously established text levels, teacher ratings, publisher's designations, and other scores. Early in the long history of readability, student performance measures as criteria were rare. Since the mid-1960s, mean student comprehension of passages has gradually come to be considered the gold standard criterion variable for developing and validating readability formulas (Authors, 2014).

Through a process of statistical modeling, a readability formula is derived and refined by relating predictors to the criterion. Once established, the formula allows a user to collect the features of a text, enter the counts of text features into the equation, and then receive a label for the text's *predicted* difficulty. The resulting label is an *estimate* of text difficulty, rather than an infallible calculation. There are longstanding trends, patterns, and issues in variables used on both the predictor and criterion side of formulas, points that we make in the section below.

As has already been discussed, both historical and current readability systems rely on predictor variables that include one or more measures of word complexity and a sentence or syntactic complexity measure, which almost always has been mean number of words per

sentence (Klare, 1984). At the word-level, developers of first-generation readability formulas (i.e., those computed by hand or mechanically) often identified a list of target high frequency words. The proportion of words in a text that did not appear on the high frequency list was used as the measure of word complexity (e.g. Dale & Chall, 1948; Spache, 1953). In digital readability systems, a continuous variable is typically calculated based on the frequency of each word in a text sample. Assigned frequencies are derived from large corpora numbering in the millions (e.g. Lexiles, ATOS). Using this approach the word *and*, for example, might be assigned a frequency of 10,000, reflecting its occurrence per million words whereas the word *beanstalk* might be assigned a frequency of 1. After obtaining the corpus frequency of each word in a text, the frequencies are usually transformed logarithmically to address the large range (e.g., 1-10,000) and then a mean log word frequency is calculated for each text to serve as the word complexity predictor in a formula.

A second word factor predictor, word length, has been measured in the following ways: a) number of syllables (Flesch,1943; Fry, 1969); b) number of words with one-syllable words (e.g., Farr, Jenkins, & Patterson, 1951); c) number of words with 3+ syllables (e.g., Gunning, 2003; McLaughlin, 1969). Many current digitized tools include measures of word length (e.g. Source Rater, Degrees of Reading Power, ATOS, Reading Maturity, Coh-Metrix).

Developmental Perspective

The focus on the utility of readability formulas at different levels in this study emanated from developmental theory and the grades typically associated with those levels. A researcher who devoted a career to the study of text difficulty—Jeanne Chall (Dale & Chall, 1948; Chall, Bissex, Conard, & Harris-Sharples, 1996)—proposed that reading development needs to be viewed as progressing through a series of six stages of reading. These stages extend from birth to

kindergarten, when students are developing the competences in oral language, phonological awareness, and concepts of print and stories, to adulthood when individuals analyze, synthesize, and use their own interpretations as readers. Although development will not always conform explicitly to grade levels, trends in normal development do tend to show grade level patterns. According to Chall, in Stages 1 and 2 (usually through grade 3) readers are first learning to decode (Stage 1) and then are becoming more fluent and efficient (Stage 2). In Stage 3, a period between grades 4 and 8, readers are "reading to learn the new" as they are mastering comprehension. At the secondary level and beyond, Stage 4, students are becoming critically literate and learning how to evaluate texts on many levels. The question remains as to whether readability formulas perform equivalently in predicting the challenge of text for readers at different grades and developmental stages. The application of developmental perspectives to the quantitative prediction of text difficulty, however, has been limited, even in Chall's own quantitative system (Dale & Chall, 1948). There is, however, evidence that certain approaches may be more or less associated with specific developmental levels and grades.

In particular, soon after use of readability formulas became widespread, Spache (1953) selected a word list different from that used by Dale and Chall (1948). Spache's list and the resulting formula were expressly designed for Grades 1 and 2. Further, an assumption underlying Fry's (1969) use of word length in his readability formula was the relationship between word length and decoding ease. In reading acquisition, shorter words are typically decoded more easily than longer ones (Ehri, 2005). By having a curvilinear rather than linear relationship in his graph between text feature amounts and estimated overall difficulty, Fry seemed to recognize that the impact of those features on difficulty varied over the developmental continuum.

Except for the efforts of Spache (1953) and Fry (1969), research to discover whether there are inconsistencies in the nature and effects of word length, word frequency, and sentence length at different grade levels has been rare. An exception is a recent analysis of seven readability formulas conducted by Nelson et al. (2012). They compared formulas' abilities to differentiate texts within a grade band (i.e., 3-5, 6-8, 9-11). The formulas included Lexile (MetaMetrics), ATOS (Renaissance Learning), Degrees of Reading Power: DRP Analyzer (Questar Assessment, Inc.), REAP (Carnegie Mellon University), SourceRater (Educational Testing Service), Pearson Reading Maturity Metric (RMM; Pearson Knowledge Technologies) and (Coh-Metrix, University of Memphis). With the exception of the Lexile measure, all of the tools included word length, sentence length, and word frequency, often along with other measures.

Using two of their reference measures, the study found that the formulas tended to have higher predictive power with texts from the lower rather than the higher grades. Most formulas performed best in the grades 3 to 5 band and were increasingly less effective at the middle (6-8) and high school grades (9-11). An exception was the data for the RMM, which proved to be a strong predictor for SAT performances at Grades 9 through 11. The finding for RMM is a critical one because it is the first readability system to replace word frequency with a factor called word maturity, a word complexity measure first introduced by Landauer, Kireyev, and Panaccione (2011), which estimates how word knowledge grows, develops, and deepens, as students gain higher levels of text exposure.

A study by Deane, et al., (2006) exposed a curious trend in intermediate grade texts with respect to word frequency. The authors contrasted texts at grades three and six and found nearly identical levels of mean log word frequency at the two grades but distinct sentence complexity.

Basically, sentences were longer in grade six texts but the frequency of words was distributed identically at the two levels. That is, texts at grades three and six did not differ with respect to word frequency.

Together word length, word frequency, and sentence length appear to be less predictive of secondary grades' text difficulty and more predictive of primary grades' difficulty despite the fact that these tools are used through grade 12. However, few studies have investigated how the word frequency measure performs at particular grade spans.

Distinctions in Text Difficulty by Text Type

For quite some time researchers have distinguished different text types and the most common distinction has been that of genre, a concept that while widely used, has been fuzzy, at best (see Karlsson, 2009). Recently, citing the overlap of text structures and genres, researchers have organized texts into disciplines or content areas such as literary, scientific, or social science (Graesser, et al., 2011; Shanahan & Shanahan, 2008). In this section, we review the limited literature regarding how readability tools perform across different text types.

The dominant model for distinguishing text types, narrative and expository, recognizes two different purposes, resulting in distinct text structures and use of language (e.g. Biber, 1988; Coté, Goldman, & Saul, 1998). Despite extensive scholarship that describes differences in genres (e.g., Biber, 1988) and numerous investigations to establish differences in readers' comprehension of different genres (e.g., Best, Floyd, & McNamara, 2008; Cervetti, Bravo, Hiebert, Pearson, & Jaynes, 2009; Saénz & Fuchs, 2002; Tun, 1989), attention to differences in genre in text complexity systems has been limited until recently (see Klare, 1984).

Recognition of the potential contribution of narrative and expository genres to the assessment of texts has increased with the Common Core (NGACBP & CCSSO, 2010a). In the

model of text in the CCSS, writers (NGACBP & CCSSO, 2010b) claimed a difference in the efficacy of quantitative measures of narrative texts, stating: "*Many current quantitative measures underestimate the challenge posed by complex narrative fiction."* (italics in original, p. 8). The primary reason for this misclassification comes from high use of very frequent words in dialogue. However, no references are cited as the basis of this conclusion.

One of the aims of the Nelson et al. (2012) study was to provide evidence on the efficacy of different text systems for assessing narrative and expository texts. Overall, the six text systems had better correlations for informational than narrative texts, although RMM correlated highly for both. Further, SourceRater showed differentiation in the complexity of narrative texts at the upper grade levels, which other systems did not do. SourceRater is the only available text analysis tool that has separate algorithms for narrative, expository, and mixed (i.e., narrative/expository) texts, a decision based on findings showing that indices such as sentence length and word frequency consistently overestimated the difficulty of expository texts and underestimated the difficulty of narrative texts (Sheehan, et al., 2008). The degree to which differences between these measures in assessing text complexity can be tied to particular features of vocabulary, syntax, or word frequency is less clear. In studies where texts have been manipulated to examine the effects of lexical and syntactic complexity, lexical complexity has affected readers' comprehension more than syntactic complexity (Arya, Hiebert, & Pearson, 2011; Droop & Verhoeven, 1998). Simple syntactic changes such as eliminating connective words (e.g., *but, while*) can increase the inference burden on readers (Ozuru, Dempsy, Sayroo, & McNamara, 2005; Pearson, 1974). Analyses of the syntactic patterns within texts of different genres have not been a research focus. Whether narrative texts have more variability in syntax,

including shorter sentences used in dialogue has been hypothesized (O'Shea, Bandar, Crockett, & McLean, 2011), but has not been investigated.

