THE ROLE OF WORD-, SENTENCE-, AND TEXT-LEVEL VARIABLES IN PREDICTING GUIDED READING LEVELS OF KINDERGARTEN AND FIRST-GRADE TEXTS

ABSTRACT

Texts classified according to guided reading levels (GRL) are ubiquitous in US beginning reading classrooms. This study examined features of texts across three grade bands (kindergarten, early first grade, final first grade) and the 10 GRLs within these bands. The 510 texts came from three programs with different functions in beginning reading instruction: core, intervention, and content areas. Text features were decoding, semantics, structure, and syntax from the Early Literacy Indicators system, mean sentence length (MSL) and mean log word frequency (MLWF) from the Lexile Framework, and word count. Five variables predicted GRLs of texts: semantics, structure, syntax, MSL, and word count. Differences in decoding and MLWF across grade bands were few and neither variable predicted levels of texts. Intervention texts had lower decoding and MLWF demands than core or content-area texts. Implications of a lack of discernible progressions in decoding and MLWF are discussed.

Elfrieda H. Hiebert
TEXTPROJECT

Laura S. Tortorelli
MICHIGAN STATE UNIVERSITY

Published online April 21, 2022

Elementary School Journal, volume 122, number 4, June 2022.
 https://doi.org/10.1086/719658

CHILDREN living during the seventeenth and eighteenth centuries may have been given adult texts to read (e.g., the Bible), but, since the mid-nineteenth century, beginning readers have been given unique texts. In 1984, Aukerman identified 165 distinct beginning reading programs in the US marketplace, which he collapsed into six main types: phonetic, symbol-sound, whole-word, natural reading, management systems, and total language arts systems. Fifteen years later, when Hiebert (1999) reviewed the prominent texts of beginning reading instruction, many of the innovative efforts of the 1960s and 1970s were no longer prominent and most texts were of three types: decodable, predictable or patterned, and high frequency.

More than 20 years after Hiebert's 1999 review, the most prominent texts used for beginning reading instruction in the United States take the form of guided reading texts or, as they are more commonly called, leveled texts (Conradi Smith et al., 2019; Fitzgerald et al., 2015b). Leveled text programs are numerous, but one common element is the manner in which the complexity of texts is reported. Across numerous publishers, the procedure for establishing the complexity of leveled texts is that of GRLs (Fountas & Pinnell, 1996, 2012). In this system, judges use 10 criteria to place texts into 26 levels ranging from A to Z. Despite the extensive application of this text-leveling system, its theoretical and empirical foundation remains infrequently described and elusively defined.

Recently, a text complexity system for young readers was empirically validated (Fitzgerald et al., 2015a). The comprehension levels of a large sample of first- and second-grade students were established using a sample of texts that included a range of text types, including leveled texts. Analyses considered the relationship between 238 text features and student performance. Four constructs representing nine primary variables contributed to the best predictive model of student performance: decoding, semantics, structure, and syntax.

These four constructs, empirically established, have added considerable resonance to a framework for the selection and creation of beginning reading texts that Mesmer et al. proposed in 2012. Mesmer et al. described the need for beginning reading texts to address variables at three levels if students are to develop the word recognition, comprehension, and motivation of proficient readers: word (e.g., decodability, frequency, and semantics), sentence (e.g., complexity of clauses), and text (e.g., topics and accessible structure). They also described the urgency of attending to program features such as word repetition and the pace of introducing new words.

The corroboration between the variables identified in the empirical work of Fitzgerald et al. (2015a) and the theoretical model of Mesmer et al. (2012) serves as the foundation for the current research. In this study, we examined the text features of three programs of leveled texts that are advertised as serving different functions in beginning reading programs: (*a*) core or Tier 1 in response to intervention (RTI), (*b*) intervention for Tiers 2 and 3 of RTI, and (*c*) science and social studies content. The leveling systems are described as representing the same constructs across programs. Consequently, one assumes that texts at the same levels of different programs have similar features. For the critical period of literacy development that extends over the kindergarten and first-grade years, compatibility in the assignments of text complexity seems essential if beginning readers are to be given appropriate texts.

To provide the context for our analysis, we first review existing systems, both qualitative and quantitative, for determining the complexity of beginning reading texts. We then review existing research on the variables that have been studied in

relation to texts leveled according to GRLs and examine the evidence for learning outcomes with leveled texts.

Research on the Complexity of Beginning Reading Texts

Over the decades when American beginning reading texts were organized by grade levels, readability formulas were the basis for text assignments (Dale & Chall, 1948; Spache, 1953). When *Becoming a Nation of Readers* (Anderson et al., 1985) called for a cessation in the use of readability formulas as the basis for text creation or selection, two ways of measuring the complexity of early texts replaced readability formulas: (*a*) computerized quantitative systems and (*b*) qualitative measures.

Computerized Quantitative Systems of Text Complexity

Second-generation text complexity formulas such as the Lexile Framework (Stenner, 1996) continue to use mean sentence length (MSL) and vocabulary, like earlier formulas (e.g., Dale & Chall, 1948; Spache, 1953). Unlike these earlier formulas, digital processing means that the average frequency of words in texts can be established rapidly by using the rankings of words from large databases. As a result of the repetition of relatively infrequent words (e.g., names of animals such as bonobos) and additive sentences, assigning Lexile levels to beginning reading texts can be perplexing. A text such as *The Cake That Mack Ate* (Robart, 1991), with additive sentences (e.g., "This is the cake that Mack ate. This is the egg that went into the cake that Mack ate." [pp. 2–6]), is assigned a Lexile of 370—close to the beginning of the second- to third-grade band of the Common Core State Standards' (CCSS) staircase of text complexity (National Governors Association Center for Best Practices & Council for Chief State School Officers [National Governors], 2010), whereas the Lexile for a first-grade text from a high-frequency core reading program (Robinson et al., 1962) is –320.

The uniqueness of beginning texts led Fitzgerald et al. (2015a) to consider an alternative scheme for text complexity that went beyond the mean log word frequency (MLWF) and MSL of the Lexile Framework. The Early Literacy Indicators (ELI) scale was based on two analyses organized around a set of 350 texts that represented six early-grade text types: code-based, whole-word, trade books, leveled books, assessments texts, and other (e.g., label books).

The first scale was derived from text complexity levels associated with comprehension scores of 1,258 first and second graders on maze passages from a subset of the 350 texts. A second logit scale was created from 90 primary-grade teachers' attributions of text complexity to random pairs of the 350 texts. These data were used to assign a text complexity level to each of the 350 texts. Next, 22 text variables were represented by 238 computerized variable operationalizations. A regression model established the nine most critical text features associated with text complexity level, which were clustered into four constructs: (*a*) decodability (decoding demands and word length in syllables), (*b*) semantic load (age of acquisition [AoA], word abstractness, and word rareness), (*c*) syntactical complexity (intersentential complexity), and (*d*) discourse structure (phrase diversity, information load or density, and noncompressibility; Fitzgerald et al., 2015a).

These constructs are closely connected to the framework of beginning texts that Mesmer et al. (2012) described as supporting word recognition and comprehension.

The Mesmer et al. framework clusters the two variables in the ELI of decoding and semantics into a single word-level variable. Syntax and discourse are the labels for the other two clusters in the Mesmer et al. framework. Thus, a framework that was based on a review of literature (i.e., Mesmer et al., 2012) and constructs emerging from an empirical evaluation (i.e., Fitzgerald et al., 2015a) provides a foundation for analyzing the features of beginning texts. We used this framework of word-, sentence-, and text-level variables to establish the basis for assignment of text complexity to leveled texts.

Qualitative Systems of Text Complexity

The initial qualitative text-leveling scheme was designed so that American teachers who were implementing the Reading Recovery (RR) tutoring program could use available books in their schools. Peterson (1988) identified four criteria that she used to assign a text to one of the 20 levels that represented grades K through 2: (*a*) book and print features; (*b*) content, themes, and ideas; (*c*) text structure; and (*d*) language and literary elements.

Fountas and Pinnell (1996) applied the leveling system used for RR texts (Peterson, 1988) to classroom texts. Their system, labeled the Fountas and Pinnell or F&P Text Level Gradient or GRLs, involved raters using descriptions of variables to assign a holistic score to a text. Their system employed the four criteria used in leveling RR texts, although themes and ideas were separated from content to create a new criterion. Five additional variables were added: genre, sentence complexity, vocabulary, words (including decoding and frequency features), and illustrations. Unlike the RR system, the GRLs used letters to designate levels and added six levels to extend the system to sixth grade and, more recently, to eighth grade (Fountas & Pinnell, 2012). Fountas and Pinnell describe RR levels as representing a finer gradient than GRLs because of the needs of struggling readers. However, the GRL also has "finer gradations for kindergarten and early first grade, slightly broader categories at later first grade" (Fountas & Pinnell, 1996, p. 115), and increasingly broader categories beyond this level.

No presentations of either RR (Peterson, 1988) or the GRL systems (Fountas & Pinnell, 1996, 2012) have included a review of the literature that verifies the role of specific variables in children's reading acquisition. Furthermore, despite the application of these text-leveling systems to thousands of texts, research has not been conducted on the relative weight of different dimensions in these holistic ratings (Pearson & Hiebert, 2014), nor has data on the reliability among coders in leveling texts been reported in the archival literature.

Research on Features of and Learning with Leveled Texts

The research literature on leveled texts, although not extensive, falls into two groups: (*a*) features that predict levels assigned to texts and (*b*) student performance as a function of different text levels.

Descriptions of Features of Leveled Texts

In a handful of studies, researchers have addressed the question of whether a progression of specific text features is apparent in texts that have been leveled by the RR

system. Hatcher (2000) examined five features of 200 texts (10 at each of the 20 RR levels): number of words with six or more letters, number of words in the longest sentence, total number of words, number of pages, and grammatical forms (contractions, negatives, and auxiliary verbs). The number of words in a book was the variable most strongly correlated with text level ($r = 0.82$). Words with six or more letters and grammatical forms were moderately correlated with text level (0.57 for the former; 0.54 for the latter). This set of variables accounted for 83% of the variance in the RR levels.

Like Hatcher (2000), Pitcher and Fang (2007) considered the number of words, number of high-frequency words, and the match between text level and grade level as determined by readability formulas (Fry and Flesch-Kincaid) in a sample of 80 RR texts, 20 at each of levels 5, 10, 15, and 20. The number of words in the texts predicted the text levels; the number of high-frequency words did not. Readability increased with text level, but readability formulas systematically rated the texts one or more grade levels above the levels indicated by RR.

The most comprehensive study of RR-leveled texts to date was conducted by Cunningham et al. (2005). They examined the predictive strength of 18 variables, including nine word-level variables that addressed word frequency and decodability (e.g., percentage of onset-rime decodable words), four sentence-level measures (e.g., words per sentence and per T-unit), and discourse-level variables (e.g., number of unique and total words). None of the nine word-level variables correlated significantly with text levels. One discourse-level variable (number of unique words) and one sentence-level variable (length of T-units, which consist of a clause and any subordinate clauses) correlated with RR levels at 0.78, accounting for 60% of the variance.

To this point, only two studies have examined features of texts where text complexity has been established with the GRL system. Neither of these studies has described the features that account for assignments of text complexity, but both studies provide some insight into features of GRL-leveled texts. The first, Murray et al. (2014), compared the presence of phonetically regular words and lesson-to-text matches between texts in a leveled and decodable text program. Leveled texts had a relatively high percentage of multisyllabic words at beginning levels (33%) and lower percentages at the final levels (10%). This pattern occurred even when the teachers' guide encouraged instruction of consistent, common letter-sound patterns in monosyllabic words in the early stages. The reverse pattern of multisyllabic words was evident in the decodable texts, where multisyllabic words appeared prominently only after phonics elements in single-syllable words had been emphasized.

In 2017, Koons et al. analyzed 974 texts that represented GRLs from A through M to determine their relationship to Lexiles. The authors reported a high degree of correlation between the assignment of a GRL and the Lexile ($r = 0.84$). The close association between text levels and Lexiles may be explained by the role of syntax in both systems (Cunningham et al., 2018; Pitcher & Fang, 2007).

Student Performance as a Function of Text Level

We located three examinations of how instruction with leveled texts relates to the reading performance of young readers. Hoffman et al. (2001) studied the reading performances of first graders at the end of the school year for 21 texts, three at each of seven GRL levels that fall into three bands: end of kindergarten (Levels C and D),

early first (Levels E, F, and G), and final first (H and I). Each child read a text at each level in each of three conditions: (*a*) sight (no prior support), (*b*) walk-through of title and key vocabulary prior to reading, and (*c*) read-aloud with follow-along prior to reading. When accuracy levels were computed across the three conditions (cold read, prior vocabulary introduction, and prior read-aloud), only high-achieving students achieved the 90% level of accuracy that Clay (1991) deemed as the minimum level in RR on all three grade bands. Middle-performing students had accuracy levels of 90%, 81%, and 79% across the kindergarten, early first, and final first levels, and accuracy levels of low-performing students were 70%, 58%, and 49%. These percentages are far from the 95% or more found necessary to support the comprehension of young adults in English as a second language studies (Hirsh & Nation, 1992).

