

## Article

# Flattening the Developmental Staircase: Lexical Complexity Progression in Elementary Reading Texts Across Six Decades

Elfrieda H. Hiebert 

TextProject, Santa Cruz, CA 95060, USA; hiebert@textproject.org

## Abstract

This study examined lexical complexity patterns in elementary reading textbooks across four pivotal decades (1957, 1974, 1995, 2014) to understand how educational reforms have influenced developmental progressions in reading materials. The study analyzed a corpus of 320,000 words from one continuously published core reading program across grades 1–4 for four copyrights. The corpus consisted of a 20,000-word sample for each grade and year, analyzed for type-token ratio, percentage of complex words, and percentage of single-appearing words. Results revealed three major shifts: (a) systematic within-grade complexity increases in earlier programs (1957, 1974) were replaced by flat progression in later programs (1995, 2014), (b) steep across-grade differentiation collapsed with grade-to-grade increases in lexical diversity declining from greater than 100% to under 10%, and (c) first-grade expectations accelerated dramatically, whereas third- and fourth-grade texts remained remarkably stable across all six decades. By 2014, first graders encountered lexical complexity levels that characterized fourth-grade texts in 1957. These findings challenge narratives of declining text complexity and reveal that contemporary elementary readers experience compressed developmental progressions with elevated starting points but minimal growth trajectories. The implications suggest the need for reconceptualizing text design to balance appropriate challenges with systematic scaffolding, particularly for students dependent on school-based literacy instruction.



Academic Editors: Emily Rodgers,  
Tracy Johnson and Jerome  
D'Agostino

Received: 11 October 2025

Revised: 9 November 2025

Accepted: 12 November 2025

Published: 17 November 2025

**Citation:** Hiebert, E. H. (2025).  
Flattening the Developmental  
Staircase: Lexical Complexity  
Progression in Elementary Reading  
Texts Across Six Decades. *Education  
Sciences*, 15(11), 1546. <https://doi.org/10.3390/educsci15111546>

**Copyright:** © 2025 by the author.  
Licensee MDPI, Basel, Switzerland.  
This article is an open access article  
distributed under the terms and  
conditions of the Creative Commons  
Attribution (CC BY) license  
(<https://creativecommons.org/licenses/by/4.0/>).

**Keywords:** text complexity; reading instruction; elementary education; vocabulary development; basal readers; lexical density

## 1. Flattening the Developmental Staircase: Lexical Complexity Progression in Elementary Reading Texts Across Six Decades

The landscape of the texts used for elementary reading instruction has undergone substantial transformation over the past six decades driven by evolving perspectives on text complexity. From [Chall's \(1967\)](#) conclusion in *Learning to Read: The Great Debate* that instructional texts were insufficiently challenging, to the subsequent movement toward authentic literature following *Becoming a Nation of Readers* ([Anderson et al., 1985](#)) and culminating in the Common Core State Standards' (CCSS; [NGA-CBP & CCSSO, 2010](#)) explicit mandate for increased text complexity across all grade levels, educators and policymakers have reconceptualized the role of textual complexity in instructional texts over this 60-year period.

These shifts in policies and practices have occurred against a backdrop of limited empirical understanding of –how text characteristics vary across the elementary grades and how these features influence students' reading. This analytical void is particularly

problematic given both the high-stakes decisions that hinge on third-grade performance and the theoretical frameworks that position third grade as a critical transition point. The influential hypothesis regarding the fourth-grade slump (Chall et al., 1990)—the notion that students encounter a dramatic increase in textual demands as they move from third to fourth grade—remains largely untested empirically in part because we lack systematic documentation of how text complexity differs between third and fourth grades and how it is distinguished from first- and second-grade texts—both currently and historically. Third grade has become the focal point for educational policy decisions ranging from retention practices to large-scale intervention programs. Yet despite this focus, third-grade texts have received surprisingly limited scholarly attention.

This study addresses these gaps by providing a systematic analysis of text complexity patterns across grades 1–4 during four pivotal points of American reading instruction, from the mid-1950s through the mid-2010s. Through careful examination of the features that make texts complex, this research considers how the progression of text complexity over the elementary grades has evolved and how third-grade materials compare to those of fourth grade and the two preceding grades.

## 2. Review of Research

American elementary reading instruction has encompassed diverse approaches over an extended period of time (Aukerman, 1984; Chall & Squire, 1996; Kurtz et al., 2020). At any period in time, schools and districts have had access to a range of materials reflecting different instructional philosophies and approaches (Center for Education Market Dynamics, 2025). However, state textbook adoption policies create leverage points for policy influence. Nineteen states maintain adoption policies; California and Texas—the two largest by student enrollment—operate recommendation-based systems that nonetheless exert substantial market influence on publisher decisions (Doan & Kaufman, 2024). Although teachers frequently modify or supplement adopted materials (Schwartz, 2025), the texts publishers design for large adoption state markets establish baseline expectations that influence broader educational practice.

This study examines one core reading program with continuous publication and market presence in adoption states across six decades. The current analysis traces how American policy shifts and research developments have shaped text complexity within one influential program, not as a comprehensive inventory of all materials available during each era. By examining a program that has remained influential through controlled vocabulary, whole language, and phonics-based pedagogical eras, this study illuminates how policy decisions—particularly those of large adoption states—have influenced the texts available to elementary students, even as alternative approaches and materials continued to coexist in American schools.

This review of research traces research from different eras sequentially to establish the historical trajectory of perspectives on text complexity and the policy landscape that shaped textbook development. This backdrop provides the context for analyzing specific textual features—lexical diversity and the percentages of complex and single-appearing words—known to influence reading proficiency—in materials from grades 1–4 across this 60-year period.