By contrast, the availability of digital corpora has led to a number of studies that begin to clarify the lexical composition of narrative and expository texts. In an examination of informational and literacy texts from a million-word corpus downloaded from the British National Corpus, Lee (2001) found that 2,000 common words accounted for 81 to 84% of the words in literary texts (including fiction, poetry, and drama); expository texts had percentages in the range of 66 to 71%. Subsequently, Gardner (2004) considered the presence of unique words beyond the common words studied by Lee. Gardner found that narrative texts had more unique words than informational texts. Further, the majority of the unique words did not overlap between the two text types.

In an analysis of words chosen for instruction in fourth-grade English Language Arts (ELA) and science programs, ELA text had substantially more unique words than the science text and more of these unique words were rare (i.e., less than one predicted appearance per million words of text) (Authors, 2012b). This finding challenges Lee's (2001) conclusions regarding the role of the core vocabulary in different genres, as does another recent study by Authors (2017) where the definition of core vocabulary was expanded to include the *morphological family members* of the 2,411 most frequent words. The text corpus in the 2017 study consisted of the exemplars identified in Appendix B of the CCSS and narrative texts had approximately 1% more words from the core vocabulary than did the expository texts—a percentage considerably smaller than that reported by Lee (2001).

Although it might seem obvious to categorize texts by expository and narrative genres, researchers have not been able to decide conclusively on those labels. In the US, the expository

label is often replaced by the term informational. In the United Kingdom, four main genres are identified, literacy, expository, procedural, and reference, which are broken into sub genres (Department of Education Sciences, 1993). Quantitative analyses complicate the issue of genre even more. In a 2007 study, Saukkonen generated factors using up to 66 linguistic variables and identified up to 21 genres as well as a more inclusive set of six genres. In a similar study, Biber (1988) used 67 linguistic variables to analyze 23 different genres of speech and writing, identified six factors/potential genres. One current readability formula, SourceRater, identifies a "mixed" expository/narrative category label, which further supports the "blurriness" of the narrative/expository distinctions noted by others (e.g., Karlsson, 2009). As Lefstein and Snell (2011) explained, "Genre is a relatively fuzzy concept, used in multiple ways and for a variety of purposes in different research traditions" (p. 41).

Given the complexities of narrative and expository genre labels, we turned to schemes that categorized texts by content area or discipline, a widely accepted idea (Shanahan & Shanahan, 2008). Within this perspective, texts are organized by subject, content, or discipline, such as social science, science, math, and literature. Indeed, in an innovative study, Graesser et al. (2011) used the multifaceted Coh-Metrix system to analyze texts by grade band and by content area, language arts, science, and social studies. Using a principal components analysis, the study showed differences amongst the content areas and grades of texts in the components of word concreteness, referential cohesion, causal cohesion, syntactic simplicity, and narrative. Language arts texts were higher on the narrativity scale, which included log word frequency, content word frequency, minimum word frequency, familiarity, and age of acquisition among many other variables.

In summary, the literature reflects categorizations of texts that address how narrative and expository genres differ in the two measures that make up most readability formulas—word and syntactic complexity and yet they have been limited and typically have taken a generic view of genre. Recent studies have analyzed texts using the more pragmatic and functional content area categories, labels that typify texts students read in schools. This latter approach shows promise and avoids ongoing disputes about which genres exist.

Research Questions

A review of current and past formulas shows a heavy reliance on word frequency, word length, and sentence length despite evidence that these features exist at different amounts in different grades and contents and thus, may influence comprehension differently. For this reason, this study used the Bormuth (1969) dataset, described below, to address the following questions:

Question 1: Text Complexity

What is the influence of grade band and content on the levels of text features (i.e., word frequency, word length, sentence length) in materials? How do texts vary on features by grade band and content?

Question 2: Text Difficulty and Reader Performance

Do grade and content area of text moderate the relationship between text features (word frequency, sentence length, word length) and comprehension?

Methods

Data Source

Bormuth (1969) used passages from published instructional materials for his criterion variable. Three hundred and thirty texts were selected from five grade bands (Grades 1-3, 4-6, 7-9, 10-12, and University) and ten school subjects (e.g., biology, chemistry, civics, current news,

economics, geography, history, literature, mathematics, and physics). Developmentally, the grade bands aligned with Chall's (1983) stages: a) Grades 1-3 (Chall Stages 1-2); b) Grades 4-6 & 7-9 (Chall Stage 3); and c) Grades 10-12, University (Chall Stage 4). A passage was randomly chosen from each text.

Bormuth (1969) used traditional cloze as the comprehension measure. Five different deletion patterns were applied to each passage to produce 1,650 cloze tests. Participants were about 2,600 middle-class students from schools in Minneapolis suburbs. About 500 were in grades 4-6, 1,000 in grades 7-9, and 1,000 in grades 10-12. Based on scores from the *California Reading Achievement Test* (Tiegs & Clark, 1963), participants were assigned to 50 matched groups. Fifty booklets each consisting of 33 randomly chosen and ordered tests were made for the groups; no booklet did repeated passages. Cloze testing was untimed. Exact replacement of a deleted word was required with spelling errors allowed. In total, there were 94,050 test protocols. To estimate reliability, Bormuth reported correlations between random halves of items across passages as .89 and .94 (corrected using the Spearman-Brown prophecy formula).

Since the 1960s, most researchers have preferred deletion items to questions when assessing comprehension for text complexity research. For example, the three most widely used readability formulas today are the Flesh-Kincaid, the DRP, and Lexile, all developed and validated using a deletion measure—traditional cloze, multiple-choice cloze, and the native item type, respectively. Bormuth (1971) defends traditional cloze: "tests [composed of comprehension questions] are subject to unpredictable variations in the size of the mean scores, variations that are due to the uncontrolled ways test writers select and phrase the items included in the tests" (p. 3). He argues that a criterion measure for text research should be affected only "by the characteristics of the passage itself and not by any other source of systematic variance" (p. 26),

because “the variance of interest [is] the between passage variance” (p. 27). In short, traditional cloze is text-dependent and lacks the test-constructor variance inherent in most multiple-choice distractors and comprehension questions.

Despite its age, Bormuth’s (1969) dataset is arguably the most rigorously developed criterion variable (measure of text difficulty) to date. For this study, we obtained Bormuth’s 330 passages as well as grade band of use (1-3, 4-6, 7-9, 10-12, or university), subject area of use (e.g., biology), and mean cloze score across test forms for each.

Coding of Texts

For each passage, the researchers obtained information on three text features. The Lexile Analyzer at lexile.com provided information on two features: the mean log word frequency (MLWF), and the mean sentence length (MSL). We transformed the MSL to the natural log (LMSL) because that is what the Lexile equation uses. The mean word length in syllables (Sylls/Wd)—came from Readability.com. To ensure the reliability of Readability.com’s counts, we randomly selected 10 of the 330 Bormuth passages and hand-counted syllables for each and then compared the results to those obtained through Readability.com. The mean number of syllables in the 10 passages was identical (167.3) in the two methods; the mean difference between the counts was 1.6 syllables per passage.

Bormuth’s (1969) texts represented ten “subject” labels (i.e., literature, math, current news, science, chemistry, biology, physics, geography, history, civics, economics). For this study, the ten categories were collapsed into five content area groups *literature*, *math, current news*, *science* (i.e., chemistry, biology, physics) and *social sciences* (i.e., geography, history, civics, economics) matching Karlsson’s (2009) criteria that groups be organized based on how they fit different contexts and settings.

Procedure

For question 1, we examined the text features, word frequency, word length, and sentence length, across grade and content area using ANOVAs. For question 2, the criterion variable in the dataset was the aggregate comprehension score (CM) across the five test forms for each of the 330 passages. We used the text features to predict CM and then handled grade and content as moderator variables in separate regressions.

Results

Question 1: Influence of Grade Band and Content Area on Word Frequency, Word Length, and Sentence Length

For this first question, we examined the influence of grade band and content on word frequency, sentence length, and word length. We used three two-way ANOVAs (grade band x content), one for each of the dependent variables (i.e., word frequency, sentence length, word length). The grade band factor had five levels (i.e., 1-3, 4-6, 7-9, 10-12, university) and the content area factor had five levels (i.e., literature, math, science, social science, and news). In each of the analyses, we used Levene's Test to assess the homogeneity of variances and applied the correction for the word length analysis. Bonferroni corrections were used as needed for post hoc tests.