As implied by the inclusion of "intervention" in its title, the Leveled Literacy Intervention (LLI) program has been positioned in the marketplace as an intervention program (Fountas & Pinnell, 2008). Two studies on the program have been rated as meeting the What Works Clearinghouse criteria for interventions (Ransford-Kaldon et al., 2010, 2013). The complete LLI instructional program comprises numerous activities and materials in addition to texts, but information is not provided in the technical reports on the length of instructional sessions or on the amount of time spent reading in either condition, nor is information provided on training or teacher expertise. Furthermore, unlike the typical mode of establishing text efficacy—in which both conditions receive the same instruction, but the programs of texts are unique (e.g., Jenkins et al., 2004; Juel & Roper/Schneider, 1985)—neither study reports on the effects of leveled texts on student performance relative to other text types.

In the first study, Ransford-Kaldon et al. (2010) reported that kindergartners gained an average of 0.78 GRLs, first graders 1.83 levels, and second graders 1.65 levels after a semester of reading LLI texts. No grade-level group performed higher on the Dynamic Indicators of Basic Early Literacy Skills text reading task. In the second study (Ransford-Kaldon et al., 2013), students showed growth on the LLI benchmark assessment at kindergarten and first grade but not at second grade. On a generalized literacy measure, Standardized Test for the Assessment of Reading (STAR-Reading), students in the treatment groups did not perform any differently from those in the control groups at any grade level.

In summary, even in the face of weak empirical verification of leveled texts, the GRL system is now used to identify instructional-level texts for students in many classrooms (Conradi Smith et al., 2019; Fitzgerald et al., 2015b). Previous studies of the text features of GRL texts did not have the benefit of an empirically validated, comprehensive theoretical model to guide variable selection, and thus we know little about how leveled texts treat the critical aspects of word complexity identified by Fitzgerald et al. (2015a) and Mesmer et al. (2012). We rectify this situation by examining leveled texts from three widely used programs.

The Present Study

This study extends previous research on the RR leveling system (Cunningham et al., 2005; Hatcher, 2000; Pitcher & Fang, 2007) through its examination of three programs of texts leveled according to the GRL system. The analysis of the constituents and progression of leveled texts representing three grade bands (kindergarten, the

first half of grade one [early first], and the last half of grade one [final first]) is also unique to the present study. Finally, this study applies unique variables, including those from an empirically validated framework of text features (Fitzgerald et al., 2015b), the constituents of the Lexile Framework, and word count, which has proven to be a powerful predictor of text complexity assignment in the three previous studies. The questions that the study addressed and the hypotheses related to these questions follow.

Research Question (RQ) 1: How do texts offered for kindergarten, early-first-, and final-first-grade bands differ in word, sentence, and text features? It would be expected that kindergarten texts would demonstrate the lowest levels of word, sentence, and discourse complexity, followed by early first texts, with final first texts demonstrating higher complexity across the board.

RQ 2: How do text levels compare in word-, sentence-, and text-level variables across three published programs, specifically in features of texts at different grade bands? The three published programs should demonstrate consistency in what constitutes a grade band in terms of word, sentence, and discourse features.

RQ 3: Which word-, sentence-, and text-level variables predict GRLs, and how much of the variance in GRLs do they explain? We hypothesize that all word-, sentence-, and discourse-level variables should predict text levels.

Method

The sample comprised 510 leveled texts with an equal number of texts (170) coming from each of three programs. The three programs were chosen because of their distinctive roles in beginning reading instruction and their widespread use (Conradi Smith et al., 2019; Simba Information, 2020). Program 1 (Reading A–Z, n.d.) is an online program that is advertised to "ensure success in your classroom and beyond with engaging, developmentally appropriate leveled books . . . leveled books support instruction in comprehension, vocabulary, close reading of text, and more" (Reading A–Z, 2020, para. 1). The website for Program 1 describes the leveling process as based on the text complexity standards of the CCSS (National Governors, 2010) and uses the following quantitative factors to assign texts to levels: total word count, number of different words, ratio of different words to total words, number of high-frequency words, ratio of high-frequency words to total words, number of low-frequency words, ratio of low-frequency words to total words, sentence length, and sentence complexity. Several qualitative factors, including predictability, text structure and organization, and concept load, are described as influencing the assignment of text complexity as well. This publisher does report that their levels mirror those of the GRL system (Reading A–Z, 2021) but provides no information as to how variables are expressed across levels.

Program 2, the LLI program (Fountas & Pinnell, 2008), is described as "an intensive, small-group, supplementary literacy intervention for students who find reading and writing difficult" (Fountas & Pinnell, 2021b, para. 1). Texts in Program 2 are written and evaluated, according to Fountas and Pinnell (2021a), by experienced educators to conform with the Fountas and Pinnell Text Level Gradient, which includes a qualitative assessment of the 10 factors described earlier in the review of literature.

The nature of description for the 10 constructs is illustrated with explanations for the categories of words and of vocabulary:

> Words are the groups of letters arranged in print that readers must recognize and solve. The challenge in a text partly depends on the number and difficulty of the words that the reader must solve by recognizing them or decoding them. A text that contains a great many of the same common words makes a text more accessible to readers.
>
> Vocabulary refers to the meaning of words and is part of our oral language. The more the words are accessible to readers in terms of meaning, the easier a text will be. An individual's reading and writing vocabularies are words that they understand and can also read or write. (Fountas & Pinnell, 2021a, para. 7–8)

As these explanations illustrate, no indication is given as to what characterizes difficult words to decode or how the variables progress from early to later levels.

Program 3, Windows on Literacy (National Geographic Learning/Cengage, 2001), provides informational text, primarily for science and social studies, drawing on images from *National Geographic*. The publisher describes the program as consisting of leveled texts and states, "Titles in each developmental stage of reading are carefully crafted to provide the supports and challenges that young readers need to build a strong foundation for literacy success" (National Geographic Learning/Cengage, 2021b, para. 1). Although descriptions of the specific leveling criteria and process are not provided, the books for beginning readers are described as having strong word-to-picture matches; predictable, repetitive sentences; familiar concepts; and natural language patterns, whereas the more difficult texts are described as having longer and more complex sentences, more variety in sentence patterns, more content-area vocabulary, fewer pictures that match text, and more structures characteristic of informational text. As is the case with Program 1, Program 3 provides a correlation chart to show how their levels match those of the GRL framework (National Geographic Learning/Cengage, 2021a).

For each program, 170 texts were evenly distributed across the first 10 GRLs, A through J. Fountas and Pinnell (2012) described A to D as covering kindergarten and E to J as covering first grade. When the number of available texts at a level of any program exceeded 17, the sample of texts was selected randomly. Because considerable changes typically occur in students' reading acquisition during first grade, the first-grade texts were divided into two groups: early first (Levels E to G) and final first (Levels H to J). The entire set of 170 texts from Programs 1 and 2 was evenly divided between fiction and nonfiction. All the texts in Program 3 are described as nonfiction or informational. Although we describe the programs here to give a sense of the database as a whole, the goal of this article is to describe how leveled text programs operationalize text complexity, rather than to critique or recommend particular programs. Consequently, we refer to Programs 1, 2, and 3 in our analyses.

Measures

Each of the 510 texts was individually run through the professional version of the Lexile Analyzer (https://la-tools.lexile.com/pro-analyze/). This version of the Lexile Analyzer provides scores for the four ELI measures—decoding, semantics, structure,

and syntax—based on the study by Fitzgerald et al. (2015a; described above). The two constituents of a Lexile measure were included: MSL and MLWF. The final measure was the number of words in the text.

Measures of the ELI system. Each of the four measures of the ELI system—decoding, semantics, structure, and syntax—is given a score on a similar scale: 1 (very low), 2 (low), 3 (medium), 4 (high), or 5 (very high). These scores represent the relative difficulty of a text in relation to the texts in the Fitzgerald et al. (2015a) study. Each of the measures represents one or more constructs that were identified in the Fitzgerald et al. study. Composite constructs will be described for each of the four measures. Further details on the procedures whereby scores were computed can be found in Fitzgerald et al. (2015a).

Two texts are used to illustrate the contributors to the ELI indicators: *Brown Bear* (Martin, 1967) and *Frog and Toad* (Lobel, 1972). According to the GRL system, the former text is assigned to Level D (end of kindergarten) and the latter to Level K (beginning of second grade). A similar number of words—69 for the former and 68 for the latter—has been chosen from each text to illustrate text features. Sentences have been numbered to clarify the computation of the syntax measure.

Example 1: *Brown Bear*: "1. Brown bear, brown bear, what do you see? 2. I see a redbird looking at me. 3. Redbird, redbird, what do you see? 4. I see a yellow duck looking at me. 5. Yellow duck, yellow duck, what do you see? 6. I see a blue horse looking at me. 7. Blue horse, blue horse, what do you see? 8. I see a green frog looking at me. 9. Green frog, green frog, what do you see?" (Martin, 1967, pp. 2–5).

Example 2: *Frog and Toad*: "1. One morning Toad sat in bed. 2. I have many things to do, he said. 3. I will write them all down on a list so that I can remember them. 4. Toad wrote on a piece of paper: A list of things to do today. 5. Then he wrote: Wake up. 6. I have done that, said Toad, and he crossed out: Wake up. 7. Then Toad wrote other things on the paper." (Lobel, 1972, pp. 2–4).

Decoding indicator. The two components of decoding are the complexity of vowel patterns within monosyllabic words and the number of syllables. Vowel patterns in monosyllabic words (except for the 50 most frequent words) are derived from a modified version of Menon and Hiebert's (2005) decodability scale in which simple, long vowel words (e.g., "go") are assigned 1 and multisyllabic words are assigned 9 (e.g., "remember"). Ratings from 2 to 8 are given to progressively more complex vowel patterns and the presence of consonant digraphs and diphthongs at the beginning and end of words. The MRC Psycholinguistic Database (Coltheart, 1981) was used to establish the numbers of syllables in words.

The average monosyllabic word in both illustrative texts has a long vowel pattern (e.g., "see" and "toad"). *Brown Bear* has a slightly larger percentage of multisyllabic words than *Frog and Toad*—21% versus 16%—but *Frog and Toad* has more three-syllable words (e.g., "remember") than *Brown Bear*. The exact weights of these two variables in the computation of the decoding score in the ELI system are proprietorial (see Fitzgerald et al., 2015a), but the scores for decoding are the same: 2 (low).

Semantic indicator. The semantic indicator represents AoA (Kuperman et al., 2012), concreteness and abstractness of words (Brysbaert et al., 2014), and the proportion of rare words in a text. *Brown Bear* has numerous concrete words that appear early in children's oral language, such as *duck* and *bear*, both of which have an AoA of 3.5 and a concreteness score of 4.9 (on a 5-point scale). Prominent words in *Frog*

and Toad such as "list" and "remember" appear later in children's oral language—5.6—and are relatively abstract (2.4 and 2.8, respectively). *Frog and Toad* is rated as somewhat harder on the semantic indicator than *Brown Bear*: 3 (medium) for the former and 1 (very low) for the latter.

Structure indicator. The three constructs of this indicator—text density, phrase diversity, and noncompressibility—all address the repetition of units (words, phrases, and letters) across the entire text. *Brown Bear* has 18 unique words, 10 of which appear in two repeated phrases throughout the text (e.g., "What do you see?"). The high level of repetition of phrases and words across the text account for a structure rating for *Brown Bear* of 1 (very low). By contrast, the 68 words in *Frog and Toad* can be compressed into 40 words where only a few words appear in phrases (e.g., "wake up," "Toad wrote," and "things to do"). On the structure indicator, *Frog and Toad* is evaluated to be considerably harder than *Brown Bear*: 4 (high).

Syntax indicator. This measure denotes intersentential complexity, which refers to the repetition of words between adjacent sentences. A higher rating indicates less repetition between pairs of sentences. In a typical set of adjacent sentences in *Brown Bear*, such as sentences 4 and 5 (see Example 1 above), three words appear in both sentences ("yellow," "duck," and "see"). In the next set of sentences (5 and 6), a single word overlaps ("see"). The syntax rating for *Brown Bear* is 3 (medium). With only a few cases of repetition of words, such as sentences 5 and 6 in Example 2, *Frog and Toad* has a rating of 4 (high) on this measure.

Measures of the Conventional Lexile Analyzer. For comparative purposes, the two measures that contribute to establishing the Lexiles of texts were included in the analysis: MLWF and MSL.