### 2.1. Behaviorist Foundation and a Controlled Vocabulary Model

Early systematic investigations into textbook vocabulary began in the 1920s with Dolch's (1928) study of elementary-grade textbook difficulty. Analyzing reading textbooks from two publishers for first through eighth grades, Dolch employed multiple vocabulary measures: vocabulary range (a term that would come to be called lexical density), new

word introduction rates, word repetitions, and difficulty levels based on established word lists. He concluded that 30–40% of words in textbooks were unknown to students at each grade level and that interviews with teachers indicated that they consistently found reading textbooks too challenging for their students. Dolch recommended that texts should use words known to students rather than the 30–40% they did not know.

The next major advance came when [Gray and Leary \(1935\)](#) prioritized high-frequency words from [Thorndike's \(1927\)](#) research and introduced them through carefully controlled repetition and pacing. Evidence for the number of repetitions for words came from [Gates and Russell \(1938\)](#), who established that average first graders required approximately 35 repetitions to learn high-frequency words such as *the*, *of*, and *what* in controlled texts. Whether the number of repetitions varied for words such as *cat*, *dog*, and *balloon*—words that are both highly concrete and present in the oral vocabularies of most children—was not examined.

The controlled vocabulary approach moved from pedagogical theory to classroom reality as Gray and Gates—positioned at Scott, Foresman ([Elson & Gray, 1931](#)) and Macmillan ([Gates & Huber, 1931](#)), respectively—built systematic word repetition into the nation's most widely adopted reading programs. For example, in the 1956 first-grade texts of Scott, Foresman, which Gray ([Gray et al., 1956](#)) authored, each unique word appeared a minimum of 12 times across one level of first-grade texts, no page introduced more than one new word, and no story contained more than three new words. The kind of text that resulted from these guidelines is illustrated in an example from Macmillan's program ([Gates et al., 1957](#)), *Mike Rides*, which consisted of six unique words (*who*, *rides*, *Jeff*, and *Mary*, *Mike*) that were repeated in variations such as "Jeff rides," "Mary rides," and "Mike rides."

[Chall's \(1967\)](#) review of reading texts fundamentally challenged the model that had been present for four decades. Chall reviewed features of first- through third-grade texts from the two reading programs that dominated the marketplace during the 1950s and 1960s: every fourth lesson in the [Ginn \(1961\)](#) program, and every eighth lesson in the [Scott Foresman \(1956\)](#) program. Chall's interest lay in the number of new words per 100 running words and the total number of words in texts. Features of words such as their decodability or frequency were not considered. Both programs introduced new words at rates of 1.0–2.1 words per 100 running words. Whereas text selections expanded from 76 words in beginning first-grade texts to 916 words in third-grade materials, new word density decreased from 1.85 to 1.35 words per 100.

[Chall \(1967\)](#) interpreted these findings as evidence that the sight-word model underlying core reading programs had severely constrained vocabulary development. However, Chall provided no data on student performance with the texts, especially that of students whose literacy experiences occurred primarily in school. Even so, her research changed the content of the next generation of core reading programs, as demonstrated by [Heitz \(1979\)](#), who found increased numbers of unique words with decreased repetitions of words in first-grade texts during the late 1970s.

Rather than examining the influence of her work on subsequent reading texts in the primary grades, Chall shifted focus to science and social studies textbooks across grades 4, 8, and 11 from 1974 to 1982 and 1985 to 1989 ([Chall et al., 1991](#)). The analysis showed that fourth-grade content textbooks demonstrated fifth- to eighth-grade readability levels, whereas eighth-grade texts ranged from seventh- to eighth- and ninth- to 10th-grade difficulty. A decline in text complexity occurred at one level and for one genre only: 11th-grade social studies texts. Drawing on the norms of the Metropolitan Achievement Test, Chall et al. concluded that only 22% of fourth-grade social studies materials and 0% of science materials matched students' reading levels, compared to 67% of reading textbooks.

Rather than viewing this ability-complexity gap as problematic, [Chall et al. \(1991\)](#) advanced the hypothesis that decades of textbook simplification had systematically weakened students' reading abilities, including declines in SAT scores. Demographics had changed in the United States, and data for this explanation for declining SAT scores was correlational. However, the perception—that lowered expectations had generated self-perpetuating declines in achievement—persisted and was particularly influential to the development of the CCSS ([NGA-CBP & CCSSO, 2010](#)).

[Hayes et al. \(1996\)](#) conducted another influential historical analysis of approximately 1.4 million words from 800 textbooks published between 1919 and 1991. They employed a unique measure titled LEX that included only content, and not common, words in texts. They calibrated baseline proficiency on levels of newspaper texts (0.0 LEX) with positive scores indicating greater complexity. The analysis revealed three distinct evolutionary phases. During 1919–1945, schoolbooks followed traditional developmental progression, advancing from –51.8 LEX at primer level to –13.9 LEX by eighth grade. The post-World War II period (1946–1962) showed primer texts dropping more than 16 LEX units to –67.9. The third era (1963–1991) showed a mixed pattern: first- and second-grade materials were restored to prewar difficulty, but fourth- through eighth-grade texts had declined on the LEX measure from the previous periods.

## 2.2. Contributions of Cognitive Science

The 1980s brought a new research perspective that focused specifically on how text characteristics affected comprehension. Unlike [Chall's \(1967\)](#) examination of instructional approaches, cognitive scientists directly studied the effects of readability formulas on students' text comprehension ([Beck et al., 1984](#); [Davison & Kantor, 1982](#)). This research provided crucial evidence that controlling vocabulary and sentence length—hallmarks of the behaviorist model—could impede rather than support reading development. However, these studies contained a critical limitation: They were conducted with second graders and older students, not the beginning readers who were the primary audience for controlled vocabulary texts.