Influence of grade band. Table 1 shows means and standard deviations for word frequency, sentence length, and word length by grade and content. Generally, as grade level increased, words became less frequent. In addition, as grade level increased, word length and sentence length increased.

The correlations between text feature variable were as expected and in keeping with decades of research. Word frequency and sentence length had a small negative correlation (r =-

.23) meaning that higher word frequencies (i.e., more familiar, easier words) were associated with shorter sentence lengths. A similar, but stronger relationship existed between word frequency (MLWF) and word length (r = -.69). Higher word frequencies (i.e., more familiar, easier words) were associated with shorter word lengths, another expected pattern. The relationship between sentence length and word length (r = .44) showed that as words got longer, sentences did as well.

Influence of grade band on word frequency. For word frequency there was a main effect for grade (F (4, 298) = 21.24, p<.001, η_p^2= .22) and content (F (4, 298) = 15.98, p<.001, η_p^2=.17) but no significant interaction between grade and content (F (16, 298) = 1.65, p = .06, η_p^2=.05). Table 1 shows means and standard deviations for word frequency by grade. Words were more frequent at lower grades and became less frequency at the higher grades.

Within readability theory, words should become less frequent as grades *increase* but post hoc analyses did not consistently reflect this pattern at all grade bands. Adjacent grade bands 1-3 and 4-6 were *not* significantly different in mean log word frequency (M= 3.68, SD= .15 vs. M=3.61, SD=.64). The words in texts in grades 1-3 were *not* more frequent (easier) than those in grades 4-6. However, texts in grades 1-3 did have significantly more frequent words (M= 3.68, SD= .15) than those in grades 7-9 (M=3.56, SD=.16), 10-12 (M=3.49 SD=.16), and university (M=3.37, SD=.20) (all ps <.001). Texts at grades 4-6 had more frequent words than those in grades 10-12 and university (all ps <.001) but not significantly more frequent words than texts in grades 7-9. Thus, word frequencies in grades 1-3 and 4-6 were not different from each other and word frequencies in grades 4-6 and 7-9 did not differ either.

Texts in grades 7-9 had significantly less frequent words (i.e., less familiar, harder words) than those in grades 1- 3 (p < .001) and significantly more frequent words (i.e., easier words)

than those at grades 10-12 (p<.05), and university (p < .001). However, as previously mentioned, grade 7-9 texts were not different in word frequency from 4-6 texts.

Texts in grades 10-12, had significantly less frequent words than those at grades 1-3, 4-6 (ps < .001) and 7-9 (p < .05), and significantly more frequency words than those at the university level (p<.05).

Lastly, texts at the university level, had significantly less frequent words than those in grades 1-3, 4-6, 7-9 (ps < .001) and 10-12 (p<.05). The overarching trend in the grade analysis was that words became less frequent as grade increased but some adjacent grade bands *did not differ* in word frequency (i.e., 1-3 vs. 4-6, 4-6 vs. 1-3, and 4-6 vs. 7-9). This suggested that word frequency was not a precise differentiator between texts at certain grade bands.

Influence of grade band on word length. Table 1 shows the mean word lengths by grade band. The trend was that of word lengths increasing by grade. Mean word length differences mirrored word frequency findings, with main effects for grade (F (4, 298) = 44.72, p <.001, η_p^2=.38) and content (F (4, 298) = 13.9, p <.001, η_p^2=.016), and an insignificant interaction between grade and content (F (16, 298) = 1.46, p<.11, η_p^2=.07).

Post hoc results showed that *all* grades were significantly different in the area of word length with p values less than .001 for all comparisons except grades 10-12 vs. university where the p value was slightly higher (p =.01). Thus, every grade band differed significantly in word length from every other grade band, making word length a completely consistent differentiator of text grade levels.

Influence of grade band on sentence length. Sentence length analyses showed a main effect for grade (F (4, 298) = 27.04, p<.001, η_p^2=.27), no effect for content (F (4, 298) = 1.75, p=.13, η_p^2=.07), but a significant grade-by-content interaction (F (16, 298) = 2.28, p<.004, η_p^2=

.11). Thus, differences in sentence length depended upon grade *and* content. Because the influence of grade band cannot be discussed irrespective of content, these interactions are detailed in the next section at the end (*Influence of content and grade on sentence length*).

Influence of content on word frequency. Since content area is not accounted for in readability theory, it would be expected that word frequency levels would *not* differ by text content area. However, as the findings showed, there were main effects for content on word frequency. From contents with the least frequent words to most frequent words, content areas were ordered in the following way: a) Literature (M =3.71, SD=.15); b) Math (M =3.61, SD=.18); c) Science (M =3.52, SD=.16) and Social Science (M =3.57, SD=.17); and d) News (M =3.43, SD=.18) (See Figure 2).

Literature texts consistently had the *most* frequent words with significantly higher word frequencies than all other contents—social science, news, science, ($p < .001$) and math ($p < .05$). Math texts were similar with significantly more frequent words than news ($p < .001$) and science texts ($p < .05$) and significantly less frequent words than literature texts.

At the other end of the spectrum with the least frequent words, were news texts which had significantly lower word frequency levels (i.e., harder words) than all other texts including literature, math, social science, ($ps < .001$), and science, ($p < .05$). Social science and science texts had word frequency levels that were not significantly different. However, science texts had significantly less frequent words than literature and math texts ($p < .05$) and more frequent words than news texts ($p < .001$). Social science texts followed the same pattern with significantly less frequent words than literature and math ($p < .001$) and more frequent words than news texts ($p < .001$).

In sum, content area trends in word frequency showed that literature texts uniformly possessed more frequent words than *all* other contents and math texts followed the same pattern with news and science texts. News texts possessed the least frequent words of all, making them distinctly different from math, literature, social science, and science texts. Falling after news texts were science texts, which had less frequent words than many other content areas.

Influence of content on word length. As previously reported there were main effects for text content on word length. Post hoc analyses of word length by content area mirrored patterns in word frequency. From content areas with the shortest words to those with the longest words, the set could be generally ordered in the following way a) literature and math; b) social studies and science; and c) news (but did not have longer words than social science).

Specifically, literature texts had significantly shorter words than social science, news (p's < .001), and science texts (p<.01) but not math texts (p =1.00). Math texts had shorter words than science, social science, and news texts (p's < .01) but not literature texts (p =1.00). Science texts had significantly longer words than both math and literature texts (p<.05) but did not differ significantly from social studies (p =1.00) and news texts (p >.11). Social science texts had significantly longer words than math and literature texts (ps < .001) but not science (p =1.00), or news texts (p =.71). News texts had significantly longer words than math or literature texts (ps < .001) as well as science texts (p<.05) but not social science texts (p =.71).

In sum, word lengths in texts became longer and harder as grades increased. In the content areas, literature and math texts tended to have significantly shorter words than all other contents. Science and social science texts had word lengths that were similar but longer than literature and math texts. News texts tended to have the longest word of all with the exception of social science texts.

Influence of content and grade on sentence length. Due to the interaction, it was not possible to discuss the impact of content on sentence length, irrespective of grade band. As is common with interactions, the results were complex but coalesced around four patterns.

First, none of the math comparisons was significant, meaning that mathematics texts at various grade ranges did not differ significantly on sentence length. Sentence lengths in mathematics texts in grades 1-3, 4-6, 7-9, 10-12, and university levels were the same.

Second, in science texts sentence lengths were significantly different in a grade wise progression, meaning that as grades increased, so also did sentence length. [Note for science comparisons all *p* values were less than .001, except at grade 7-9 vs. 4-6; 7-9 vs. 10-12; 7-9 vs. university, which all had *p* values of less than .05]. The one exception with science texts was that the sentence lengths of university and grades 10-12 texts were not significantly different ($p = .81$).

Third, in social science texts, sentence lengths at all grades increased except grades 4-6 vs. 7-9 ($p = .38$) and 10-12 vs. university which were not different ($p = .78$) [Note. The *p* values for all comparison were at least less than .01 with the exception of 7-9 vs. university ($p < .05$).]

With news texts prior to grades 7-9, sentence lengths increased with significant differences. Thus, sentence lengths in grades 1-3 were shorter than in grades 4-6 and sentence lengths in grades 4-6 were shorter than 7-9 (*ps* <.001). After grades 7-9, sentence lengths in news texts did not differ between grades. That is, sentence lengths in news texts in grades 7-9 and 10-12 were not different ($p = .16$) and news texts in grades 10-12 and university were not different ($p = .25$). Essentially, after grades 7-9, sentence lengths in news texts did not increase by grade.