Mean log word frequency. MLWF is a logarithm of the average frequency for every word based on a word's ranking in the MetaMetrics word bank. Higher MLWFs mean that texts have, on average, more frequent words than texts with lower MLWFs and indicate easier text. The MLWF for *Brown Bear* is 3.75 and for *Frog and Toad*, 3.67. *Brown Bear* is predicted to have words with an approximate average frequency of 6,000 appearances (per five million words), whereas the approximate average appearance of words in *Frog and Toad* is 4,250 appearances (per five million words).

Mean sentence length. This measure is the average number of words per sentence. *Brown Bear* has an MSL of 8.35 words, whereas *Frog and Toad* has an MSL of 7.49.

Word count. Word count is the total number of words in the text. *Brown Bear* has 192 words, and *Frog and Toad* has 483 words.

Analytic Plan

To answer RQ 1, we analyzed descriptive statistics for MLWF, MSL, word count, and the measures of decoding, semantic load, syntax, and structure of the ELI system for texts at each GRL to observe patterns of complexity across the K–1 grades. Correlations were used to describe relationships among the seven text variables and GRLs.

To answer RQ 2, we used a series of nine multivariate analyses of variance (MANOVAs) to examine the differences across programs in variables at the word, sentence, and text levels in each grade band (kindergarten, early first, and final first). We chose MANOVAs to examine each subsystem of variables representing word-,

sentence-, and text-level complexity because the goal of the analyses was to identify any differences across programs in how each grade band was defined in terms of our theoretical model (Huberty & Morris, 1992).

To answer RQ 3, we used regression analysis to create and evaluate a model using text variables to predict numeric values corresponding to the GRLs (A = 1, J = 10). To evaluate the applicability of the model to leveled texts outside the current sample, the model was cross validated by randomly dividing the sample into a calibration sample and a validation sample, fitting the model on the calibration sample, and evaluating it in the validation sample (Browne, 2000). To expand on previous studies of leveled texts, we examined the unique contributions of each text predictor to GRLs through dominance analysis (Azen & Budescu, 2003; Budescu, 1993; Hayes & Darlington, 2017). Dominance analysis compares the contribution of each variable to the explanatory power of the full regression model with the contributions of each other variable across all possible models with all possible subsets of variables. Although previous studies of leveled texts have found relationships between a handful of text predictors and text levels (e.g., Cunningham et al., 2005; Hatcher, 2000; Pitcher & Fang, 2007), no studies have determined the relative importance of these predictors (Pearson & Hiebert, 2014). Therefore, a goal of the study was to rank text predictors of GRLs by their importance to better understand how the GRL framework aligns with theoretical and empirical research on text complexity for beginning readers.

To determine dominance, we compared each variable's additional contributions to R^2 across all possible models. For example, we evaluated the R^2 for a model including semantic load as the only text predictor and for a model that included both semantic load and syntax (yielding the unique contribution of semantic load after accounting for syntax), then repeated the procedure pairing semantic load with each other variable, each other pair of variables, and so forth. This procedure was repeated over all possible subsets of the text predictors to determine the average contribution to R^2.

Azen and Budescu (2003) describe three types of possible dominance. *Complete dominance* means that one variable is a stronger predictor than another in every model analyzed. *Conditional dominance* means that one variable has a larger average contribution to R^2 than another across models with the same number of predictors for each subset. *General dominance* means that one variable contributes more than another to the R^2 on average across all models and subset sizes. Complete dominance is inclusive of conditional and general dominance, and conditional dominance is inclusive of general dominance. We obtained reproducibility coefficients indicating the proportion of 1,000 bootstrapped samples that confirmed the levels of dominance found in the analysis. The analyses were completed in *R* using the dominance analysis package. For more detailed information on dominance analysis procedures, see Azen and Budescu (2003).

Results

RQ 1: How Do Texts Offered for Kindergarten, Early-First-, and Final-First-Grade Bands Differ in Word, Sentence, and Text Features?

Descriptive statistics for the variables of all GRLs are provided in Table 1. As demonstrated in Figure 1, a linear increase was apparent for all measures except MLWF

Table 1. Means and Standard Deviations of Seven Measures of Text Difficulty by Level (A to J) and Grade Band (Kindergarten, Early First, Final First)

	Lexile Components		Lexile Early Literacy Indicators[a]				Other
Guided Reading Level	MLWF	MSL	Decoding	Semantic	Syntax	Structure	Word Count
			Kindergarten				
A	3.58 (.51)	4.32 (.99)	2.47 (1.42)	1.77 (1.07)	1.06 (.24)	1.26 (.74)	34.82 (13.87)
B	3.65 (.35)	5.23 (1.43)	2.59 (1.45)	1.90 (1.19)	1.18 (.44)	1.12 (.38)	48.90 (18.11)
C	3.61 (.33)	5.15 (1.44)	2.77 (1.48)	1.88 (.97)	1.45 (.61)	1.28 (.50)	64.10 (22.18)
D	3.61 (.30)	5.61 (.98)	2.35 (1.16)	1.98 (1.10)	1.78 (.70)	1.63 (.75)	86.28 (38.06)
Mean	3.62 (.38)	5.08 (1.31)	2.54 (1.38)	1.88 (1.08)	1.37 (.59)	1.32 (.64)	58.53 (31.17)
			Early First				
E	3.52 (.28)	5.71 (1.17)	2.49 (1.12)	2.19 (1.10)	2.12 (.71)	2.26 (1.02)	117.22 (58.57)
F	3.61 (.22)	6.31 (1.01)	2.67 (1.14)	2.35 (1.00)	2.37 (.56)	2.43 (.83)	138.12 (56.21)
G	3.61 (.24)	6.80 (1.29)	2.84 (.95)	2.55 (.94)	2.73 (.67)	2.88 (.84)	164.20 (70.77)
Mean	3.58 (.25)	6.27 (1.24)	2.67 (1.08)	2.37 (1.02)	2.41 (.70)	2.52 (.93)	139.84 (64.71)
			Final First				
H	3.60 (.15)	7.05 (1.21)	2.90 (1.04)	2.78 (.92)	2.96 (.72)	3.10 (.92)	198.00 (73.35)
I	3.60 (.19)	7.10 (1.23)	3.18 (.82)	3.10 (.78)	3.12 (.74)	3.41 (.90)	254.14 (106.86)
J	3.61 (.16)	7.66 (1.53)	3.22 (.90)	3.51 (1.14)	3.37 (.72)	3.61 (.92)	315.18 (124.65)
Mean	3.60 (.17)	7.27 (1.35)	3.10 (.93)	3.14 (.92)	3.12 (.74)	3.37 (.93)	255.77 (113.76)

Note.—The sample consisted of 50 texts for each of Levels A through I and 46 texts for Level J. A similar distribution at each level of texts came from each text program. MLWF = mean log word frequency; MSL = mean sentence length.

[a] The decoding, semantic, syntactic, and text structure measures come from the Lexile Analyzer revised for beginning reading texts (Fitzgerald et al., 2015a). For each measure, 1 on the scale means "few demands" and 5 means "demanding."

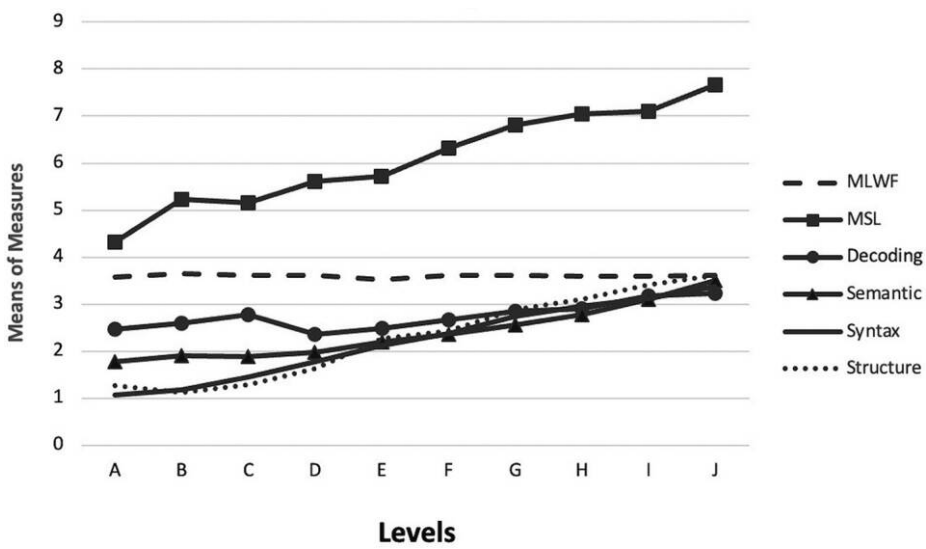

Figure 1. Progression of six variables from Level A to J in three programs. MLWF = mean log word frequency; MSL = mean sentence length.

and decoding, which remained relatively flat across the 10 levels. Level J texts, representing the end of first grade, had an MLWF of 3.61, which indicates more frequent words (on average) than the metric of 3.58 for the lowest kindergarten level, Level A. The mean for the decoding variable remained fairly consistent across all 10 levels, beginning at the midpoint of the scale (2.47) for Level A and increasing by less than one point (3.22) by Level J, the last level.

Data in Table 2 show that the GRL assigned to texts was significantly correlated with all text variables except MLWF. As can be seen in Table 2, the correlations varied from small (decoding and semantics) to moderate (syntax, word count, text structure, and MSL; Ferguson, 2016). These results indicate that lower text levels were characterized primarily by shorter texts and simpler text structures, shorter and simpler sentences, and simpler vocabulary. In addition, lower-level texts had somewhat more decodable words than higher-level texts but not more high-frequency words.

RQ 2: How Do Text Levels Compare in Word-, Sentence-, and Text-Level Variables across Three Published Programs, Specifically in Features of Texts at Different Grade Bands?

MANOVA analyses were used to evaluate differences in word-, sentence-, and text-level variables across the three published programs at each grade band (Table 3). Contrary to predictions, text features showed statistically significant differences across programs at the word and discourse levels in all three grade bands and at the sentence level at the end of first grade.

At the word level, there were significant differences across programs ($p < .001$) in the decoding and semantic indicators across all three grade bands and in MLWF for the kindergarten and early-first-grade texts. Effect sizes for the word-level variables, as indicated by partial eta squared, were small to medium (Ferguson, 2016). Bonferroni post hoc tests indicated that Program 2 had significantly lower decoding and semantic demands than the other two programs in kindergarten and lower decoding demands at the end of first grade. Program 1 had significantly fewer high-frequency words at the beginning of first grade and greater semantic demands throughout first grade compared with the other two programs.

At the sentence level, the programs differed significantly in syntax demands at the kindergarten level and syntax and sentence length at the end of first grade, with small

Table 2. Correlations between Guided Reading Levels (GRLs) and Text Variables

Variables	1	2	3	4	5	6	7	8
1. GRL	1	−.01	.62**	.20**	.47**	.78**	.73**	.77**
2. MLWF		1	.21**	−.17**	−.33**	.03	−.03	.09
3. MSL			1	.24**	.34**	.71**	.45**	.48**
4. Decoding				1	.42**	.23**	.06	.09
5. Semantic					1	.44**	.32**	.35**
6. Syntax						1	.68**	.68**
7. Structure							1	.75**
8. Word count								1

Note.—MLWF = mean log word frequency; MSL = mean sentence length.