The cognitive science critique culminated in *Becoming a Nation of Readers* ([Anderson et al., 1985](#)), which called for loosening the control of readability formulas built on the controlled vocabulary model. The report argued that rigid adherence to readability formulas “destroyed the flexibility needed to write interesting, meaningful stories” (p. 47). Crucially the report identified a fundamental gap in the field: “What the field does need is an understanding of the concepts at work” (p. 47) regarding how text characteristics support beginning reading acquisition.

This recommendation, which was intended to spur activity on text features (e.g., engaging content) as well as program features (e.g., repetition of vocabulary), was not as compelling as the recommendation to loosen readability formulas on text creation. Policymakers chose to eliminate controlled vocabulary altogether. Two years after the report, [California English/Language Arts Committee's \(1987\)](#) mandated that reading textbooks contain only authentic literature for core reading programs to be purchased with state funds. Texas followed with similar guidelines in 1990. The size and centralized adoption processes of these states meant their mandates effectively drove national textbook development.

The movement of core reading programs to mandate only authentic text in reading programs represented the first major policy breaks from the Gates and Gray models. Research supporting the use of only authentic texts, especially with young children, was notably limited when these policies were formalized. The recommendation to lessen the hold of readability formulas on texts *Becoming a Nation of Readers* ([Anderson et al., 1985](#))

did not elicit research to examine how the features of authentic texts influenced reading acquisition and development.

Rather, research focused on features of instructional context related to reading engagement (Alvermann & Guthrie, 1993). Another notable strand in text complexity research within the cognitive science movement is exemplified by Graesser et al.'s (2003) development of Coh-Metrix, a computational tool that analyzes textual complexity such as cohesion. Although this body of research provides valuable insights for supporting comprehension across different student proficiency levels, the studies were conducted exclusively with middle through high school populations rather than at the elementary level.

Research on the effects of abandoning controlled vocabulary in primary grades was scarce, although available evidence indicated that first-grade texts had grown more challenging. Hiebert's (2005) analysis revealed that the 1993 Scott Foresman program—developed in response to California's (California English/Language Arts Committee, 1987) and Texas's (Texas Educational Agency, 1990) mandates for authentic literature—represented a dramatic watershed in beginning reading texts. Two fundamental changes distinguished the 1993 texts from earlier programs: The number of new unique words per 100 running words increased by 500% at first grade (from 5 in 1983 to 29 in 1993), whereas the percentage of single-appearing words increased ninefold at the beginning of first grade (from 5% to 46%).

### 2.3. Movement to Decodable Text

When California tied for 39th place out of 40 states on the National Assessment of Educational Progress's first state-by-state comparison (Campbell et al., 1996), the poor performance of California's fourth graders was attributed to the whole language policy of the textbook mandates (Levine, 1996). This interpretation prompted a shift in beginning reading instruction from authentic texts toward decodable texts. Texas (Texas Education Agency, 1997) adopted a particularly distinctive approach to decodability, one that focused on individual phoneme-grapheme correspondences and required reading programs to include lessons and accompanying texts for each of the 44 phonemes in English. Even though a small set of high-frequency words could be taught by sight, texts were considered decodable only when students had already been taught the phoneme-grapheme patterns for all other words—either in the current lesson or in previous lessons (Stein et al., 1999). This approach became known as Lesson-to-Text Match (LTTM) and was designed to enable young readers to decode every word accurately. Both Texas and California established minimum decodability thresholds: 80% of words in Texas and 90% in California (California English/Language Arts Committee, 1999) had to meet the decodability criterion with the remainder taught as sight words.

However, these policy mandates outpaced the available research base. When the LTTM model was implemented in textbooks in Texas and California, fundamental questions remained unaddressed: How many lessons do children require to recognize a letter-sound correspondence in unfamiliar words? How does the complexity and frequency of letter-sound correspondences influence learning, particularly for children whose primary literacy instruction occurs in school? More than 25 years later, during which the LTTM has become even more dominant in instructional and intervention programs (Lane et al., 2025), these and other essential questions about the model remain unanswered.

The programs adopted by Texas as complying with LTTM requirements revealed vestiges of the authentic literature perspective in which vocabulary control, specifically word repetition—had been eliminated. Foorman et al. (2004) found that first graders in 2000 encountered approximately 32 new words per week—double the rate of the mid-1980s—with significantly less repetition to support word learning. Critically, across four of the six programs examined, approximately 70% of nondecodable words appeared only once within each 6-week instructional block, and only 14–27% of these words were repeated two to five times. This represented a stark departure from earlier basal readers, in which words typically appeared 20 times across the first 10 passages. These findings suggest that first-grade texts in 2000, even under decodable text mandates, placed considerably greater demands on a beginning reader’s memory, vocabulary knowledge, and decoding skills than did the controlled vocabulary texts of previous decades.

Another policy response to fourth graders’ poor performances on the National Assessment of Educational Progress (Campbell et al., 1996) was the initiation of the No Child Left Behind Act (U.S. Congress, 2002). A particularly consequential—indeed, cataclysmic—shift within this legislation was the inclusion of kindergarten within federal reading standards for the first time. Understanding current first-grade texts thus requires recognizing this concurrent transformation: the movement of formal reading instruction downward to kindergarten. Reading expectations experienced a downward push across grade levels, effectively making kindergarten the new first grade, according to Bassok et al. (2016).