Lastly, shorter sentence lengths characterized literature texts at the very low end of the developmental spectrum in literature, grades 1-3, but not much afterward. The literature texts in

grades 1-3 had significantly shorter sentences than those in grades 4-6 (p=.002), grades 10-12 (p =.001) and university texts (p< .001). In all other comparisons, literature texts were not significantly different on sentence length (e.g., 4-6 vs 7-9, 7-9 vs. 10-12, 10-12 vs. 4-6).

In sum, according to most readability theory, sentence lengths would be expected to increase by grade regardless of content area. The results here showed that the pattern held true in certain content areas but not in others. For instance, in mathematics texts sentence lengths did not increase across grade. In literature texts, sentence lengths were only different at the very lowest levels (grades 1-3). In science, texts sentence lengths increased by grade level, through grade 10-12. Both social science and news texts followed the overall trend of increasing sentence lengths at increased grades but more broadly. In both social science and news, texts there were several cases in which adjacent grade levels (e.g., 1-3, 4-6) did not differ in sentence length.

Section summary. Grade level influenced word frequency and word length in texts but interacted with content area in the area of sentence length. The clear patterns in terms of the impact of grade level were a) as grade levels increased word lengths increased and b) as grade levels increased, word frequency generally decreased. Word length was the only text feature to differ consistently and increase at each grade band significantly (e.g. 1-3 < 4-6 < 7-9 < 10-12 < university). Specifically, at each progressive grade band, word lengths systematically and significantly increased. Texts at progressing grade levels also tended to have decreasing word frequencies but the pattern was looser. For example, word frequencies tended to differ at distal grade levels (e.g. 1-3 vs. 7-9) but not adjacent ones (e.g. 1-3 vs. 4-6 or 4-6 vs. 7-9). We can say that increases in grade level clearly and consistently corresponded to increases in word length and that grade level increases were generally associated with word frequency decreases as grades progressed.

In terms of content, word frequency and word length patterns tended to be similar to each other, whereas sentence length patterns were highly variable. Literature and math texts, for instance had some of the most frequent and shortest words. In contrast, news texts had some of the longest and least frequent words. Science and social science texts fell in the middle in terms of word length and word frequency, and, in terms of word length, social science texts had word lengths that were similar to news. Due to the grade-by-content interaction in the sentence length comparisons, what can be said is that sentence length was highly variable and dependent about both content and grade. For example, although sentence lengths did not differ in math texts as grades increased, they did in science texts. Whereas literature texts had shorter sentences, the pattern was confined to only materials at grades 1-3. News texts did have sentence lengths that increased by grade, but the pattern did not hold true after grade 7. Sentence length, then, appeared to be a highly dependent variable.

Question 2: Grade Band and Content Moderating the Influence of Word Frequency, Word Length, and Sentence Length on Comprehension

We were interested in whether grade and content moderated the effects of word frequency, word length, and sentence on comprehension. For example, knowing that there is a relationship between word frequency in a passage and a student's comprehension of that passage, we wanted to examine the degree to which the grade level of the text or its content area would change that relationship. In other words, perhaps word frequency would influence comprehension differently at grades 1-3 as compared to grades 7-9 or have more of influence in science texts and less influence in literary texts. Figure 1 illustrates the hypothesized relationship of grade or content on the relationship between word frequency, word length, sentence length and comprehension.

Table 2 shows the mean comprehension scores by grade and content. As is typical in a true

cloze procedure, wherein only precise word replacements are correct, accuracy rates do not surpass 60%. In order to examine if students comprehended different contents at different levels by grade or content, we first compared mean comprehension scores by content and found that there was a main effect for content area (F (4, 318) = 2.53, p=.04). However, post hoc tests on content were not significant. We also compared mean comprehension by grade and found a main effect (F (4, 318) = 79.73, p<.001) with all post hoc tests on grade being significant (ps <.001) except the university vs. 10-12 comparison. Cloze accuracy rates decreased by grade. This trend is common in a true cloze procedure and reflects the fact that highly specific, content-based words would be challenging to replace at higher-grade levels.

In keeping with previous moderator analyses in reading, we ran hierarchical linear regression equations, using each text feature (e.g., word frequency, sentence length, word length) to predict comprehension. Then we constructed dummy variables for either grade or content, and then created interaction terms with grade or content. We entered the independent text feature predictors at Step 1 (i.e., word frequency or word length or sentence length) as well as the dummy variables for our moderators (i.e., grade or content area). At Step 2, we added a product term for our moderators (i.e., text feature x dummy grade/content). Significant product terms at the second step indicated a moderation effect.

Grade band moderator. For the grade band moderator analysis, word frequency significantly predicted comprehension (F (6, 323) =90.59, p <.001) with an R^2 of .58. None of the interaction terms in the word frequency analysis was statistically significant, indicating that grade did *not* moderate the impact of word frequency on comprehension. The relationship between word frequency and comprehension was the same across grade. Similarly, word length significantly predicted comprehension (F (6, 323) =93.86, p <.001) with an R^2 of .60 but none of

the interaction terms were statistically significant indicating that grade did not moderate the impact of word length on comprehension. In both of these equations, grade level did predict increases in word frequency and word length, which in turn consistently predicted comprehension, and the relationship did not differ at various grades.

In contrast, the grade level moderator analysis for sentence length *was* significant. Sentence length significantly predicted mean comprehension (F (6, 323) =78.14, p <.001) with an R^2 of .55, but the addition of the grade moderators improved the prediction significantly (See Table 3). Specifically, at grades 1-3, 4-6, 10-12, and university product terms were significant, meaning that the impact of sentence length on comprehension differed in these grades. At grades 7-9, there were no significant effects. At grades 1-3 and grades 4-6 sentences were shorter and this drove higher levels of comprehension and the reverse was true at grades 10-12 vs. university where sentences were longer and comprehension lower.

Content moderator. Mean comprehension scores by content are shown in Table 2. For the content moderator analysis, word frequency significantly predicted comprehension (F (5, 323) =44.51, p <.001) with an R^2 of .41. None of the moderator interaction terms in the word frequency analysis was statistically significant; indicating that content did not moderate the impact of word frequency on comprehension.

For our second content moderator analysis, we examined word length, finding that it significantly predicted comprehension (F (5, 323) =63.25, p <.001) with an R^2 of .49. The moderation *was* significant for literature (See Table 4). This indicated that the relationship between word length and comprehension was different in literature texts that in other texts and that this distinction differentially influenced comprehension. In terms of comprehension levels at grades 1-3, comprehension was highest in literature than any other content but at grades 10-12

and university it was the lowest. What these results suggest is that the shorter words in literature texts at grades 1-3 may have facilitated easier comprehension but by grades 10-12 and university, they did not.

In the last content moderator analysis, sentence length significantly predicted comprehension (F (5, 317) =32.96, p <.001) with an R^2 of .34. The moderation of content on sentence length *was* significant for all contents, meaning that the relationship between comprehension and sentence length was different for each content—literature, math, science, social science, and news (See Table 5).

Section summary. In sum, the analysis showed that grade band and content did moderate the effects of sentence length and word length on comprehension. In terms of word frequency, there were no moderator effects. The impact of word frequency on comprehension was the same regardless of grade or content.

The relationship between sentence length and comprehension was moderated by both grade band and content. Grade band moderated the effects of sentence length on comprehension in all grades but 7-9. In addition, content area moderated the effects of sentence length on comprehension in *all* contents. Thus, the longer sentences in science, social science, and news texts did influence comprehension, at specific grades. Similarly, the shorter sentence lengths in literature did as well.

Content also moderated the effects of word length on comprehension in literature texts, indicating that shorter words in literature were influencing comprehension. Interestingly, however, comprehension in literature was highest among contents in the grades 1-3, but declined comparatively such that by grades 10-12 and university, comprehension in literary texts was *lowest*, amongst contents.

Discussion

The purpose of this study was to address gaps in knowledge about the measurement of text complexity, especially the ability of current, digital formulas to measure text complexity and predict student-reading performance. The study had several findings. First, was that the three text dimensions used in today's most influential digital formulas—word length, word frequency, and sentence length—were present in different, relative amounts at various grades and contents. Whereas readability theory would suggest a singular approach to estimating text difficulty wherein consistent, linear relationships between text features and grades existed, this pattern did not always hold true. Some content areas possessed more or less frequent words and some grade levels were the same. One size did not necessarily fit all grade levels or contents. Particularly in the areas of word frequency and sentence length, patterns differed widely by content, grade, or the interaction of both.