** $p < .01$.

Table 3. Summary of Multivariate Analysis of Variance Results

Variable	Program 1 *M* (*SD*)	Program 2 *M* (*SD*)	Program 3 *M* (*SD*)	*F*	*p*	η^2/η_p^2
			Word Level—Kindergarten			
		Wilk's $\Lambda = .71$, $F(6, 398) = 12.15$, $p < .001$				.15
n	68	68	68			
MLWF	3.47 (.38)	3.77 (.30)	3.61 (.40)	12.06	<.001	.11
Decoding	2.75 (1.33)	1.84[a] (.99)	3.04 (1.49)	16.23	<.001	.14
Semantics	2.34 (1.13)	1.21[a] (.53)	2.10 (1.12)	25.87	<.001	.21
			Word Level—Early First			
		Wilk's $\Lambda = .61$, $F(6,296) = 14.01$, $p < .001$				.22
n	51	51	51			
MLWF	3.46[a] (.23)	3.66 (.17)	3.62 (.29)	10.29	<.001	.12
Decoding	3.24 (1.07)	2.14 (.80)	2.63 (1.06)	15.95	<.001	.18
Semantics	3.20[a] (.92)	1.77 (.65)	2.14 (.87)	41.66	<.001	.36
			Word Level—Final First			
		Wilk's $\Lambda = .64$, $F(6,296) = 12.17$, $p < .001$				.20
n	51					
MLWF	3.55 (.16)	3.65 (.15)	3.62 (.18)	4.35	.015	.06
Decoding	3.57 (.76)	2.43[a] (.73)	3.35 (.87)	27.63	<.001	.27
Semantics	3.67[a] (.86)	2.67 (.77)	3.10 (.85)	18.65	<.001	.20
			Sentence Level—Kindergarten			
		Wilk's $\Lambda = .93$, $F(4,400) = 3.58$, $p = .007$				.04
n	68	68	68			
MSL	4.94 (1.21)	5.20 (1.06)	5.09 (1.60)	.68	.507	.01
Syntax	1.21 (.44)	1.57 (.68)	1.32 (.58)	7.23	.001	.07
			Sentence Level—Early First			
		Wilk's $\Lambda = .89$, $F(4,298) = 4.52$, $p < .001$				.06
n	51	51	51			
MSL	6.59 (1.41)	5.94 (.83)	6.29 (1.33)	3.60	.030	.05
Syntax	2.57 (.76)	2.43 (.54)	2.21 (.73)	3.48	.033	.04
			Sentence Level—Final First			
		Wilk's $\Lambda = .75$, $F(4,298) = 11.64$, $p < .001$				.14
n	51	51	51			
MSL	7.91 (1.43)	6.63 (.90)	7.26 (1.37)	13.36	<.001	.15
Syntax	3.61[a] (.72)	2.92 (.48)	2.92 (.77)	17.79	<.001	.19
			Text Level—Kindergarten			
		Wilk's $\Lambda = .73$, $F(4,400) = 16.87$, $p < .001$				.14
n	68	68	68			
Structure	1.22 (.54)	1.28 (.54)	1.46 (.78)	2.55	.08	.03
Word count	50.27 (17.88)	79.57[a] (37.22)	45.74 (23.55)	30.46	<.001	.23
			Text Level—Early First			
		Wilk's $\Lambda = .41$, $F(4,298) = 41.18$, $p < .001$				.36
n	51	51	51			
Structure	2.37 (.94)	3.08[a] (.72)	2.12 (.86)	101.44	<.001	.58
Word count	135.43[a] (33.45)	201.82[a] (54.73)	82.28[a] (36.02)	17.72	<.001	.19

Table 3. (*Continued*)

Variable	Program 1 *M* (*SD*)	Program 2 *M* (*SD*)	Program 3 *M* (*SD*)	*F*	*p*	η^2/η_p^2
	Text Level—Final First					
	Wilk's $\Lambda = .47$, $F(4,298) = 34.31$, $p < .001$					.32
N	51	51	51			
Structure	3.92 (.52)	3.69 (.71)	2.51[a] (.83)	59.62	<.001	.44
Word count	300.18 (93.99)	312.28 (104.66)	154.86[a] (62.46)	49.56	<.001	.40

Note.—MLWF = mean log word frequency; MSL = mean sentence length.

[a] Significantly different from the other two programs, $p < .005$ in Bonferroni post hoc tests. Alpha levels were reduced to account for the multiple analyses conducted, so results with $p < .005$ were considered significant.

to medium effects. Post hoc analyses indicated that Program 1 had higher syntax demands than the other two programs at the end of first grade. At the text level, the programs differed significantly in structure demands in first grade and in word count across the grade bands, with small to medium effects. Program 2 had significantly greater structure demands than the other two programs at the beginning of first grade, and Program 3 had significantly lower structure demands than the other two programs at the end of first grade. Program 2 texts were significantly longer than the other two programs in kindergarten and early first grade, and Program 3 texts were significantly shorter than the other two programs in first grade.

Box-and-whisker plots (Fig. 2) show the variability of text complexity within grade bands and across programs. In each plot in Figure 2, a text feature is depicted side by side for each grade band of the three programs. The boxes represent the middle two quartiles of the texts, and the whiskers represent the top and bottom quartiles. The black bar indicates the average for a group of texts, and outliers are represented as individual points. The heights of the boxes, whiskers, and outlier points in each column represent the variability of each text feature within a grade band and program. The degree to which boxes and whiskers progress from the left panel to the right panel indicates how the texts increase in complexity from kindergarten to the end of first grade for each text feature. The degree to which they align within each panel indicates how well the three programs agree on the grade-level range for each text feature.

Variability within grade bands. Overall, the measures of word complexity demonstrated a high degree of variability. MLWF scores (Fig. 2*a*) had an extensive range in kindergarten (1.6–4.4, including outliers), meaning that kindergarten texts had more high- and low-frequency words than first-grade texts, which ranged from 2.7 to 4.1. For decoding and semantics (Fig. 2*b* and 2*c*), every grade band included texts at all five levels of complexity (although programs showed differences, as discussed below).

On the MSL measure (Fig. 2*d*), kindergarten texts showed a larger range than first-grade texts (1.5–10 words per sentence vs. 4–12 words per sentence). That is, although first-grade texts typically had longer sentences, kindergarten texts did not necessarily have short sentences. Scores for the syntax measure of the ELI system were closely aligned by grade band and showed a clear progression, ranging from 1 to 3 in kindergarten, 1 to 4 in early first grade, and 2–5 in final first grade (Fig. 2*e*).

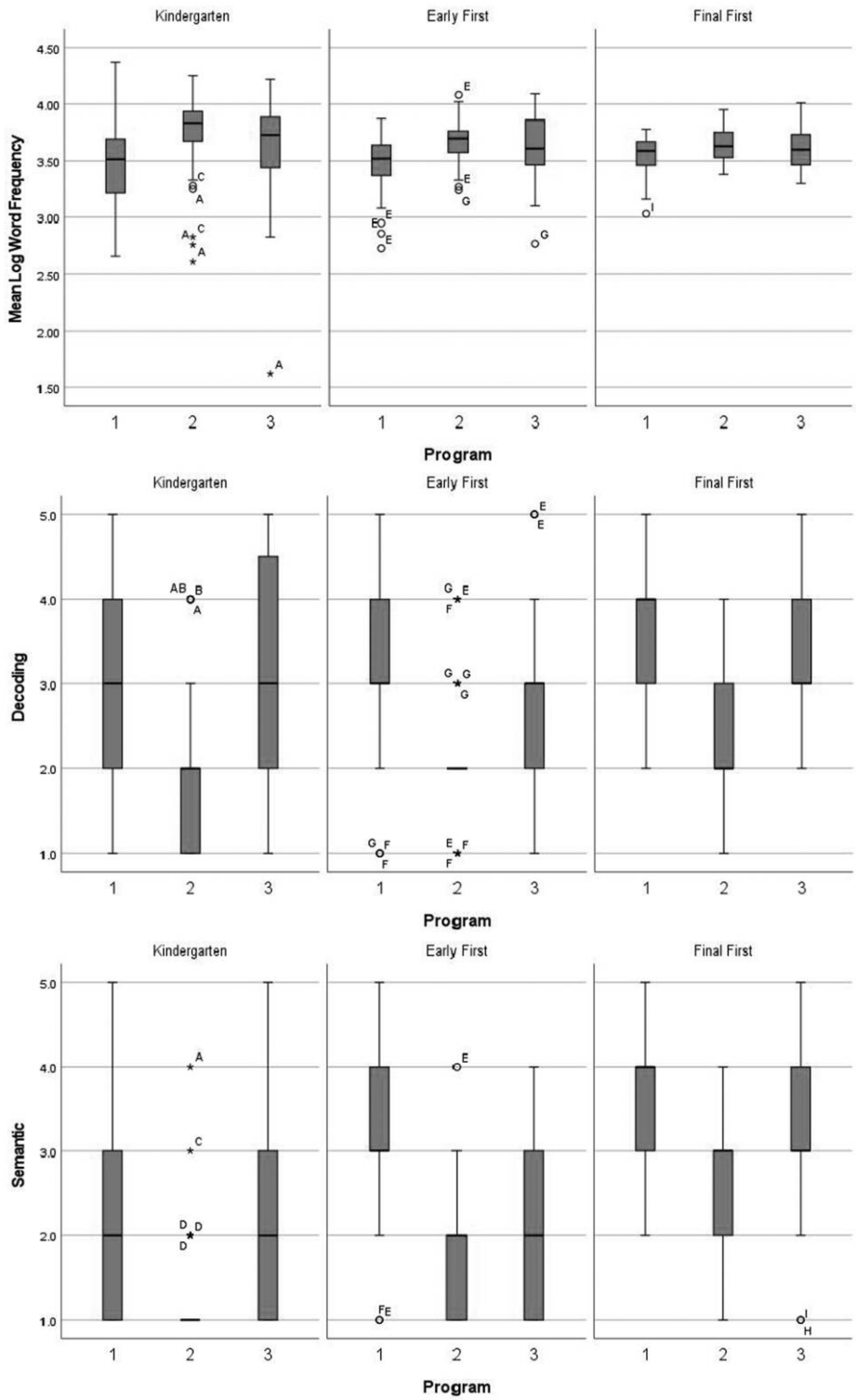

Figure 2. Box plots for all variables by grade bands and programs. Word-level variables: (*a*) mean log word frequency, (*b*) decoding early literacy indicator, and (*c*) semantic early literacy indicator. Sentence-level variables: (*d*) mean sentence length and (*e*) syntax early literacy indicator. Text-level variables: (*f*) structure early literacy indicator and (*g*) word count.

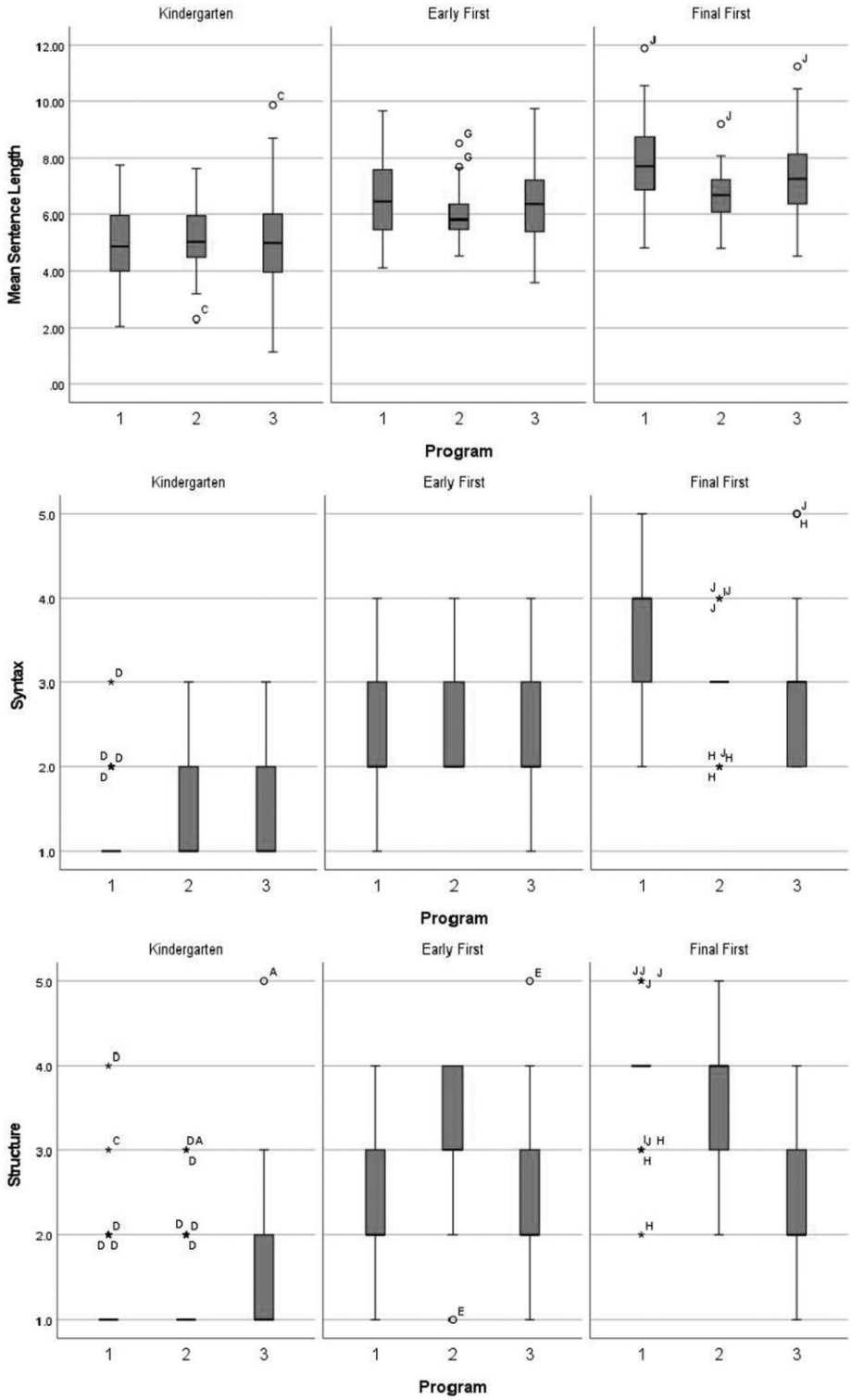

Figure 2. (*Continued*)