Ironically, when the CCSS (NGA-CBP & CCSSO, 2010) cited a decline in text complexity over a 50-year period—a narrative that implicitly included kindergarten—core reading programs had offered kindergarten anthologies for less than a decade. These materials emerged only after NCLB mandated the instructional shift, making critiques of their declining rigor particularly premature.

The emphasis on decodable texts, initially prompted by state mandates in the late 1990s and early 2000s, would prove foundational to subsequent policy developments. Over the past dozen years, as science of reading initiatives gained momentum across states, the clarion call for decodable texts intensified substantially (Schwartz, 2023) becoming the resounding theme of contemporary early literacy instruction and shaping how programs would respond to the next major policy shift: the CCSS.

#### 2.4. The Common Core State Standards

The adoption of the CCSS (NGA-CBP & CCSSO, 2010) represented another policy-driven transformation in beginning reading texts. Unlike previous reforms that provided general guidance, the CCSS devoted an entire standard to increasing students’ capacity with complex text, establishing a “text complexity staircase” with accelerated Lexile levels beginning at the grade 2–3 band. The CCSS text complexity framework rested on two research pillars. First, Williamson’s (2008) analysis revealed a significant gap between the complexity of texts typically used in high schools and those required for college and career success. ACT’s (2006) study, *Reading Between the Lines*, provided additional grist for increased text complexity. The findings of that study showed that those who encountered more complex texts in high school were significantly more likely to succeed in college coursework. On the basis of these two reports, CCSS developers argued that text complexity needed to be recalibrated across the grades to ensure that students attain levels necessary for college and careers by the end of 12th grade. The most consequential changes in text complexity were at the primary level, where text complexity was to increase to levels formerly assigned to the grade 4–5 range.

The second research source for justifying the increase in text complexity across the grades, but particularly the primary levels, came from [Chall \(1967\)](#), [Chall et al. \(1991\)](#) and [Hayes et al. \(1996\)](#), the contents of which were summarized earlier in this review of literature. CCSS developers asserted that “while the reading demands of college, workforce training programs, and citizenship have held steady or risen over the past fifty years or so, K–12 texts have, if anything, become less demanding” ([NGA-CBP & CCSSO, 2010](#), p. 2). Notably absent from this evidence base were contradictory findings ([Hiebert, 2005](#); [Foorman et al., 2004](#)) that first-grade levels had increased substantially after Texas and California text mandates in the 1990s and early 2000s, respectively.

Shortly after the CCSS publication and subsequent adoption by the majority of American states, [Gamson et al. \(2013\)](#) published a comprehensive challenge to the text decline narrative. Their analysis examined a corpus of roughly 10 million words from textbooks spanning 1905–2004, a sample approximately 10 times the size of [Hayes et al.’s \(1996\)](#). Gamson et al.’s findings directly contradicted the CCSS claims: Text complexity had either risen or stabilized over the past century. For third-grade texts specifically, despite some declines in the early decades of the 20th century, the data showed that textbook difficulty had increased steadily over the past 70 years. Perhaps most significantly, they noted that Hayes et al.’s reliance on McGuffey’s Eclectic Readers for pre-World War I standards was problematic because these readers were essentially compilations of adult reading materials repurposed for younger readers rather than texts originally written for children.

#### *2.5. Post-Common Core State Standards Research on Text Complexity*

Research on text complexity following the CCSS adoption remains limited, although two studies illuminate important trends. [Fitzgerald et al. \(2016\)](#) extended [Hiebert’s \(2005\)](#) analysis of one core reading program from 1960 to 2000 by incorporating the 2007 and 2013 copyrights of the same program. The researchers identified two consequential patterns. First, in contemporary programs (1995, 2007, 2013), first-grade texts began with relatively easy syllable and decoding patterns, but increases in orthographic features resulted in texts with substantially more complex word structures than in earlier eras. In contrast, earlier programs distributed complexity more gradually across the year. Second, whereas earlier programs provided extensive word repetition and redundancy that gradually decreased across the school year, recent programs offered minimal repetition from the outset.

[Kearns and Hiebert \(2022\)](#) conducted the most comprehensive empirical examination of post-CCSS text complexity in grade 1 and grade 3 texts, analyzing word-level features in three widely used U.S. reading programs. They employed factor analysis to identify four empirically distinct dimensions—orthography, length, familiarity, and morphology from a set of 14 measures of word recognition. The words in third-grade texts were more complex than first-grade texts on dimensions of word length and familiarity, as would be expected. But even in first-grade texts, complex words were common. Multisyllabic words comprised 48% of unique words in first-grade texts and 65% in third-grade texts. These patterns suggest post-CCSS texts introduce morphological challenges far earlier than developmental progressions predict.

### **3. The Current Study**

The present investigation addresses the nature of lexical density, presence of complex words, and presence of single-appearing words between the first and second term of texts for four grades that were published by the same company during four distinct decades. This study fills several research voids. First, it traces patterns of these variables in texts from grades 1–4. Studies of first- ([Hiebert, 2005](#); [Fitzgerald et al., 2016](#)) and third-grade texts ([Gamson et al., 2013](#); [Kearns & Hiebert, 2022](#)) from specific years have been

conducted. However, no research has been conducted to determine within- and across-grade progressions of texts over the entire span of grades 1–4.

Second, the four copyrights of texts followed or preceded a major shift in policies related to textbook reform: 1957, prior to Chall's (1967) critique; 1974, subsequent to Chall's critique; 1995, following *Becoming a Nation of Readers'* recommendation for a loosening of controlled vocabulary; and 2014, following the CCSS mandates for more complex text.