Second, when the relationship between text features and comprehension—student performance--was examined, grade and content influenced or changed the relationship. Intriguingly, these moderating relationships did not always occur in places where the texts themselves varied. For example, although the word frequencies in *texts* did not always differ at adjacent grades, grade level did not change or moderate the relationship between word frequency and comprehension. In contrast, however, sentence length, did play out quite distinctly in texts and did moderate the relationship between sentence length and comprehension.

Word Frequency: Different Levels in Texts, Consistent Prediction across Grades and Contents.

The word frequency findings in this study produced some very interesting trends. As described in the sections below, there were instances where texts really differed in terms of word

frequency levels and then those in which they did not. Notably, however, the word frequency variable performed consistently in the prediction of comprehension across all grades 1-university. In the sections below, we hypothesize about the word frequency findings turning to relevant research and findings.

Lack of grade band distinction highlights word frequency limitations. In texts, word frequency tended to differentiate broad grade bands (e.g. grades 1-6 vs 7-12) but not adjacent ones (e.g., grades 1-3 & 4-6 or 4-6 vs. 7-9). The expected pattern in readability theory would be that words would become less frequent at each increasing grade band but in this data set, texts in grades 1-6 were essentially the same in terms of word frequency.

Theoretically, the use of word frequency as a text feature to differentiate levels fits within a model of reading development such as Chall's (1983), where readers are expected to learn to recognize harder, longer words of decreasing frequency as they move through the elementary and the secondary grades. In the *Educators Word Frequency Guide* (Zeno, Ivens, Millard, & Duvvuri, 1995), words that are prominent in primary-level texts have a predicted frequency of 340 appearances per million words, while words that are added during the middle grades of elementary school have 37 predicted appearances per million. Why, then, were texts at different grade bands not consistently distinguished by word frequency at the elementary level, especially? We believe that the explanation lies in the use of word frequency as a proxy for word difficulty, the averaging of word frequencies within texts, and repetition of words.

One explanation for the lack of differentiation in word frequency at adjacent grade bands points to the limitations of word frequency as a proxy differentiating complex words. It is possible for two words to be of similar frequency but different in true difficulty. For example, within the *Educators Word Frequency Guide* (Zeno, Ivens, Millard, & Duvvuri, 1995) the words

dunk, dowager, and *Mesolithic* all occur 4 times and have Standard Frequency Indexes (SFIs) of around 30. From a word frequency and readability perspective, these words would all be treated similarly, but they are clearly different in true complexity. Thus, it is possible for words in texts at grade 6 and grade 2 to have similar levels of frequency but different levels of difficulty or complexity.

Another explanation for the lack of differentiation in word frequency in various grade bands may lie with the manner in which word frequency is measured in second-generation readability formulas, specifically Lexile, ATOS, and DRP. In earlier formulas (e.g., Dale & Chall, 1948; Spache, 1953), the word frequency measure assessed the proportion of word/s in texts that were not on a designated word list of frequent and/or familiar words—a dichotomous variable identifying a word as "familiar/frequent" vs. "unfamiliar/infrequent."

In an apparent "improvement," digital or second-generation formulas use averages of word frequency to estimate text complexity—a continuous variable. The collapsing of word frequencies within a text into a *mean* log word frequency is not sensitive to the skewed distribution of words in English (Adams, 2009). In the Zeno et al. (1995) database, a small group of words—107—is predicted to appear 1,000 times or more per million, while approximately 85,000 words (60% of the list) is predicted to appear less than once per 10 million.

In the case of texts where rare words are prominent and repeated frequently, the average word frequency is likely influenced. For example, in a Grade 1-3 in Bormuth (1969) sample, the word *bald*—a word predicted to appear 7 times per million words (Zeno et al., 1995)—was repeated five times in a 100-word segment about the *bald eagle*. The distribution of words in English is extremely skewed, making the mean log word frequency a measure that shows little variation across grade levels.

More frequent words in math and literature, less frequent words in science, social studies, and news. Patterns of word frequency and length showed differences in content areas with math and literature texts having the most frequent and shortest words and news texts having the least frequent and longest words, findings that both challenge (e.g. Gardner, 2004) and replicate (Lee, 2001) previous work. The science and social science findings were not surprising, since these texts tend to have content-specific words that are rare and difficult (e.g., *bicameral, psychogenic, societal pressures*). News texts are likely populated with some of these very same words as well as, rare proper nouns.

Given the differences in technical content, the word frequency similarities between math and literature texts were unexpected, but on closer inspection, there were commonalities in word use in these two contents. Both contained many common, highly frequent words used in unique and sophisticated ways. In math, polysemy is evident where common words take on content-specific, technical meanings (e.g., *line, set, balance*). In recent innovations, polysemy is addressed in a variable called "word maturity," in Pearson tool's RMM, which estimates the number of different meanings of a word that students develop over text exposure (Landauer et al., 2011). The RMM was a better predictor of performance in state and standardized tests than word frequency in other text complexity systems (Nelson et al., 2011) study.

As asserted in Appendix A of the CCSS (NGA & CCSSO, 2010b), literature texts do tend to include more frequent words. In addition, to having high levels of dialogue, which explains this trend, in literature, common words are used in figurative devices such as analogy, personification, and metaphor as when Steinbeck (1952) compares clouds to the "grey of rats" and air as "raw and wounded" in *East of Eden*. Many of the recommendations related to the close reading that has been an emphasis in literature of the Common Core State Standards (NGA &

CCSSO, 2010a, 2010b) focus precisely on the figurative language of literature. It may be the figurative use of language in literature, rather than long or rare words renders texts more difficult. Hence, word frequency may not be the best indicator of word complexity with mathematics and literature texts because both rely on a simple lexicon to convey complex, content-specific meanings.

Text analyses showed that, science and social science texts had less frequent and longer words and that news text had the least frequent words. Science and social science texts are filled with content-specific words, many of which are also multimorphemic (e.g., *macromolecule, bureaucracy, civilization*). Additionally, social science texts include numerous proper names, as do news texts. Tracking lengthy sentences, managing long, multimorphemic words, and coping with infrequent, unfamiliar words is important in these content areas.

Neither grade nor content area moderated the impact of word frequency on comprehension: Examining an irony. Despite the different levels of word frequency in different texts, neither content nor assigned grade band moderated the impact of word frequency on comprehension. In other words, there were differences in texts that did not matter when predicting comprehension. The word frequency metric predicted comprehension in the same way regardless of grade or content. To explain this finding we explored three hypotheses.

The first relates to the traditional cloze procedure, which, as discussed earlier, has the advantage of items that are completely text-dependent and lack test-constructor variance. However, a true cloze requires a precise word replacement, which may be more representative of vocabulary knowledge than proficiency with inferences, main ideas, or details. That being the case we would expect student comprehension as measured by a cloze procedure to be *more*

sensitive to word frequency differences in text, than another measure and yet, this was not the case. The influence of word frequency was not moderated by grade or content differences.

It is possible that one reason that the cloze comprehension was not sensitive to word frequency differences is due to the rigorous way that Bormuth designed the measure, with multiple versions of each passages. This design meant that no one word was replaced more than any other word. For example, in one version of a passage a simple word like "other" might be eliminated but in another version of the same passage a complex word like "electromagnetic spectrum" might be eliminated. Thus, the comprehension scores associated with each of the passages was averaged across versions with different omitted words, so that the deletion of a "hard" would not be too influential.

Word Length: A Reliable Variable Distinguishing Grade Levels and Predicting Comprehension

In a recent comparison of the validity of readability tools, six out of seven current tools included word length (Nelson, et al., 2012). In this study, word length was the most reliable text feature in differentiating grade. Texts at each grade band systematically differed with words becoming increasingly longer at each grade band and the moderator analysis showed word length to predict comprehension similarly across grades. Thus, the performance of word length by grade was solid, reinforcing its recent use by formula developers.

Despite the fact that word lengths were different in many texts (e.g., longer in science, social studies and news, shorter in math & literature), only in literature texts did word length moderate comprehension. The findings suggested that word length, rather than word frequency was impactful in comprehending literature texts. In other words, if there were two words of the same frequency but one was shorter, the shorter word would impact comprehension, but only in

literature. In addition, the results suggested that the comprehension of math texts, which also had shorter and more frequent words, was not influenced by the word frequency or word length factors.

Sentence Length: A Highly Dependent Text Variable

Perhaps more than any other findings this study, those surrounding sentence length produced some of the most interesting trends. Sentence length is included in readability formulas to serve as a proxy for syntactic complexity, because, in general, longer sentences are more complex with more dependent and independent clauses, phrases, modifiers, and conjunctions, and thus, more difficult to comprehend. There is more for the mind to "hold onto" in a longer sentence as the meaning is constructed and worked out. What became clear in this analysis was the highly variable and dependent nature of sentence length both as a differentiator of texts and as a predictor of comprehension. Unlike word frequency, which consistently predicted comprehension regardless of grade or content, the influence of sentence length on comprehension was moderated by both grade and content area.