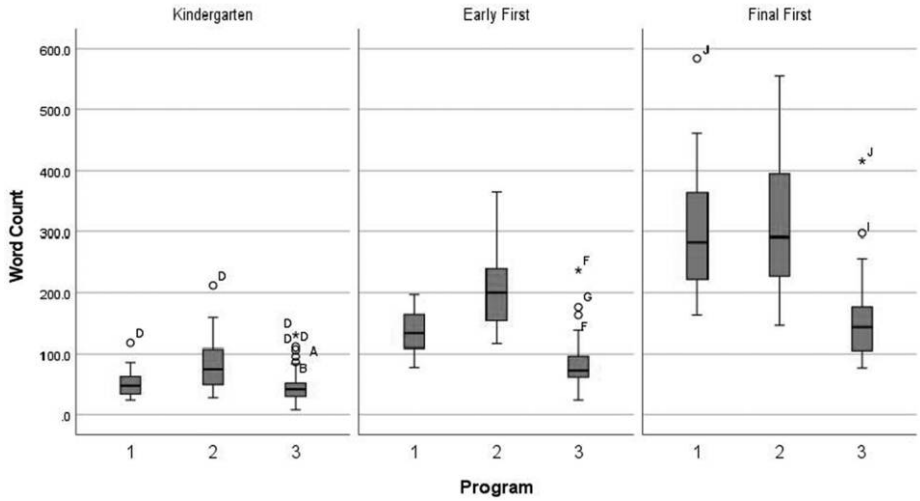

Figure 2. (*Continued*)

Text structure (Fig. 2*f*) had a clearly defined range for kindergarten texts, with most texts scoring a 1. First-grade texts were more variable in text structure, ranging from 1 to 4 in early first grade (with one outlier of 5) and from 1 to 5 in final first grade. Kindergarten texts ranged in overall length from 20 to 200 words, with most texts below 100 words (Fig. 2*g*).

Kindergarten texts ranged in overall length from 8 to 212 words, with most texts below 100 words. Early-first-grade texts ranged from 24 to 365 words, and final-first-grade texts from 77 to 584 words, demonstrating a clear progression.

Variability across programs. In keeping with the variability within grade bands, the three text programs differed with one another about which level of complexity constituted each grade band. In particular, Programs 1 and 3 were more challenging in decoding and semantics than Program 2 (Fig. 2*b* and 2*c*). Programs 1 and 3 included kindergarten texts that covered the entire range of decoding (1–5), whereas Program 2 texts were concentrated below scores of 3, with no texts scoring 5. First-grade texts continued this pattern, with Program 2 texts clustered below the other programs in the 2–3 range and no texts scoring 5. Similarly, the kindergarten texts of Programs 1 and 3 ranged from 1 to 5 in semantics, but only outlier Program 2 texts scored above 1. The first-grade Program 2 texts were also concentrated in the lower end of the range (1–3 in early first, 1–4 in final first) compared with Programs 1 and 3.

MSL and syntax were relatively similar across programs (Fig. 2*d* and 2*e*). MSL was similar across programs for all three grade bands, although Program 2 had narrower ranges than the other two programs. For syntax, most texts concentrated in the lower levels in kindergarten (1–2) and early first grade (2–3) across programs. By the end of first grade, Program 1 had higher syntax scores (3–4) than Programs 2 and 3.

The closest agreement across the three programs was for the structure indicator (Fig. 2*f*) in the kindergarten texts (concentrated in the 1–2 range for all three programs). By the end of first grade, however, Program 1 texts were again the most difficult (4) compared with Programs 2 and 3. Interestingly, the three programs also disagreed on the appropriate length of first-grade texts (Fig. 2*g*). Program 3 literacy had the shortest early-first- and final-first-grade texts, followed by Program 1 and Program 2, in that order.

RQ 3: Which Word-, Sentence-, and Text-Level Variables Predict GRLs, and How Much of the Variance in GRLs Do They Explain?

Multiple regression analysis was used to evaluate the contribution of the seven variables to text levels. The differences in these variables across programs meant that a dummy-coded categorical variable for program was included as a covariate in all regression analyses. The linear regression equation was cross validated to estimate how well the model might apply to texts not included in the current sample. The 510 texts in the sample were randomly assigned to two subsets of 255 texts each (with each set including either 25 or 26 books at each level). One of these two sets was randomly selected as the calibration sample. ANOVA analyses were used to confirm that there were no significant differences between the two text sets on the key variables (see Table 4).

The model was built in the calibration sample using a backward-stepping procedure and then tested in the validation sample. All seven text-level predictors were entered simultaneously along with the dummy-coded variable for program as a control variable. Two word-level predictors, decoding and MLWF, were not significant predictors of GRLs ($p > .05$). Models including and excluding each of these variables were compared using Akaike's information criterion (AIC) and Bayesian information criterion (BIC) values to identify the most parsimonious model that best fit the data. For the full model, the AIC was 139.44 and the BIC was 174.85. Removing either decoding or MLWF reduced both the AIC (137.72 and 139.81, respectively) and BIC (169.69 and 171.69, respectively); removing both further reduced the BIC (166.39), whereas the AIC remained in the same range (138.06). Thus, a model including five text variables (semantics, MSL, syntax, structure, and word count) was used to produce a multiple regression equation in the calibration sample. The constant and beta weights from this analysis were then used to compute a predicted GRL (1–10) for each of the 255 texts in the validation sample. A simple correlation was computed between the predicted and actual GRLs for the 255 texts in the second subset. This correlation ($r = 0.886$, $p < .01$) was compared with the correlation between the actual and predicted values from the calibration sample ($r = 0.897$, $p < .01$). The decline in the correlation from the training set to the testing set was less than 0.1 (0.011), supporting the probable stability of the multiple regression equation. Furthermore, the difference in the R^2 between the calibration (0.80) and validation samples (0.78) was tested for statistical significance by computing the standard errors for

Table 4. Characteristics of the Two Samples for Cross-Validation of Regression Model and Results of ANOVA Analyses

Variable	Calibration Sample	Validation Sample	F	p	η^2
n	255	255			
MLWF	3.61 (.29)	3.60 (.29)	.03	.86	0
Decoding	2.76 (1.17)	2.74 (1.21)	.03	.86	0
Semantic	2.37 (1.11)	2.44 (1.17)	.44	.51	0
MSL	6.15 (1.56)	6.03 (1.62)	.67	.42	0
Syntax	2.18 (1.00)	2.25 (1.01)	.70	.40	0
Structure	2.29 (1.20)	2.31 (1.19)	.03	.85	0
Word count	140.38 (106.36)	143.81 (114.47)	.12	.73	0

Note.—MLWF = mean log word frequency; MSL = mean sentence length.

the R^2s and computing the 95% confidence interval using the formula provided by Alf and Graf (1999). The resulting confidence interval (–0.05, 0.09) included zero, indicating that the change in R^2 was not statistically significant. The results of the regression analyses in each sample and the full sample are presented in Table 5.

Dominance analysis was used to evaluate the relative contribution of the five text predictors (semantics, MSL, syntax, structure, and word count) to the full model. The overall R^2 for the five factors predicting text level was 0.80. Table 6 presents the additional contribution of each variable to R^2 in each possible subset model (which determines complete dominance), and Table 7 presents the average contribution for each variable at each subset of the model (which determines conditional dominance) and overall (which determines general dominance). Table 8 provides the dominance analysis coefficients for each variable pair (*Dij*) for complete, conditional, and general dominance as well as the means, *SD*s, and reproducibility of these coefficients across the 1,000 bootstrapped samples. $Dij = 1$ indicates that i dominates j, 0 indicates that j dominates i, and 0.5 means that dominance could not be determined at that level. Reproducibility coefficients indicate the proportion of bootstrapped analyses that confirmed the dominance result.

Word count appeared be the most important predictor, contributing the most unique variance for all models analyzed (0.25 on average, range 0.06–0.68). Word

Table 5. Results from Regression Analyses Predicting Guided Reading Levels

Variable	*B*	*SE*	*t*	R^2
	Calibration Sample			
				.80
Constant	.47	.42	1.13	
Semantic load	.38**	.10	3.91	
Mean sentence length	.11	.07	1.51	
Syntax	.80**	.14	5.56	
Structure	.49**	.11	4.35	
Word count	.01**	0	7.91	
	Validation Sample			
				.78
Constant	−.23	.41	−.55	
Semantic load	.19*	.09	2.14	
Mean sentence length	.33**	.08	4.23	
Syntax	.59**	.15	3.84	
Structure	.48**	.11	4.35	
Word count	.01**	0	8.99	
	Full Sample			
				.80
Constant	.10	.29	.34	
Semantic load	.27**	.07	4.15	
Mean sentence length	.22**	.05	4.21	
Syntax	.70**	.11	6.68	
Text structure	.48**	.08	6.20	
Word count	.01**	0	12.02	

Note.—Program was included as a dummy-coded categorical covariate.

* $p < .05$.

** $p < .01$.

Table 6. Dominance Analysis Results from Text Predictors of Guided Reading Levels

		Unique Contribution of Predictor to Guided Reading Level				
Subset Model	R^2	Sem.	Syn.	Str.	MSL	Word Ct.
	Models with One Text Predictor					
Sem.	.27		.38	.35	.24	.43
Syn.	.62	.03		.09	.01	.15
Str.	.56	.06	.15		.10	.17
MSL	.39	.11	.23	.26		.34
Word Ct.	.68	.02	.09	.04	.05	
	Models with Two Text Predictors					
Sem. + Syn.	.65			.08	.01	.13
Sem. + Str.	.62		.10		.08	.12
Sem. + MSL	.51		.15	.19		.24
Sem. + Word Ct.	.70		.07	.04	.05	
Syn. + Str.	.71	.02			.01	.08
Syn. + MSL	.63	.03		.09		.14
Syn. + Word Ct.	.77	.01		.01	.01	
Str. + MSL	.66	.04	.06			.10
Str. + Word Ct.	.73	.02	.06		.04	
MSL + Word Ct.	.73	.01	.04	.03		
	Models with Three Text Predictors					
Sem. + Syn. + Str.	.73				.01	.06
Sem. + Syn. + MSL	.66			.08		.12
Sem. + Syn. + Word Ct.	.78			.01	.01	
Sem. + Str. + MSL	.70		.04			.08
Sem. + Str. + Word Ct.	.74		.05		.04	
Sem. + MSL + Word Ct.	.75		.03	.03		
Syn. + Str. + MSL	.72	.02				.07
Syn. + Str. + Word Ct.	.78	.01			.01	
Syn. + MSL + Word Ct.	.78	.01		.02		
Str. + MSL + Word Ct.	.77	.01	.02			
	Models with Four Text Predictors					
Sem. + Syn. + Str. + MSL	.74					.06
Sem. + Syn. + Str. + Word Ct.	.79				.01	
Sem. + Syn. + MSL + Word Ct.	.78			.02		
Sem. + Str. + MSL + Word Ct.	.78		.02			
Syn. + Str. + MSL + Word Ct.	.79	.01				
	Model with Five Text Predictors					
Sem. + Str. + Syn. + MSL + Word Ct.	.80					

Note.—Sem. = semantic; Str. = structure; Syn. = syntax; MSL = mean sentence length; Word Ct. = word count. All models included program as a control variable.

count demonstrated complete dominance over the other four text variables, with reproducibility from 0.86 (over syntax) to 0.99 (over structure) and 1.0 (over semantics and MSL). Syntax contributed the next greatest amount of unique variance (average = 0.19, range = 0.02–0.62). Syntax demonstrated complete dominance over MSL and semantics (reproducibility = 0.89 and 0.88) and general dominance over structure (reproducibility = 0.78). Structure followed, averaging a contribution of 0.18 (range = 0.02–0.56). Structure demonstrated complete dominance over semantics (reproducibility = 0.89) and conditional dominance over MSL (reproducibility = 0.89). MSL

Table 7. Dominance Analysis Average Predictor Contributions of Each Text Predictor to Each Model Size

	Average R^2 Contribution					
Variable	Overall Average	0 Text Predictors	1 Text Predictor	2 Text Predictors	3 Text Predictors	4 Text Predictors
Semantic	.07	.27	.06	.02	.01	.01
Syntax	.19	.62	.21	.08	.04	.02
Structure	.18	.56	.19	.08	.04	.02
MSL	.11	.39	.10	.03	.02	.01
Word count	.25	.68	.27	.14	.08	.06

Note.—MSL = mean sentence length. All models included program as a control variable.