By employing consistent measures of lexical density and word frequency within the texts of four grades from 4 years, this study provides the longitudinal perspective necessary to understand how educational reforms—from behaviorist controlled vocabulary through authentic literature mandates to CCSS complexity acceleration—have influenced the texts that students encounter. This comprehensive approach allows for examination of both the developmental progression within any given era and the evolution of that progression across multiple reform cycles.

This study addresses two primary questions:

1. How does the complexity of texts vary from the first to second term of a grade in type-token and complex vocabulary? Does this variation differ across grades and years?
2. What is the nature of differences in lexical density, percentage of complex words, and percentage of single-appearing words across the texts for grades 1–4 at four points in time: 1957, 1974, 1995, and 2014?

## 4. Methods

### 4.1. Sample Selection and Characteristics

This study examined text complexity changes in a single core reading program across four copyright periods: 1957, 1974, 1995, and 2014. These periods represent significant developments in reading research and policy as previously outlined. Copyrights will be referred to as *years* and the texts for the 1957 and 1974 years will be referred to as *earlier*, whereas those for the 1995 and 2014 years will be referred to as *later*.

The program was selected because it is one of three published continuously over this period with substantial market presence (Chall & Squire, 1996; Schwartz, 2019). Focusing on a single program allowed controlled comparison of complexity changes while minimizing confounding variables related to different publishers' approaches or editorial philosophies. Prior research has indicated substantial congruity across large-scale programs with comparison points available through another prominent program's first-grade component (Fitzgerald et al., 2016) and analyses of the first- and third-grade components of this specific program with two other primary core reading programs in the marketplace (Kearns & Hiebert, 2022).

For each grade-year combination, we systematically analyzed 20,000 words, yielding a comprehensive corpus of 320,000 words across grades 1–4 and the 4 years. This sampling threshold was established based on the minimum word count available in first-grade texts from the 1995 and 2014 copyrights, ensuring consistent analytical parameters that preclude potential bias in lexical density and word frequency calculations. The resulting corpus provides robust statistical power for detecting meaningful complexity variations while representing instructional materials encountered in classroom settings.

Within each grade-year cell, the 20,000-word sample was evenly distributed across instructional terms (10,000 words per term). For first-grade texts in later copyright years, where 20,000 words constituted the complete program, no intermediate division existed; however, materials from higher grades maintained the temporal bifurcation between first and second terms.

Analysis focused exclusively on core instructional materials—specifically, primary textbooks (termed “anthologies” in later editions)—rather than supplementary or specialized texts. Whereas later years included additional resources such as leveled and decodable texts, the essential programmatic components purchased by states and districts have remained the primary readers or anthologies. This methodological approach ensures that findings accurately reflect elementary students’ predominant experiences with texts over the years.

#### 4.2. Text Complexity Measures

Three features of texts were analyzed: (a) type-token ratio (TTR), (b) percentage of complex words, and (c) percentage of single-appearing words. The first measure, TTR, indicates lexical diversity by calculating the number of unique words relative to total words in a text. Higher TTRs suggest greater vocabulary demands and more sophisticated language use. This measure has been validated as an indicator of lexical sophistication and correlates with traditional readability assessments (Malvern et al., 2004).

The second measure examined word complexity through the percentage of complex words—those with frequencies of nine or fewer per million words. This draws on the low-frequency and rare word zones (Hiebert et al., 2018) from the Zeno et al. (1995) corpus. The low-frequency zone (1–9 per million) contains 7858 words when morphological family members are eliminated, whereas the rare zone (less than 1 per million) contains 124,405 words, although this varies by text content. Low-frequency and rare words were combined to represent complex words.

The unit for analyzing complex words was the token (number of total words) rather than type (number of unique words). Token analysis better reflects the reading experience students encounter. Each instance of a complex word presents a cognitive challenge regardless of whether students have encountered it previously in the text. A passage with a high percentage of complex vocabulary creates cumulative cognitive demand that type analysis would underestimate.

The third measure was the number of single-appearing words or singletons. The frequency of single-appearing words serves as a critical indicator of text complexity. Specifically, single-appearing words present unique cognitive challenges because they afford no opportunity for within-text consolidation or incremental comprehension building (Nagy & Herman, 1985). Unlike repeated vocabulary that allows readers to refine understanding through multiple contextual encounters, single-appearing words require readers to construct meaning from a solitary instance, placing greater demands on prior knowledge and contextual inference skills.

#### 4.3. Statistical Procedures

##### 4.3.1. Research Question 1: Progression of Text Features Within Grades

To consider shifts in lexical complexity between terms and years, percentages of changes in TTR from first to second term were established as were percentages of changes from the second term of one grade to the first term of the subsequent grade. Examining TTR changes from first to second term reveals the degree to which texts become more lexically diverse as students advance through an academic year. Increasing TTR would indicate that curriculum designers assume students develop greater vocabulary sophistication within a single grade, whereas decreasing or stable TTR might suggest emphasis on consolidating familiar vocabulary or focus on complexity dimensions (e.g., syntactic) rather than lexical diversity.

Comparing the second term of one grade to first term of the subsequent grade established the transition between grade levels. A substantial TTR increase from one grade to

the next would indicate clear developmental stepping, whereas minimal change might suggest either conservative progression or reliance on nonlexical complexity factors.