In terms of texts, themselves, the results suggested that longer sentences characterized social studies and science texts at increasing grades and that, given the significant moderating effects in these contents, these longer sentences also influenced comprehension. We can then expect that, when reading science and social studies, students would be challenged by longer sentences, regardless of grade. Paying attention to sentence-level comprehension would likely be useful in comprehension of science and social studies texts. Similarly, sentences in news texts were also longer progressively by grade but not above the 7-9 grade level. Since grade and content moderated comprehension, we can draw the conclusion that longer sentences rendered news texts more difficult to comprehend, especially at grades 1-3, 4-6, and 7-9. After grade 7,

news texts were similar in terms of sentence lengths, and likely comprehension not affected differently. Instructionally, this suggests that in working in news texts at grades 1-7, teachers might work with students on sentence level comprehension skills, including understanding connectives.

In terms of math and literature texts, sentence lengths did not distinguish texts at different grades much, if at all. A conclusion that might be drawn from these results is that sentence length is not important in understanding math texts and only important at the lower grades in literature. Of course, mathematics educators would likely agree that the dense conceptual knowledge required to understand mathematics texts does not lie within lengthy sentences. With respect to literary texts, our findings support the hypothesis that narrative texts possess shorter sentences due, perhaps to dialogue (O'Shea, Bandar, Crockett, & McLean, 2011).

These results intersect with previous readability results to suggest sentence length is a peculiar differentiator of texts and predictor of comprehension. In this study, it did not prove to be a predictor that performed uniformly across grades and contents. In content-heavy disciplines, such as science and social studies, it was consistently predictive but in literature and math not so much. The results fit a recent study showing that that word-level predictors made a larger contribution than sentence length (Authors, 2018). These findings along with other trends in readability research, suggest that sentence length should be reevaluated as a major predictor in text difficulty.

Implications

The main implication of the study's findings is that one unified measure of text complexity may not be appropriate at all grade levels and across all content areas. Specifically, we noted four additional implications. First, the failure of word frequency to distinguish between

texts at the grade 1-6 range or the grade 7-12 range is a problem, given the importance of vocabulary in comprehension (Ricketts, Nation, & Bishop, 2007; Sénéchal, Ouellette, & Rodney, 2006). When word (vocabulary) demands fail to be influential in formulas, text difficulty labels can be dominated by sentence length. Thus, we suggest alternatives to the current use of an average of word frequency of all words in a text, including the use of designated groups of words that differ in features such as age-of-acquisition and frequency.

Second, the findings suggested that formula results should be interpreted differently based on subject matter and grade. Teachers of literature and mathematics would do better to concentrate on the figurative and/or discipline-specific uses of words rather than frequency. Within social studies or science, however, it is quite likely that word frequency is making a text more difficult and a teacher may want to make sure that students recognize and understand lengthy, content-specific words.

Third, these results suggested the importance of digitally adjusting formulas based on grade and content area. At present one formula, SourceRater, has separate equations for different genres, but no others do. In fact, the digital advantages of today's second-generation formulas are grossly underutilized; in no other era could formulas and their results be more quickly and easily adjusted than today. Adaptive tests are virtually the norm in K-12 education and there is no reason that adaptive readability formulas could not also be. One adjustment, based on the present findings, would be to diminish the influence of word frequency in literature texts. Another adjustment might be to adjust the influence of sentence length in secondary texts, where we found no differences in grades 7-12 and word complexity better predicts reading comprehension (Arya, et al., 2011; Droop & Verhoeven, 1998).

Lastly, the conventional approach in developing formulas is to predict criterion variables that reflect age, grade, or reading experience and to create difficulty labels that reflect the same. This study suggested that formula criterion variables should also include content area. Readability formulas are developed solely on grade-level data without regard for text content, it is likely that critical aspects of text difficulty variation will be ignored.

References

Authors, 2012a

Authors, 2012b

Authors, 2014

Author, 2017

Adams, M. J. (2009). The challenge of advanced texts. In E. H. Hiebert (Ed) *Reading more, reading better: Are American students reading enough of the right stuff?* (163-189) New York: Guilford Publications.

Anderson, R.C., Hiebert, E.H., Scott, J.A., & Wilkinson, I.A.G. (1985). *Becoming a Nation of Readers: The Report of the Commission on Reading.* Champaign, IL: The Center for the Study of Reading, National Institute of Education, National Academy of Education.

Arya, D. J., Hiebert, E. H., & Pearson, P. D. (2011). The effects of syntactic and lexical complexity on the comprehension of elementary science texts. *International Electronic Journal of Elementary Education*, *4*(1), 107.

Best, R. M., Floyd, R. G., & McNamara, D. S. (2008). Differential competencies contributing to children's comprehension of narrative and expository texts. *Reading Psychology*, *29*(2), 137-164.

Biber, D. (1988). *Variation across speech and writing*. Cambridge, MA: Cambridge University Press.

Bormuth, J. R. (1969). *Development of readability analyses* (U.S. Office of Education Final Rep., Proj. No. 70052, Contract No. OEC-3-7-070052-0326). Chicago, IL: University of Chicago.

Bormuth, J. R. (1971). *Development of standards of readability: Toward a rational criterion of passage performance*. (U.S. Office of Education Final Rep., Contract No. OEC-0-9-230237-4125). Chicago, IL: University of Chicago.

Cervetti, G. N., Bravo, M. A., Hiebert, E. H., Pearson, P. D., & Jaynes, C. A. (2009). Text genre and science content: Ease of reading, comprehension, and reader preference. *Reading Psychology*, *30*(6), 487-511.

Chall, J.S. (1983). *Stages of reading development.* New York: McGraw-Hill.

Chall, J. S., Bissex, G. L., Conard, S. S., & Harris-Sharples, S. H. (1996). *Qualitative assessment of text difficulty: A practical guide for teachers and writers*. Brookline, MA: Brookline Books.

Coté, N., Goldman, S. R., & Saul, E. U. (1998). Students making sense of informational text: Relations between processing and representation. *Discourse Processes*, *25*(1), 1-53.

Dale, E. & Chall, J.S. (1948). *A formula for predicting readability.* Columbus, OH: Ohio State University Bureau of Educational Research. (Reprinted from Educational Research Bulletin). 27, 11-20.

Deane, P., Sheehan, K. M., Sabatini, J., Futagi, Y., & Kostin, I. (2006). Differences in text structure and its implications for assessment of struggling readers. *Scientific Studies of Reading*, *10*(3), 257-275.

Droop, M., & Verhoeven, L. (1998). Background knowledge, linguistic complexity, and second-language reading comprehension. *Journal of literacy research*, *30*(2), 253-271

Ehri, L. C. (2005). Learning to read words: Theory, findings, and issues. *Scientific Studies of rea*ding, 9(2), 167-188.

Farr, J. N., Jenkins, J. J., & Paterson, D. G. (1951). Simplification of Flesch Reading Ease Formula. *Journal of Applied Psychology*, *35,* 333-357.

Flesch, R. (1943). Marks of readable style; a study in adult education. *Teachers College Contributions to Education*.

Fry, E. B. (1969). The readability graph validated at primary levels. *The reading teacher*, *22*(6), 534-538.

Gamson, D. A., Lu, X., & Eckert, S. A. (2013). Challenging the research base of the Common Core State Standards: A historical reanalysis of text complexity. *Educational Researcher*, *42*(7), 381-391.

Gardner, D. (2004). Vocabulary input through extensive reading: A comparison of words found in children's narrative and expository reading materials. *Applied Linguistics*, *25*(1), 1-37.

Graesser, A. C., McNamara, D. S., & Kulikowich, J. M. (2011). Coh-Metrix providing multilevel analyses of text characteristics. *Educational researcher*, *40*(5), 223-234.

Gunning, T. G. (2003). The role of readability today's classroom. *Top Language Disorders. 23*(3) 175-189.

Joyce, J. (1939/2015). *Finnegans wake*. London, UK: Penguin.

Karlsson, A. M. (2009). Positioned by reading and writing: Literacy practices, roles, and genres in common occupations. *Written Communication*, *26*(1), 53-76.

Klare, G. R. (1984). Readability. In P.D. Pearson, R. Barr, M. Kamil, & P. Mosenthal (Eds.), *Handbook of reading research* (*Vol. 1, pp. 681-744).* New York, NY: Longman.