Table 8. Predictor Dominance Relations and Reproducibility

i	*j*	*Dij*	*M (SD)Dij*	*Pij*	*Pji*	*Pn0ij*	Reproducibility
Complete Dominance							
Semantic	Syntax	0	.06 (.16)	0	.88	.11	.88
Semantic	Structure	0	.05 (.10)	0	.89	.11	.89
Semantic	MSL	.5	.47 (.10)	0	.04	.96	.96
Semantic	Word count	0	0 (0)	0	1	0	1
Syntax	Structure	.5	.51 (.06)	.02	0	.98	.98
Syntax	MSL	1	.95 (.16)	.89	0	.11	.89
Syntax	Word count	0	.07 (.17)	0	.86	.14	.86
Structure	MSL	.5	.66 (.23)	.32	0	.68	.68
Structure	Word count	0	00 (.04)	0	.99	.01	.99
MSL	Word count	0	0 (.02)	0	1	0	1
Conditional Dominance							
Semantic	Syntax	0	.04 (.13)	0	.93	.07	.93
Semantic	Structure	0	.04 (.13)	0	.92	.08	.92
Semantic	MSL	0	.27 (.26)	.01	.47	.52	.47
Semantic	Word count	0	0 (0)	0	1	0	1
Syntax	Structure	.5	.69 (.32)	.47	.08	.45	.45
Syntax	MSL	1	.95 (.15)	.90	0	.10	.90
Syntax	Word count	0	0 (.04)	0	.99	.01	.99
Structure	MSL	1	.94 (.16)	.89	0	.11	.89
Structure	Word count	0	0 (.02)	0	1	0	1
MSL	Word count	0	0 (0)	0	1	0	1
General Dominance							
Semantic	Syntax	0	0 (0)	0	1	0	1
Semantic	Structure	0	0 (0)	0	1	0	1
Semantic	MSL	0	.03 (.17)	.03	.97	0	.97
Semantic	Word count	0	0 (0)	0	1	0	1
Syntax	Structure	1	.78 (.41)	.78	.22	0	.78
Syntax	MSL	1	1 (0)	1	0	0	1
Syntax	Word count	0	0 (0)	0	1	0	1
Structure	MSL	1	1 (.04)	1	0	0	1
Structure	Word count	0	0 (.03)	0	1	0	1
MSL	Word count	0	0 (0)	0	1	0	1

Note.—*Dij* = dominance of *i* over *j*; *M* (*SD*) = Mean and *SD* of *Dij* over 1,000 bootstrapped samples; *Pij* = the proportion of bootstrapped analyses in which *i* dominated *j*; *Pji* = the proportion of bootstrapped analyses in which *j* dominated *i*; *Pn0ij* = the proportion of bootstrapped analyses in which dominance could not be established; MSL = mean sentence length.

had an average unique contribution of 0.11 (range = 0.01–0.39) and demonstrated conditional dominance over semantics with low reproducibility (0.47) and general dominance with high reproducibility (0.97). Semantics was the least important predictor, with an average contribution of 0.07 (range = 0.01–0.27). Semantics did not exhibit any form of dominance over any other factor in the model.

Discussion

As in previous studies (Cunningham et al., 2005; Hatcher, 2000; Pitcher & Fang, 2007), features at the sentence and text levels predicted the assignment of texts to levels in the current study. A single word-level variable, semantics, which was not examined in the three prior studies, also showed differences across levels but accounted for only a small portion of the variance and was the least important predictor of text level. Similar to the findings of Cunningham et al., the present study found that neither the word-level measures of word frequency nor decodability predicted the placement of texts in levels.

In addition to replicating the results of previous studies, the current study makes several unique contributions to the literature. First, the current study is the first to consider the variables accounting for the GRL gradient of text complexity (Fountas & Pinnell, 2012). The RR leveling system, the focus of prior studies (Cunningham et al., 2005; Hatcher, 2000; Pitcher & Fang, 2007), continues to be used in the RR intervention, but it is the GRL system that has become the dominant method for reporting the complexity of beginning reading texts in the marketplace. Teachers' guides can be purchased for the three programs analyzed in this study, but texts are offered as a primary mechanism for student learning in all three programs. The text-leveling system is central to claims of efficacy in children's learning (Ransford-Kaldon et al., 2010, 2013). By contrast, no evaluations of RR have highlighted the lists of recommended texts as a source of the efficacy of RR for student achievement (D'Agostino & Harmey, 2016; Sirinides et al., 2018).

Second, the current study is the first to examine the consistency of levels within grade bands, specifically the three that comprise a critical period of reading development: kindergarten, early first grade, and final first grade. Both Cunningham et al. (2005) and Pitcher and Fang (2007) examined texts at levels 5, 10, 15, and 20 that span grades K–2, whereas Hatcher's (2000) interest lay in the general progression across all 20 RR levels. System developers describe the text-level gradient as ensuring that students' placement in a level matches their reading proficiency (Fountas & Pinnell, 2012). The attribution of a gradient to the system suggests that a student with a designation of Level E (early first grade) differs in reading capacity from a student with a designation of the subsequent levels of F or G. The findings of this study question this attribution of a text gradient within and across grade bands.

Third, new measures of text features have become available in the past decade, such as large databases of AoA (Kuperman et al., 2012) and concreteness norms (Brysbaert et al., 2014). Thus, we were able to analyze the text-leveling system with a comprehensive set of measures that has been validated empirically (Fitzgerald et al., 2015a) and that has been anchored in a theoretical model (Mesmer et al., 2012). Furthermore, we were able to compare the importance of the text predictors to GRLs by conducting

dominance analysis, which has not been used in previous studies. We were particularly interested in whether, in light of the new measures and extensive research on reading acquisition conducted over the past 2 decades (Cain et al., 2017; Yap & Balota, 2015), refinements had been made in the GRL system. We also believed that the widespread use of this leveling system and the consequential decisions for which it is used, such as intervention participation, made it imperative to examine the literature to determine whether new empirical evidence has been generated to validate the 10 constructs of the GRL system.

In the next section of the article, we apply the criterion of evidence that supports reading acquisition first to the five variables that accounted for the distinctions in text levels and, second, to the two variables that did not predict the levels of texts.

The Five Variables That Account for Text Progression

Five variables predicted assignment of texts to levels, accounting for 80% of the variance: semantics, syntax, structure, MSL, and word count. Word count was highly predictive of text placement, showing dominance over the other four predictors. Early word recognition acquisition is aided when the words in lessons are prominent in the texts that students are given for application (Juel & Minden-Cupp, 2000; Lesgold et al., 1985). There is simply no evidence, however, that either fewer or more words in a text will determine whether children recognize words. Reading an entire text is necessary in an assessment context and may influence the reading rate sustained by readers (Valencia et al., 2010), but, in an instructional context, a page or even a single sentence can be read, and the task can be curtailed or shared with a peer or teacher. The length of the text or, more likely, the length of the task (Hayden et al., 2019) can influence some students' willingness to persevere in reading. However, there is simply no empirical or theoretical support for the number of words in texts as a major mechanism for word recognition.

The next variable of prominence in the prediction of text levels was the syntax measure, which indicates intersentential complexity in the ELI system, not sentence length. As the descriptions of *Brown Bear* and *Frog and Toad* illustrate, intersentential complexity is low when adjacent sentences share numerous words and high when they do not. The question of how or whether intersentential complexity supports word recognition proficiency is complex. In the context of an immediate text, children may read words more quickly when they recognize the pattern of repeated words or phrases from sentence to sentence (Mesmer, 2009). It may be that, when words are repeated between and across sentences, children could be receiving some of the repetition that underlies automaticity in recognizing specific words and patterns (Ehri, 2005). However, a competing argument can be raised as to whether the form of word repetition illustrated in *Brown Bear* supports independent word recognition. Young children who quickly pick up on the repetition of words between sentences may recite the pattern. This activity of repeating a pattern may be an important one as part of early print awareness for children, but, as a mechanism for word recognition acquisition and improving reading proficiency over time, its role is less clear.

The variable of intersentential complexity is closely tied to the third variable that accounted for text levels: text structure. All but a handful of the 510 texts in the present sample received the same score for syntax and structure. The structure measure

in the ELI system was an amalgam of three constructs: phrase diversity, text density, and noncompressibility. The constituents of structure in the ELI system that capture the degree of repetition of words and phrases are different from the text structure variable in the GRL system, which is described in terms of narrative or factual text structures (Fountas & Pinnell, 2012). In other contexts, however, Fountas and Pinnell (1996) have described predictable structures such as that in *Brown Bear* as characterizing the structure of texts at early levels. Frequent claims have been made that predictable structures support reading development (e.g., Holdaway, 1984), but empirical validation that the repetition of words, phrases, and sentences in predictable structures supports independent word recognition has not been forthcoming. Rather, students without rudimentary decoding skills appear to not increase in word recognition proficiency when instruction emphasizes predictable texts (Boylin, 1998; Johnston, 2000).

The fourth variable—sentence length—was not as powerful in predicting text levels as other variables, but it did have general dominance over semantics. Existing studies showing that decreasing sentence length can depress comprehension (Lupo et al., 2019) have been conducted with older students, not with beginning readers. In this study, the difference of an average of 3.4 more words at Level J compared with Level A represents the presence of phrases or clauses in the higher levels of text, a variable that predicts text levels in the RR system (Cunningham et al., 2005). The length and the number of T-units in sentences could be predicted to influence comprehension for young readers, but the influence of this variable on students' word recognition is less clear. In a study of features of words known and unknown to first graders, children in the bottom quartile in spring read the same percentage of words correctly in sentences with 6 words as in sentences with 12 words (Hiebert et al., 2020). What did influence word recognition was the frequency and length of words, not sentence length or where in a sentence (beginning or end) a word occurred.

Of the five predictors of text levels, the semantics variable has the strongest empirical base but was the least dominant predictor. Each of the three constructs of the ELI semantics variable has been shown to influence the speed of recognizing a word—AoA (Morrison & Ellis, 1995), abstractness (Kroll & Merves, 1986), and rareness (Nagy & Scott, 1990). These three variables influence readers' access to the high-quality lexical representations that underlie word recognition and comprehension (Perfetti, 2007). Although the semantic features of a word underlie the word recognition process, they do not compensate for a lack of decoding (Gerhard & Barry, 1999). Relative to this discussion, the role of the picture-text match, one of the 10 variables of the GRL system, merits discussion. The picture-text match may aid in eliciting the lexical representation of a word, but the use of pictures as a mechanism for recognizing words has not been shown to support independent word recognition (Singer et al., 1973).

In sum, the typical features of texts at the early levels, such as few words, short sentences in repetitive text patterns, and words that represent familiar concepts, may be useful in aiding young children with few prior text experiences in understanding the association between spoken and written words and the directional nature of written English. Continued reliance on these features to differentiate texts as students move into independent word recognition (Ehri, 2005) in early-first- and final-first-grade levels has little grounding in either research or theory.

Sentence- and text-length variables can be critical in comprehension, once readers have at least a modicum of word recognition (Graesser et al., 1996). The Mesmer et al. (2012) framework of beginning texts emphasizes sentence- and text-level variables, in addition to word-level ones, and draws on research to emphasize that word recognition occurs in the context of sentences and texts and depends on lexical access. Furthermore, the Fitzgerald et al. (2015a) findings were based on first and second graders' comprehension scores, not on word recognition scores. The students in the sample were able to read sufficiently well to perform a silent reading maze comprehension task. The performances of students who were not successful with the task were excluded. Our emphasis on word recognition as a foundation for comprehension is compatible with the often validated simple view of reading (Hoover & Gough, 1990). Word recognition is in the service of comprehension. For comprehension to occur, however, readers need to recognize at least a critical portion of the words in texts.

Variables That Do Not Predict GRLs

By contrast to the measures that predicted text levels, the two variables that failed to predict text levels—MLWF and decoding—are substantiated by sizable literatures as features of words that affect word recognition proficiency. English, although having a quasiregular orthography, is an alphabetic language, and alphabetic knowledge is needed for success in reading (Seymour et al., 2003).

Proponents of leveled texts may argue that the texts themselves are not intended to serve as the sole sources of word recognition instruction. Programs 1 and 2 provide teachers' guides with lessons and activities on specific letter-sound patterns. The connections between lessons and texts have not been examined for Program 1, but in the case of Program 2, Murray et al. (2014) found that the words recommended for instruction in the teachers' guide had a tenuous association with the words in the program's texts. Furthermore, for children who need support in developing or strengthening decoding skills, as is the case for many children in Tier 2 and Tier 3 interventions, reading texts at lower GRL levels will not necessarily provide additional opportunities to apply decoding skills as compared with higher-level texts. The clear lack of differentiation in decoding demands and in the word frequencies of the levels assigned to texts calls for caution in the use of leveled texts for Tier 2 and Tier 3 interventions.