To determine whether token distributions by frequency level differed from the first to second half within each grade, we conducted chi-square ( $\chi^2$ ) tests for each combination of year, grade, and term. For instance, the analysis compared percentages of complex tokens between the first and second half of grade 1 in the 1957 copyright sample. Statistical significance was set at  $\alpha = 0.05$  with Bonferroni correction for multiple comparisons and phi ( $\varphi$ ) for effect sizes.

To determine whether token distributions by frequency level differed across years within each grade half, we conducted  $\chi^2$  tests by grade, half, and year. For the first half of grade 1, for example, we compared complex token percentages across all 4 years using Bonferroni correction for multiple comparisons and Cramer's  $V$  for effect sizes. We also conducted pairwise year comparisons for each grade and half combination (e.g., comparing complex token percentages in the first half of grade 1 between the 1957 and 1974 samples, 1957 and 1995, etc.) with Bonferroni correction and  $\varphi$  for effect sizes.

#### 4.3.2. Research Question 2: Nature of Text Features Across Grades and Years

To examine the progression of vocabulary demands from grade to grade within each year, we examined TTRs, percentages of complex tokens, and percentages of singletons for each year and grade combination.

We analyzed distributions of tokens by frequency level, by grades and years, with  $\chi^2$  tests of complex vocabulary (vs. frequent vocabulary) by year and grade. For example, percentages of complex tokens in 1957 texts across all four grades were analyzed. Statistical significance was set at  $\alpha = 0.05$ , and we used Bonferroni correction for multiple comparisons and Cramer's  $V$  for effect sizes. We also conducted pairwise comparisons of grades by year. For example, we compared the percentages of complex tokens in the 1957 sample for grade 1 versus grade 2, grade 1 versus grade 3, and so on. We applied Bonferroni correction again for multiple comparisons and  $\varphi$  for effect sizes.

We conducted similar analyses to determine the nature of percentages of singletons by grade for each year, using  $\chi^2$  tests to examine differences across the sets of four grades across each year. We applied Bonferroni correction for multiple comparisons and Cramer's  $V$  for effect sizes. Again, we made pairwise comparisons of grades within each year, using Bonferroni correction and  $\varphi$  for effect sizes.

We conducted analyses for RQ1 and RQ2 in *R* (Version 4.5.1) using the *tidyverse* package for data manipulation and the *ggplot2* package to generate plots. Coding assistance was provided by Claude Opus 4 ([Anthropic, 2025](#)).

## 5. Results

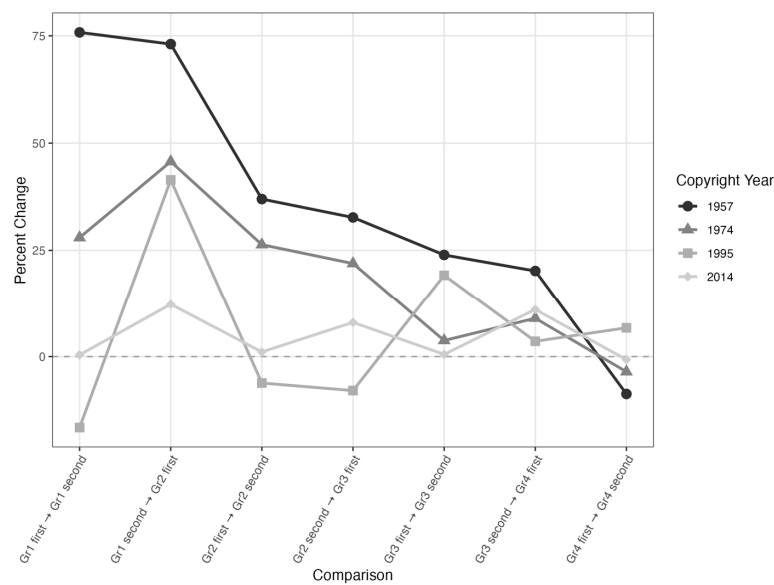
**RQ1.** How does the complexity of texts vary from the first to second term of a grade in type-token and complex vocabulary? Does this variation differ across grades and years?

### 5.1. Type-Token Ratio Within and Between Grades

Data on TTRs are presented in Table 1. The patterns of the percentages of change on this metric from the first to second terms of a grade and year are depicted in Figure 1. Changes in TTRs across grade halves were dissimilar across years, especially for grades 1 and 2. The samples from 1957 and 1974 showed substantial increases in TTR in grades 1 and 2 with increases steadily but modestly declining until the second half of grade 4. The changes in ratios for the 1995 sample were erratic, vacillating between decreases and increases of varying magnitudes. The changes in the ratios from the 2014 sample were more predictable with essentially no within-grade changes and modest increases between grades.