Klare, G. R., & Buck, B. (1954). *Know your reader: The scientific approach to readability*. Hermitage.

Kintsch, W. (1988). The role of knowledge in discourse comprehension: A construction-integration model. *Psychological review*, *95*(2), 163.

Landauer, T.K., Kireyev, K., & Panaccione, C. (2011). Word maturity: A new metric for word knowledge. Scientific Studies of Reading, 15(1), 92–108. doi:10.1080/10888438.2011.536130

Lee, D. Y. W. (2001). Genres, registers, text types, domains and styles: Clarifying the concepts and navigating a path through the BNC jungle. *Language Learning and Technology*, *5*(3), 37-72.

Lefstein, A., & Snell, J. (2011). Promises and problems of teaching with popular culture: A linguistic ethnographic analysis of dis-course genre mixing in a literacy lesson. Reading Research Quarterly, 46(1), 40–69. doi:10.1598/RRQ.46.1.3

Lupo, S. (2017). Comprehension, text difficulty, background knowledge, talk: A comparison of KWL and listen read discuss. Unpublished doctoral dissertation, University of Virginia.

McLaughlin, G. H. (1969). SMOG grading-a new readability formula. *Journal of reading*, 12(8), 639-646.

National Governors Association Center for Best Practices & Council of Chief State School Officers. (2010a). *Common Core State Standards for English language arts and literacy in history/social studies, science, and technical subjects.* Washington, DC: Authors. Retrieved from www.corestandards.org/assets/CCSSI_ELA%20Standards.pdf

National Governors Association Center for Best Practices & Council of Chief State School Officers. (2010b). *Common Core State Standards for English language arts & literacy in history/social studies, science, and technical subjects*, Appendix A. Washington, DC: Author. Retrieved from http://www.corestandards.org/assets/Appendix_A.pdf

Nelson, J., Perfetti, C., Liben, D., & Liben, M. (2012). *Measures of text difficulty: Testing their predictive value for grade levels and student performance*. *Council of Chief State School Officers, Washington, DC*.

O'Shea, J., Bandar, Z., Crockett, K., & McLean, D. (2011). A comparative study of two short text semantic similarity measures (pp. 172-181). In J. O'Shea, N.T. Nguyen, K. Crockett, R.J., Howlett, & L.C. Jain (Eds.). *Agent and Multi-Agent Systems: Technologies and Applications*. New York, NY: Springer.

Ozuru, Y., Dempsey, K., Sayroo, J., & McNamara, D. S. (2005). Effects of text cohesion on comprehension of biology texts. In *Proceedings of the 27th Annual Meeting of the Cognitive Science Society* (pp. 1696-1701).

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

Pearson, P. D., & Hiebert, E. H. (2014). The state of the field: Qualitative analyses of text complexity. *The Elementary School Journal*, *115*(2), 161-183.

Ricketts, J., Nation, K., & Bishop, D. V. (2007). Vocabulary is important for some, but not all reading skills. *Scientific Studies of Reading*, *11*(3), 235-257.

Sáenz, L. M., & Fuchs, L. S. (2002). Examining the reading difficulty of secondary students with learning disabilities expository versus narrative text. *Remedial and Special Education*, *23*(1), 31-41.

Saukkonen, P. (2007). Cognitive schemas behind statistics: Towards a system of text typology. *Journal of Quantitative Linguistics*, 14(2/3), 242â€“264. doi:10.1080/09296170701514197

Sénéchal, M., Ouellette, G., & Rodney, D. (2006). The misunderstood giant: On the predictive role of early vocabulary to future reading. *Handbook of early literacy research*, *2*, 173-182.

Shanahan, T., Kamil, M. L., & Tobin, A. W. (1982). Cloze as a measure of intersentential comprehension. *Reading Research Quarterly*, *17* (2), 229-255.

Shanahan, T., & Shanahan, C. (2008). Teaching disciplinary literacy to adolescents: Rethinking content-area literacy. *Harvard Educational Review*, *78*(1), 40-59.

Sheehan, K. M., Kostin, I., & Futagi, Y. (2008). When do standard approaches for measuring vocabulary difficulty, syntactic complexity and referential cohesion yield biased estimates of text difficulty. In *Proceedings of the 30th Annual Conference of the Cognitive Science Society, Washington DC*.

Spache, G. (1953). A new readability formula for primary-grade reading materials. *The Elementary School Journal*, *53*(7), 410-413.

Steinbeck, J. (1952/2002). *East of Eden*. London, UK: Penguin

Thorndike, E. L. (1921). *The teacher's word book.* New York: Teachers College, Columbia University.

Tiegs, E. W., & Clark, W. W. (1963). California Achievement Test. Monterey, CA: CTB.

Tun, P.A. (1989). Age differences in processing expository and narrative text. *Journal of Gerontology*, *44*(1), 9-15.

Zeno, S. M., Ivens, S. H., Millard, R. T., & Duvvuri, R. (1995). *The educator's word frequency guide.* Brewster, NY: Touchstone Applied Science Associates.

Table 1

Means and Standard Deviations for Word Frequency, Sentence Length, and Word Length by Total, Grade, and Content

		Word Frequency (Log)		Sentence Length (Log)		Word Length (Syllables)	
	Content	**Mean**	***SD***	**Mean**	***SD***	**Mean**	***SD***
Grades 1-3	Literature	3.83	.15	2.45	.51	1.23	.05
	Social Studies	3.70	.13	2.46	.39	1.30	.13
	Math	3.68	.18	2.76	.22	1.24	.12
	News	3.51	.12	2.25	.26	1.31	.06
	Science	3.70	.13	2.18	.31	1.25	.09
	Total	3.68	.15	2.40	.39	1.28	.11
Grades 4-6	Literature	3.79	.12	2.98	.23	1.28	.08
	Social Studies	3.61	.11	2.73	.32	1.41	.11
	Math	3.68	.12	2.70	.18	1.26	.08
	News	3.56	.15	2.58	.14	1.38	.09
	Science	3.54	.13	2.59	.30	1.36	.13
	Total	3.61	.14	2.69	.30	1.36	.12
Grades 7-9	Literature	3.75	.11	2.78	.40	1.31	.07
	Social Studies	3.56	.15	2.79	.29	1.46	.13
	Math	3.59	.22	2.75	.25	1.41	.14
	News	3.36	.17	2.92	.20	1.57	.17
	Science	3.55	.11	2.80	.24	1.44	.11
	Total	3.56	.16	2.80	.27	1.45	.13
Grades 10-12	Literature	3.63	.16	3.03	.34	1.48	.10
	Social Studies	3.51	.16	3.02	.27	1.56	.14
	Math	3.54	.17	3.02	.46	1.44	.17
	News	3.40	.10	3.14	.19	1.62	.12
	Science	3.44	.15	3.06	.30	1.51	.08
	Total	3.49	.16	3.04	.30	1.53	.13
University	Literature	3.44	.01	3.31	.02	1.46	.02
	Social Studies	3.29	.15	3.07	.29	1.71	.13
	Math	3.66	.17	2.66	.62	1.36	.15
	News	3.17	.06	3.16	.14	1.70	.06
	Science	3.38	.18	3.04	.22	1.71	.15
	Total	3.37	.20	3.02	.36	1.64	.18

Table 2

Means and Standard Deviations for Comprehension by Grade and Content

Grades	Content	Mean	*SD*
Grades 1-3	Literature	.57	.03
	Social Studies	.49	.09
	Math	.50	.05
	News	.51	.07
	Science	.51	.05
	Total	.50	.09
Grades 4-6	Literature	.42	.10
	Social Studies	.42	.07
	Math	.52	.07
	News	.39	.06
	Science	.42	.08
	Total	.43	.08
Grades 7-9	Literature	.36	.06
	Social Studies	.37	.08
	Math	.44	.11
	News	.33	.06
	Science	.39	.06
	Total	.38	.08
Grades 10-12	Literature	.26	.06
	Social Studies	.30	.07
	Math	.33	.07
	News	.31	.06
	Science	.32	.05
	Total	.31	.06
University	Literature	.19	.04
	Social Studies	.23	.04
	Math	.41	.20
	News	.27	.05
	Science	.24	.03
	Total	.27	.11
Total	Literature	.38	.13
	Social Studies	.39	.11
	Math	.44	.12
	News	.37	.10
	Science	.40	.10
	Total	.39	.11

Table 3

Hierarchical Regression Predicting Comprehension with Sentence Length--Grade Moderator