Consistencies in Text Complexity within and across Levels and across Programs

At present, the GRLs provide the framework for classroom instruction, Tier 2 and Tier 3 interventions, and school libraries. On the official site where information on the GRL of a text can be purchased, each of the 72,483 books that have been assigned GRLs are described as having been "meticulously reviewed and leveled" by the two developers of the leveling system, "in conjunction with their team of hand-selected levelers" (Fountas & Pinnell, 2021c, para. 2). Based on this statement, a high level of consistency would be expected in the levels of texts within and across programs.

Our findings suggest that these assumptions of comparability of levels within and across programs of texts are not supported.

Consistency within and across levels. On measures other than word count, less than a handful of features showed any differences across levels within grade bands, and these differences were not consistent. The box plots showed that word-level features displayed a high degree of variability within grade bands. Variation within and across programs was considerable for the word-level measure of decodability. For kindergarten, where the average for decodability was 2.6 (medium level), a standard deviation of 1.2 means that decoding demands could range from very high to low.

The one variable that showed consistent patterns within and across levels was MLWF. Word frequency has long been recognized as influencing word recognition, especially in the early stages of word acquisition (Yap & Balota, 2015). Word frequency would be expected to be higher (i.e., easier) in the kindergarten band than in the early first and final first bands. That was not the case, indicating that students get a similar diet of words across grade bands of leveled texts.

Consistency across programs. Despite varying functions in a beginning reading classroom, the three programs in this study are advertised as offering texts sequenced in a similar progression of text complexity. If a school were using all three programs for their advertised functions, teachers would find that expectations as to what students should be able to read at a particular level vary considerably across core instruction, interventions, and content-area instruction. Struggling readers assigned to the intervention texts (Program 2) would find words in the core program (Program 1) and content-area instruction (Program 3) to be more challenging than those used in Tier 2 or Tier 3 instruction.

The presence of more semantically familiar content and less challenging decoding demands in the intervention program than in the core or content-area program might reflect a perception of program developers that struggling readers need less challenging content. First, evidence does not exist that giving struggling readers easier and less text will bring them to higher reading levels (Amendum et al., 2018). Second, the leveling system is presented as representing the same constructs at each level. The variation in word-level features of texts within and across levels of programs leaves uncertainty as to what proficient reading means at specific milestones, such as the end of kindergarten and first grade, when decisions are made about grade promotions or assignments to interventions. The variability of text features across programs and the texts within levels of a program suggests that the use of these measures for student placement into interventions should be a pressing topic for further investigation.

Caveats

As a quantitative analysis of text complexity, this study was limited to features of texts for which tools have been empirically validated. Consequently, constructs identified in the text-leveling system (Fountas & Pinnell, 2012), including genre and form, content, themes and ideas, language and literary features, illustrations, and book and print features, were not addressed. This shortcoming, we argue, does not diminish the importance of this study's findings. First, none of these features addresses the

word-level demands of texts. Furthermore, research to support these features as facilitating word recognition has not been reported. In addition, the factors explored in this study accounted for a large proportion of the overall variance in text levels. Therefore, conclusions regarding the word recognition demands of the target sets of texts would not be changed by analyzing these additional characteristics.

This study's sample of 510 texts distributed across 10 levels is considerably larger than samples from the three previous studies, which represented 20 RR levels with either 80 texts or 200 texts (Cunningham et al., 2005; Hatcher, 2000; Pitcher & Fang, 2007). Moreover, texts in these earlier studies either came from disparate publishers or from an early leveled text program rather than recently published programs. However, even with a sizably larger sample than previous studies, our sample represents only a fraction of the 72,483 texts for which levels can be obtained on the GRL website (Fountas & Pinnell, 2021c) or of the many texts leveled by other systems that claim compliance with the GRL. Even so, several patterns lead us to conclude that the present results can be generalized to leveled texts beyond this sample.

First, the consistency of our findings with those of the three previous studies that examined RR leveling suggests a shared perception across text-leveling systems of what supports young children in becoming proficient readers. Text and sentence features determine the assignment of levels to texts; decodability and word frequency do not. This finding has been consistent across studies using both RR and GRL systems and multiple ways of measuring decodability.

Second, the lack of detailed information on the reliability and validity of text-leveling systems from publishers, including the GRL developers, contributes to the conclusion that the leveling system relies on global evaluations of text complexity, not fine-tuned analyses of features such as the decodability and frequency of words. To date, we have been unable to find reviews of research that validate the roles of such variables as word count, sentence length, and text and sentence predictability in supporting reading acquisition. These variables may support students' comprehension once they have fundamental word recognition, as was the case with the first and second graders in the ELI study (Fitzgerald et al., 2015a). The manner in which these variables support young children in developing the independent word recognition that ensures comprehension is less certain.

Questions

The prominence of leveled texts in beginning reading instruction, despite a lack of research on their efficacy, has recently been described as evidence of a disregard for the science of reading within the educational community (Schwartz, 2019). We propose that such a conclusion fails to address a fundamental issue: the lack of research on beginning reading texts in general. Leveled texts are not the only text type that lacks a research foundation. A paucity of research also exists for the decodable texts that are currently advocated as foundational to successful reading acquisition (Reading League, 2020; Reading Rockets, 2019).

Cheatham and Allor (2012), in the only existing review of decodable texts, concluded that there is very little evidence of a long-term impact on reading growth resulting from practice with decodable texts. They based this conclusion on the two studies that have directly addressed the effects of decodable texts relative to another

text type with the same instructional routine (Jenkins et al., 2004; Juel & Roper/Schneider, 1985), not studies where decodable texts were used as part of interventions with different instructional components (e.g., Foorman et al., 1998; Mathes et al., 2005).

Seidenberg (2017), in his frequently cited book, makes no reference to decodable text, but he does describe the need for beginning readers to have large amounts of text if they are to learn correspondences between spelling and sound. According to Seidenberg, a combination of encounters with large samples of words and timely instruction leads to the emergence in readers of "major statistical patterns" (p. 113). As children encounter more and more words, more fine-tuned details about orthography are acquired. This perspective raises several questions about the use of leveled texts during the reading acquisition period. First, because lower-level GRL texts are significantly shorter than higher-level texts, restricting struggling readers to these lower-level texts could be counterproductive to the goal of providing children with exposure to a large sample of words.

Second, although Seidenberg (2017) does not focus on the number of words that need to comply with particular patterns, proposals and even state mandates have identified percentages of decodability, such as the 75% identified by the California State Board of Education (2002) as necessary in texts purchased for use with beginning and struggling readers. Connections between the words of lessons and the words in texts do appear to aid children in statistical learning (Juel & Minden-Cupp, 2000; Lesgold et al., 1985). Yet, in the Juel and Roper/Schneider (1985) study that is often given as evidence for the efficacy of current decodable texts, 49% of the words in the decodable texts for the early-first-grade period matched the curriculum of words with short-vowel patterns; the figure was 30% in texts at a comparable level in the core reading program. Understanding the degrees of decodability in texts required for students to develop automaticity with critical patterns in words should be a priority of future research on reading acquisition.

Answering questions about degrees of decodability at specific points in reading acquisition and as a function of volume of text can be challenging if new texts are to be created for these studies. An alternative was suggested by Mesmer et al. (2012), who asked whether the thousands of beginning reading texts now available, including leveled and decodable texts, could be sorted according to criteria that support word recognition acquisition. Two such efforts, where a program of leveled texts was re-sorted to comply with a phonics curriculum and not the levels assigned by the publisher, have shown increased reading acquisition for students in the re-sorted text condition (Ehri et al., 2007; Menon & Hiebert, 2005). Digital capacity makes it possible to conduct such a re-sorting on a large scale. Similar efforts seem worthy of future research investigation.

Conclusion

Available evidence points to educators' widespread trust in leveled texts and reliance on levels to differentiate their reading instruction and to track their students' progress (Conradi Smith et al., 2019; Fitzgerald et al., 2015b). The lack of significant differences in levels within grade bands for most measures suggests that students could

be placed in any level and receive a similar learning experience. That is, beginning first graders would receive similar learning opportunities if they had access to texts in the range of Levels E through G rather than to a single level. For some students, chunking longer texts into smaller units may be necessary, but teachers could select texts from any level within a grade band and word-level demands would be similar.

To these concerns we add the equity concerns raised by Hoffman (2017) and Tatum (2013), who have argued that text levels could contribute to Matthew effects in reading (Stanovich, 1986) by giving good readers access to more text and poorer readers less text as well as solidifying readers' self-perceptions in a fixed hierarchy. These concerns, although needing empirical verification, are potentially more alarming given that the text levels themselves do not accurately reflect the best theory and empirical data on text complexity for young readers. Thus, the heavy emphasis on leveled texts in the elementary grades could potentially take up a large proportion of instructional time but have few positive benefits for young readers (Glasswell & Ford, 2011). Additional studies that examine this cost-benefit trade-off in classrooms are needed to fully understand the potential impact of continuing to rely on leveled texts.

Note

Elfrieda H. Hiebert is president/CEO of the nonprofit TextProject; Laura S. Tortorelli is an assistant professor at Michigan State University. Correspondence may be sent to Elfrieda H. Hiebert at hiebert@textproject.org.

References

Alf, E. F., Jr., & Graf, R. G. (1999). Asymptotic confidence limits for the difference between two squared multiple correlations: A simplified approach. *Psychological Methods*, **4**(1), 70.

Amendum, S. J., Conradi, K., & Hiebert, E. (2018). Does text complexity matter in the elementary grades? A research synthesis of text difficulty and elementary students' reading fluency and comprehension. *Educational Psychology Review*, **30**(1), 121–151.

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*. Center for the Study of Reading.

Aukerman, R. C. (1984). *Approaches to beginning reading* (2nd ed.). Wiley.

Azen, R., & Budescu, D. V. (2003). The dominance analysis approach for comparing predictors in multiple regression. *Psychological Methods*, **8**(2), 129.

Boylin, M. (1998). *Effect of predictable and decodable texts and strategy instruction on literacy acquisition* (UMI No. 9840481) [Doctoral dissertation, University of Virginia]. Dissertation Abstracts International, **59**(X), 7.

Browne, M. W. (2000). Cross-validation methods. *Journal of Mathematical Psychology*, **44**(1), 108–132.

Brysbaert, M., Warriner, A. B., & Kuperman, V. (2014). Concreteness ratings for 40 thousand generally known English word lemmas. *Behavior Research Methods*, **46**(3), 904–911.

Budescu, D. V. (1993). Dominance analysis: A new approach to the problem of relative importance of predictors in multiple regression. *Psychological Bulletin*, **114**(3), 542–551. https://doi.org/10.1037/0033-2909.114.3.542

Cain, K., Compton, D. L., & Parrila, P. K. (Eds.). (2017). *Theories of reading development*. John Benjamins.

California State Board of Education. (2002). *2002 reading/language arts/English language development adoption (final report)*. Sacramento.

Cheatham, J. P., & Allor, J. H. (2012). The influence of decodability in early reading text on reading achievement: A review of the evidence. *Reading and Writing*, **25**(9), 2223–2246.

Clay, M. M. (1991). *Becoming literate: The construction of inner control*. Heinemann.

Coltheart, M. (1981). The MRC psycholinguistic database. *Quarterly Journal of Experimental Psychology Section A*, **33**(4), 497–505.

Conradi Smith, K. C., Parsons A. W., Vaughn, M., & Yatzek, J. C. (2019). *Elementary students' text diets: Results from a nationwide survey* [Conference presentation]. Literacy Research Association Conference, Tampa, FL.

Cunningham, J. W., Hiebert, E. H., & Mesmer, H. A. (2018). Investigating the validity of two widely used quantitative text tools. *Reading and Writing*, **31**(4), 813–833.

Cunningham, J. W., Spadorcia, S. A., Erickson, K. A., Koppenhaver, D. A., Sturm, J. M., & Yoder, D. E. (2005). Investigating the instructional supportiveness of leveled texts. *Reading Research Quarterly*, **40**, 410–427. https://doi.org/10.1598/RRQ.40.4.2

D'Agostino, J. V., & Harmey, S. J. (2016). An international meta-analysis of Reading Recovery. *Journal of Education for Students Placed at Risk (JESPAR)*, **21**(1), 29–46.

Dale, E., & Chall, J. S. (1948). A formula for predicting readability. *Educational Research Bulletin*, **27**, 11–20.

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

Ehri, L. C., Dreyer, L. G., Flugman, B., & Gross, A. (2007). Reading Rescue: An effective tutoring intervention model for language-minority students who are struggling readers in first grade. *American Educational Research Journal*, **44**(2), 414–448.