**Table 1.** Changes in type-token ratios across halves of each grade by year.

Copyright Year	Grade/Half	Type-Token Ratio	Comparison Grade/Half	Comparison Type-Token Ratio	Percent Change
1957	Gr1 first	0.02	Gr1 second	0.04	75.79
	Gr1 second	0.04	Gr2 first	0.06	73.07
	Gr2 first	0.06	Gr2 second	0.09	36.96
	Gr2 second	0.09	Gr3 first	0.12	32.68
	Gr3 first	0.12	Gr3 second	0.14	23.94
	Gr3 second	0.14	Gr4 first	0.17	20.19
	Gr4 first	0.17	Gr4 second	0.16	−8.74
1974	Gr1 first	0.06	Gr1 second	0.07	27.98
	Gr1 second	0.07	Gr2 first	0.11	45.70
	Gr2 first	0.11	Gr2 second	0.14	26.34
	Gr2 second	0.14	Gr3 first	0.17	21.93
	Gr3 first	0.17	Gr3 second	0.17	3.76
	Gr3 second	0.17	Gr4 first	0.19	8.98
	Gr4 first	0.19	Gr4 second	0.18	−3.55
1995	Gr1 first	0.17	Gr1 second	0.14	−16.59
	Gr1 second	0.14	Gr2 first	0.20	41.43
	Gr2 first	0.20	Gr2 second	0.18	−6.19
	Gr2 second	0.18	Gr3 first	0.17	−7.95
	Gr3 first	0.17	Gr3 second	0.20	19.16
	Gr3 second	0.20	Gr4 first	0.21	3.57
	Gr4 first	0.21	Gr4 second	0.22	6.73
2014	Gr1 first	0.16	Gr1 second	0.17	0.40
	Gr1 second	0.17	Gr2 first	0.19	12.27
	Gr2 first	0.19	Gr2 second	0.19	1.08
	Gr2 second	0.19	Gr3 first	0.20	8.00
	Gr3 first	0.20	Gr3 second	0.20	0.51
	Gr3 second	0.20	Gr4 first	0.23	11.05
	Gr4 first	0.23	Gr4 second	0.22	−0.72



**Figure 1.** Difference in TTR by Terms Within Grade and Between Grades.

5.2. Complex Vocabulary Within and Between Grades

We conducted two analyses to determine the progression of complex tokens between terms: (a) between terms of a grade for a year and (b) across grades of a year by term.

### 5.2.1. Between Terms of a Grade for a Year

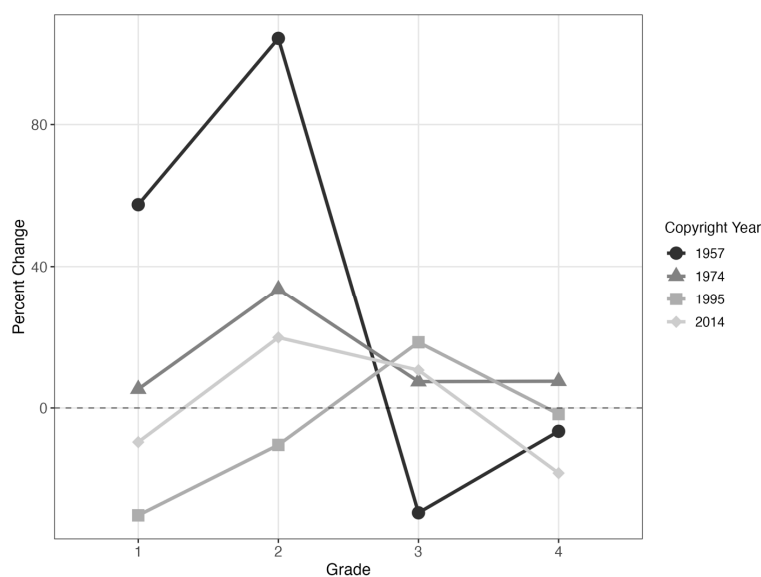
For each grade, the percentage of change in complex tokens from the first to second term of a year is presented in Table 2, as are goodness-of-fit tests. The 1957 texts showed the most differentiation from the first to second terms with all but the fourth-grade comparison significantly. For subsequent years, differences were erratic and showed inconsistent patterns with effects for all comparisons either negligible or nonexistent.

The patterns for percentages of differences in complex words by grade and year in Figure 2 showed a pattern similar to that for TTR. The texts for 1957 showed the most variation from grade to grade. Differences in the 1974 texts showed a somewhat similar pattern to that of 1957, although the size of the percentage differences were substantially smaller.

**Table 2.** Changes in percentages of complex tokens across halves of each grade by year.

Copyright Year	Grade	$\chi^2$	$p$	Adjusted $p$	$\varphi$	1st Half (%)	2nd Half (%)	Change (%)
1957	1	16.1	<0.001 ***	<0.001 ***	0.03	1.27	2.00	57.45
	2	81.01	<0.001 ***	<0.001 ***	0.06	2.19	4.48	104.24
	3	31.73	<0.001 ***	<0.001 ***	0.04	5.95	4.20	−29.49
	4	1.36	ns	ns	0.01	6.05	5.66	−6.57
1974	1	0.54	ns	ns	0.01	4.16	4.38	5.30
	2	25.91	<0.001 ***	<0.001 ***	0.04	5.08	6.79	33.60
	3	1.62	ns	ns	0.01	5.99	6.43	7.41
	4	1.61	ns	ns	0.01	5.83	6.27	7.52
1995	1	58.31	<0.001 ***	<0.001 ***	0.05	9.97	6.96	−30.21
	2	4.73	0.030 *	ns	0.02	7.80	6.99	−10.41
	3	14.44	<0.001 ***	0.002 **	0.03	8.48	10.04	18.50
	4	0.13	ns	<0.001 ***	0	9.10	8.94	−1.72
2014	1	3.09	ns	ns	0.01	6.20	5.61	−9.61
	2	12.73	<0.001 ***	0.006 **	0.03	6.67	7.99	19.85
	3	3.42	ns	ns	0.01	6.14	6.79	10.64
	4	17.74	<0.001 ***	<0.001 ***	0.03	8.94	7.30	−18.34

**Notes:** ns = non-significant.  $df = 1$  for all comparisons. Interpretation of  $\varphi$  as effect size:  $0.1 \leq \varphi < 0.3$  (small). \*\*\* denotes  $p < 0.001$ ; \*\* denotes  $p < 0.01$ ; \* denotes  $p < 0.05$ .