Step 1 Sentence Length (Grade)	ΔR^2	**SE (*B*)**	**β**	***p***
	.55			***
(Constant)	.61	.13		***
Sentence Length	.01	.00	.29	
Grade 1-3 dummy	.20	.10	.76	
Grade 4-6 dummy	.19	.11	.70	
Grade 10-12 dummy	.07	.14	.24	
Grade University dummy	.55	.14	1.46	***
Step 2 Grade Moderator	.04			***
Grade 1-3 dummy x Sentence Length	-.15	.05	-1.42	**
Grade 4-6 dummy x Sentence Length	-.17	.06	-1.69	**
Grade 7-9 dummy x Sentence length	-.11	.06	-1.25	
Grade 10-12 dummy x Sentence length	-.16	.08	-1.76	*
University dummy x Sentence length	-.33	.07	-2.67	***

=p < .05, ** = *p <.01* *** = *p <.001*

Table 4

Hierarchical Regression Predicting Comprehension with Word Length-Content Moderator

Step 1 Word Length (Content)	**ΔR2**	**B**	**SE (B)**	**β**	**p**
	.49				***
(Constant)		1.04	.06		***
Word Length Text (Syl/word)		-.45	.04	-.71	***
Literature dummy		.32	.16	.87	*
Math dummy		.15	.13	.42	
News dummy		-.05	.10	-.14	
Science dummy		-.10	.09	-.40	
Step 2	.02				*
Lit dummy x word length		-.28	.12	-1.04	*
Math dummy x word length		-.10	.09	-.39	
News dummy x word length		.04	.09	.15	
Science dummy x word length		.07	.06	.40	

* =p < .05, ** = *p <.01* *** = *p <.001*

Table 5

Hierarchical Regression Predicting Comprehension with Sentence Length-Content Moderator

Step 1 Sentence Length (Content)	***ΔR²***	**B**	**SE (B)**	β	***p***
	.34				
(Constant)		1.05	.01		***
Mean Sentence Length		.007	.00	.42	*
Literature dummy		-.02	.12	-.04	
Math dummy		.29	.15	.82	
News dummy		.03	.12	.08	
Social studies dummy		-.00	.08	-.02	
Step 2	.08				
Lit dummy x sentence length		-.28	.06	-2.16	***
Math dummy x sentence length		-.37	.07	-2.90	***
News dummy x sentence length		-.30	.06	-2.29	***
Sci dummy x sentence length		-.30	.05	-3.23	***
SS dummy x sentence length		-.30	.05	-3.49	***

/

** =p < .05, ** = p <.01 *** = p <.001*

Figure 1

Grade and Content Moderate the Effects of Text Variables on Comprehension

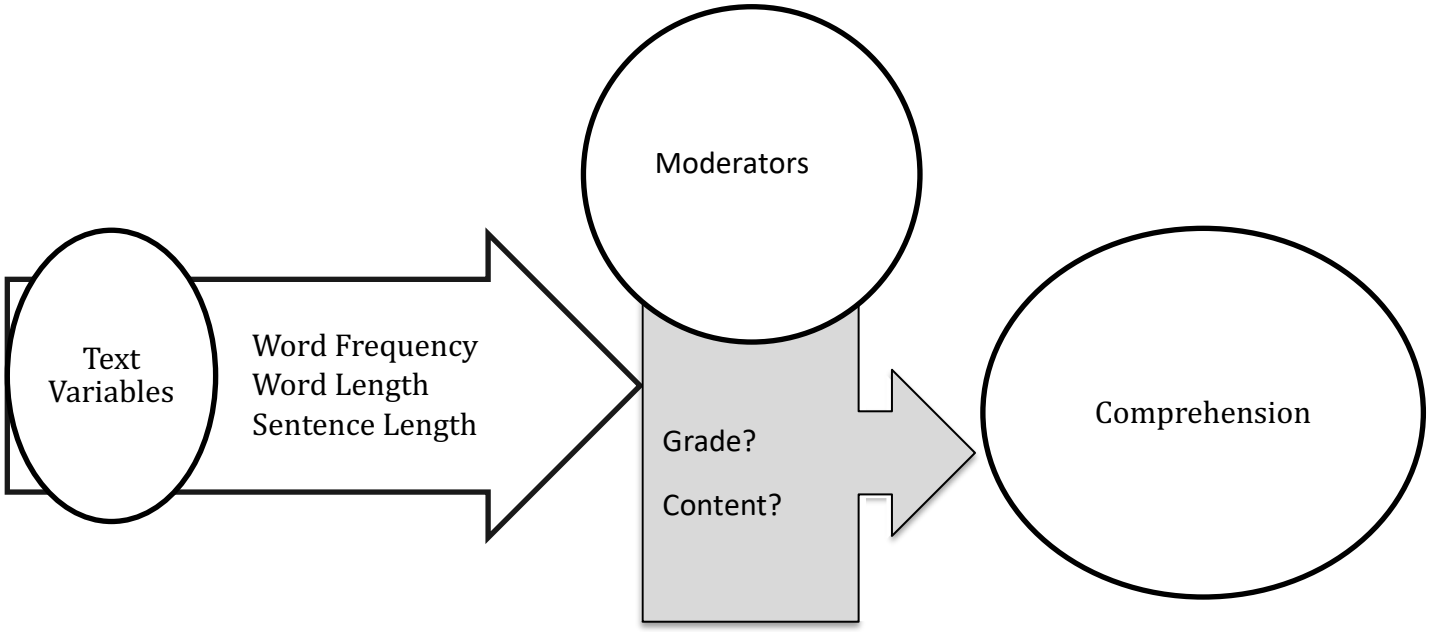
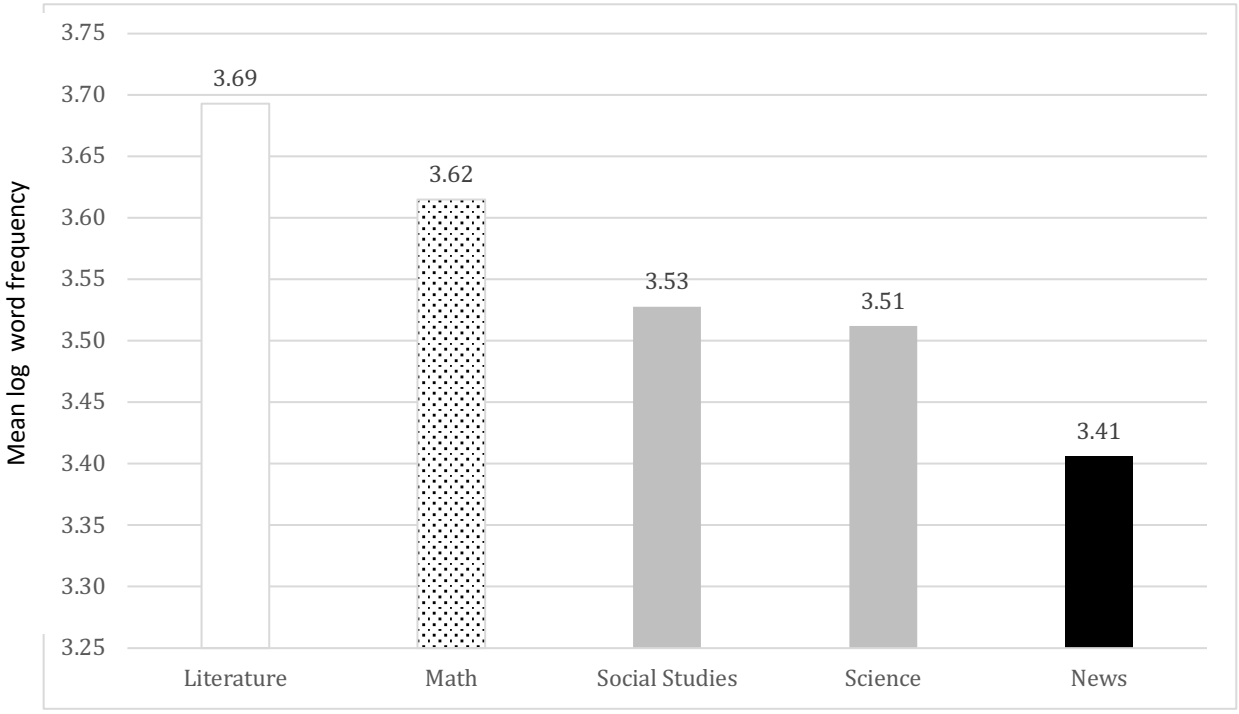


Figure 2

Mean Log Word Frequency by Text Content Area



Notes. Higher bars for mean log word frequencies indicate more frequent/less rare words (e.g., Literature has the most frequent words.)
Content areas bars that are different colors are significantly different at a .05 level (e.g., Literature > Math, Social Studies & Science, News; Math > Social Studies & Science, News; Social Studies & Science > News)