Ferguson, C. J. (2016). An effect size primer: A guide for clinicians and researchers. *Professional Psychology: Research and Practice*, **40**(5), 532–538.

Fitzgerald, J., Elmore, J., Koons, H., Hiebert, E. H., Bowen, K., Sanford-Moore, E. E., & Stenner, A. J. (2015a). Important text characteristics for early-grades text complexity. *Journal of Educational Psychology*, **107**(1), 4–29.

Fitzgerald, J., Hiebert, E. H., Bowen, K., Elmore, J., Relyea-Kim, E. J., & Kung, M. (2015b). Text complexity: Primary teachers' views. *Literacy Research and Instruction*, **54**(1), 19–44.

Foorman, B. R., Francis, D. J., Fletcher, J. M., Schatschneider, C., & Mehta, P. (1998). The role of instruction in learning to read: Preventing reading failure in at-risk children. *Journal of Educational Psychology*, **90**(1), 37.

Fountas, I. C., & Pinnell, G. S. (1996). *Guided reading: Good first teaching for all children*. Heinemann.

Fountas, I. C., & Pinnell, G. S. (2012). *The F&P Text Level Gradient: Revision to recommended grade-level goals*. Heinemann. http://www.heinemann.com/fountasandpinnell/pdfs/WhitePaperTextGrad.pdf

Fountas, I. C., & Pinnell, G. S. (2021a). *Factors related to text difficulty*. https://www.fountasandpinnell.com/Authenticated/ResourceDocuments/FP_FPL_Chart_Factors-Related-to-Text-Difficulty.pdf

Fountas, I. C., & Pinnell, G. S. (2021b). *Leveled Literacy Intervention (LLI)*. https://www.fountasandpinnell.com/lli/

Fountas, I. C., & Pinnell, G. S. (2021c). *Welcome to the Fountas & Pinnell leveled books website*. https://www.fandpleveledbooks.com

Gerhard, S., & Barry, C. (1999). Age of acquisition, word frequency, and the role of phonology in the lexical decision task. *Memory and Cognition*, **27**(4), 592–602.

Glasswell, K., & Ford, M. (2011). Let's start leveling about leveling. *Language Arts*, **88**(3), 208–216.

Graesser, A. C., Swamer, S. S., Baggett, W. B., & Sell, M. A. (1996). New models of deep comprehension. In B. K. Britton & A. C. Graesser (Eds.), *Models of understanding text* (pp. 1–32). Erlbaum.

Hatcher, P. J. (2000). Predictors of Reading Recovery book levels. *Journal of Research in Reading*, **23**(1), 67–77.

Hayden, E., Hiebert, E. H., & Trainin, G. (2019). Patterns of silent reading rate and comprehension as a function of developmental status, genre, and text position. *Reading Psychology*, **40**(8), 731–767.

Hayes, A. F., & Darlington, R. B. (2017). *Regression analysis and linear models: Concepts, applications, and implementation*. Guilford.

Hiebert, E. H. (1999). Text matters in learning to read. *Reading Teacher*, **52**, 552–568.

Hiebert, E. H., Toyama, Y., & Irey, R. (2020). Features of known and unknown words by first graders of different proficiency levels in winter and spring. *Education Sciences*, **10**(12), 389.

Hirsh, D., & Nation, P. (1992). What vocabulary size is needed to read unsimplified texts for pleasure? *Reading in a Foreign Language*, **8**, 689–696.

Hoffman, J. V. (2017). What if "just right" is just wrong? The unintended consequences of leveling readers. *Reading Teacher*, **71**(3), 265–273.

Hoffman, J. V., Roser, N., Patterson, E., Salas, R., & Pennington, J. (2001). Text leveling and little books in first-grade reading. *Journal of Literacy Research*, **33**, 507–528.

Holdaway, D. (1984). *The foundations of literacy*. Heinemann.

Hoover, W. A., & Gough, P. B. (1990). The simple view of reading. *Reading and Writing*, **2**(2), 127–160.

Huberty, C. J., & Morris, J. D. (1992). Multivariate analysis versus multiple univariate analyses. *Psychological Bulletin*, **105**(2), 302–308.

Jenkins, J. R., Peyton, J. A., Sanders, E. A., & Vadasy, P. F. (2004). Effects of reading decodable texts in supplemental first-grade tutoring. *Scientific Studies of Reading*, **8**(1), 53–85.

Johnston, F. R. (2000). Word learning in predictable text. *Journal of Educational Psychology*, **92**(2), 248.

Juel, C., & Minden-Cupp, C. (2000). Learning to read words: Linguistic units and instructional strategies. *Reading Research Quarterly*, **35**(4), 458–492.

Juel, C., & Roper/Schneider, D. (1985). The influence of basal readers on first grade reading. *Reading Research Quarterly*, **20**(2), 134–152.

Koons, H., Elmore, J., Sanford-Moore, E., & Stenner, A. J. (2017). *The relationship between Lexile text measures and early grades Fountas & Pinnell reading levels (MetaMetrics research brief)*. MetaMetrics. https://support.lexile.com/s/article/The-Relationship-Between-Lexile-Text-Measures-and-Early-Grades-Fountas-Pinnell-Reading-Levels

Kroll, J. F., & Merves, J. S. (1986). Lexical access for concrete and abstract words. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, **12**(1), 92–107.

Kuperman, V., Stadthagen-Gonzalez, H., & Brysbaert, M. (2012). Age-of-acquisition ratings for 30,000 English words. *Behavior Research Methods*, **44**(4), 978–990.

Lesgold, A., Resnick, L. B., & Hammond, K. (1985). Learning to read: A longitudinal study of word skill development in two curricula. In G. E. Mackinnon & T. G. Waller (Eds.), *Reading research: Advances in theory and practice* (Vol. **4**, pp. 107–138). Academic Press.

Lupo, S. M., Tortorelli, L., Invernizzi, M., Ryoo, J. H., & Strong, J. Z. (2019). An exploration of text difficulty and knowledge support on adolescents' comprehension. *Reading Research Quarterly*, **54**(4), 457–479.

Mathes, P. G., Denton, C. A., Fletcher, J. M., Anthony, J. L., Francis, D. J., & Schatschneider, C. (2005). The effects of theoretically different instruction and student characteristics on the skills of struggling readers. *Reading Research Quarterly*, **40**(2), 148–182.

Menon, S., & Hiebert, E. H. (2005). A comparison of first graders' reading with little books or literature-based basal anthologies. *Reading Research Quarterly*, **40**(1), 12–38.

Mesmer, H. A. E. (2009). Textual scaffolds for developing fluency in beginning readers: Accuracy and reading rate in qualitatively leveled and decodable text. *Literacy Research and Instruction*, **49**(1), 20–39.

Mesmer, H. A. E., Cunningham, J. W., & Hiebert, E. H. (2012). Toward a theoretical model of text complexity for the early grades: Learning from the past, anticipating the future. *Reading Research Quarterly*, **47**(3), 235–258.

Morrison, C. M., & Ellis, A. W. (1995). Roles of word frequency and age of acquisition in word naming and lexical decision. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, **21**(1), 116–133.

Murray, M., Munger, K. A., & Hiebert, E. H. (2014). An analysis of two reading intervention programs: How do the words, texts, and programs compare? *Elementary School Journal*, **114**(4), 479–500.

Nagy, W. E., & Scott, J. A. (1990). Word schemas: Expectations about the form and meaning of new words. *Cognition and Instruction*, **7**(2), 105–127.

National Geographic Learning/Cengage. (2021a). *Leveled book finder*. https://ngl.cengage.com/search/showresults.do?N=201+4294918395&Ntk=NGL&Ntt=leveled+book+finder&Ntx=mode%2Bmatchallpartial&homePage=false

National Geographic Learning/Cengage. (2021b). *Program overview*. https://ngl.cengage.com/search/productOverview.do?N=201+4294891794+4294918395&Ntk=P_EPI&Ntt=1583539573448049747153934695220842165 40&Ntx=mode%2Bmatchallpartial&homePage=false

National Governors (National Governors Association Center for Best Practices & Council of Chief State School Officers). (2010). *Common Core State Standards for English language arts & literacy in history/social studies, science, and technical subjects*. Common Core State Standards Initiative. http://www.corestandards.org/the-standards

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

Perfetti, C. (2007). Reading ability: Lexical quality to comprehension. *Scientific Studies of Reading*, **11**(4), 357–383.

Peterson, B. L. (1988). *Characteristics of texts that support beginning readers* [Unpublished doctoral dissertation]. Ohio State University.

Pitcher, B., & Fang, Z. (2007). Can we trust levelled texts? An examination of their reliability and quality from a linguistic perspective. *Literacy*, **41**(1), 43–51.

Ransford-Kaldon, C. R., Flynt, E. S., Ross, C. L., Franceschini, L., Zoblotsky, T., Huang, Y., & Gallagher, B. (2010). *Implementation of effective intervention: An empirical study to evaluate the efficacy of Fountas & Pinnell's Leveled Literacy Intervention System (LLI)*. Center for Research in Educational Policy. https://files.eric.ed.gov/fulltext/ED544374.pdf

Ransford-Kaldon, C. R., Ross, C., Lee, C., Flynt, E. S., Franceschini, L., & Zoblotsky, T. (2013). *Efficacy of the Leveled Literacy Intervention system for K–2 urban students: An empirical evaluation of LLI in Denver Public Schools*. Center for Research in Educational Policy.

Reading A–Z. (2020). *Leveled text*. https://www.raz-kids.com/main/RazQuizRoom

Reading A–Z. (2021). *Level correlation chart*. https://www.readinga-z.com/learninga-z-levels/level-correlation-chart/

Reading League. (2020, May). *Decodable text sources*. https://www.thereadingleague.org/wp-content/uploads/2020/11/Decodables-Update-November-2020.pdf

Reading Rockets. (2019). *Decodable text sources*. https://www.readingrockets.org/article/decodable-text-sources

Schwartz, S. (2019, December 3). The most popular reading programs aren't backed by science. *Education Week*. https://www.edweek.org/teaching-learning/the-most-popular-reading-programs-arent-backed-by-science/2019/12

Seidenberg, M. (2017). *Language at the speed of sight: How we read, why so many can't, and what can be done about it*. Basic.

Seymour, P. H. K., Aro, M., & Erskine, J. M. (2003). Foundation literacy acquisition in European orthographies. *British Journal of Psychology*, **94**(2), 143–174. https://doi.org/10.1348/000712603321661859

Simba Information. (2020). *K–12 reading market survey report 2020*. https://www.simbainformation.com/Reading-Survey-12938533/

Singer, H., Samuels, S. J., & Spiroff, J. (1973). The effect of pictures and contextual conditions on learning responses to printed words. *Reading Research Quarterly*, **9**(4), 555–567.

Sirinides, P., Gray, A., & May, H. (2018). The impacts of Reading Recovery at scale: Results from the 4-year i3 external evaluation. *Educational Evaluation and Policy Analysis*, **40**(3), 316–335.

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

Stanovich, K. (1986). Matthew effects in reading: Some consequences of individual differences in the acquisition of literacy. *Reading Research Quarterly*, **21**(4), 360–406.

Stenner, A. J. (1996). *Measuring reading comprehension with the Lexile Framework* (Paper presentation). North American Conference Adolescent/Adult Literacy 4th Annual Meeting, Washington, DC, February. https://eric.ed.gov/?id=ED435977

Tatum, A. (2013, May 22). *Alfred Tatum on literacy education for African American and Latino students* [Video]. YouTube. https://www.youtube.com/watch?v=DwSKB-mg-mU

Valencia, S. W., Smith, A. T., Reece, A. M., Li, M., Wixson, K. K., & Newman, H. (2010). Oral reading fluency assessment: Issues of construct, criterion, and consequential validity. *Reading Research Quarterly*, **45**(3), 270–291.
Yap, M. J., & Balota, D. A. (2015). Visual word recognition. In A. Pollatsek & R. Treiman (Eds.), *Oxford library of psychology: The Oxford handbook of reading* (pp. 26–43). Oxford University Press.

Programs Evaluated

Fountas, I. C., & Pinnell, G. S. (2008). *Leveled literacy intervention*. Heinemann.
National Geographic Learning/Cengage. (2001). *Windows on literacy*. Author.
Reading A–Z. (n.d.). Readingatoz.com

Instructional Texts for Students

Lobel, A. (1972). *Frog and toad together*. Harper & Row.
Martin, B. (1967). *Brown bear, brown bear*. Puffin.
Robart, R. (1991). *The cake that Mack ate*. Little, Brown Books for Young Readers.
Robinson, H. M., Monroe, M., & Artley, A. S. (1962). *The new basic readers*. Scott Foresman.