**Figure 2.** Complex Tokens: Across Grade Halves.

The differences for 1995 and 2014 followed a similar direction with differences for grades 1 and 2 similar, as was the case with grades 3 and 4. However, the percentage of differences for the first and second terms in these years was considerably smaller than in 1957. The profile (Figure 2) does show an interesting pattern: By fourth grade for all years, differences between the two terms were inconsequential.

5.2.2. Across Grades of a Year by Term

Data for the percentage of change in numbers of complex words across the 4 years for a grade, by term, are presented in Table 3. Goodness-of-fit analyses showed significant differences for all comparisons for both first and second terms. Grade 1, first term exhibited the greatest progression, with complex tokens more than doubling between 1974 and 1995. Pairwise comparisons across all years for each grade half (Table 4) revealed statistically significant differences in 83% of cases, though only 15% yielded effect sizes exceeding negligible thresholds (all small).

The similarity of size of direction of the changes in percentages of complex words by grade and year was similar for the first and second terms, leading to inclusion of Figure 3 only for the first term. Figure 3 shows that the percentages of tokens that were complex words are similar in grades 1 and 2 for 1957 and 1974. The percentage almost doubled for grade 1 from 1974 to 1995. Patterns of interest include the similarity in percentages of complex words in third grade to fourth grade in the earlier 2 years. The percentages were also similar in the later 2 years for third grade, although the absolute number of complex words increased substantially.

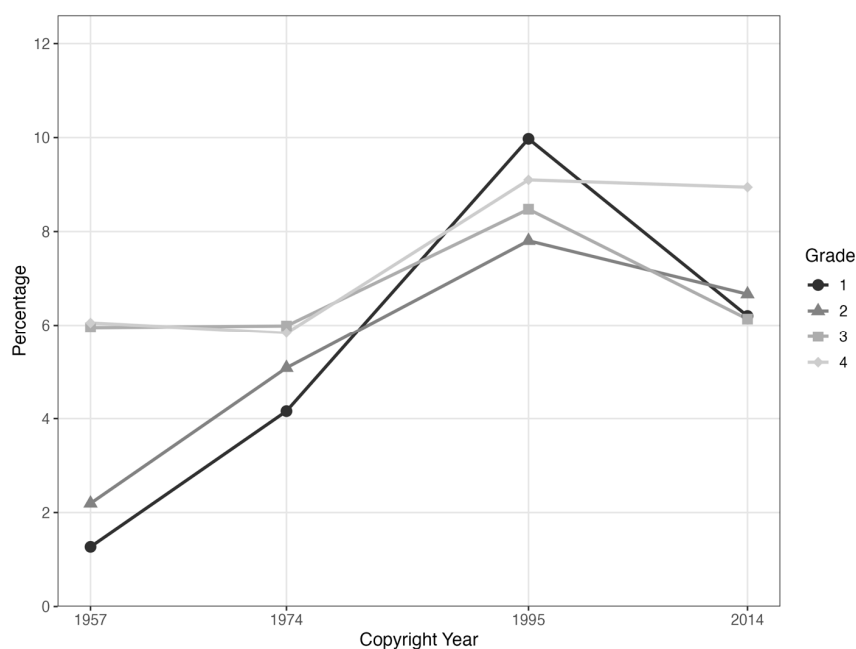


Figure 3. Complex tokens: First half by grade by year.

Table 3. Changes in percentages of complex tokens across years by halves of each grade.

Grade	Half	$\chi^2$	<i>p</i>	Adjusted <i>p</i>	<i>V</i>	1957 (%)	1974 (%)	1997 (%)	2014 (%)
1	First	654.14	<0.001 ***	<0.001 ***	0.14	1.27	4.16	9.97	6.20
2		204.58	<0.001 ***	<0.001 ***	0.09	2.19	5.08	7.80	6.67
3		60.85	<0.001 ***	<0.001 ***	0.04	5.95	5.99	8.48	6.14
4		44.64	<0.001 ***	<0.001 ***	0.06	6.05	5.83	9.10	8.94

Table 3. Cont.

Grade	Half	$\chi^2$	$p$	Adjusted $p$	$V$	1957 (%)	1974 (%)	1997 (%)	2014 (%)
1	Second	164.81	<0.001 ***	<0.001 ***	0.09	2.00	4.38	6.96	5.61
2		69.92	<0.001 ***	<0.001 ***	0.05	4.48	6.79	6.99	7.99
3		101.55	<0.001 ***	<0.001 ***	0.08	4.20	6.43	10.04	6.79
4		29.43	<0.001 ***	<0.001 ***	0.05	5.66	6.27	8.94	7.30

Notes.  $df = 3$  for all comparisons. Interpretation of Cramer’s  $V$  as effect size:  $0.06 \leq V < 0.17$  (small). \*\*\* denotes  $p < 0.001$ .

Table 4. Pairwise comparisons of complex token percentages in years by halves of grades.

Grade	Half	Comparison	$\chi^2$	$p$	Adjusted $p$	$\phi$
1	First	1957 vs. 1974	156.55	<0.001 ***	<0.001 ***	0.09
		1957 vs. 1995	712.52	<0.001 ***	<0.001 ***	0.19
		1957 vs. 2014	336.44	<0.001 ***	<0.001 ***	0.13
		1974 vs. 1995	256.72	<0.001 ***	<0.001 ***	0.11
		1974 vs. 2014	42.04	<0.001 ***	<0.001 ***	0.05
		1995 vs. 2014	95.21	<0.001 ***	<0.001 ***	0.07
2	First	1957 vs. 1974	120.18	<0.001 ***	<0.001 ***	0.08
		1957 vs. 1995	336.73	<0.001 ***	<0.001 ***	0.13
		1957 vs. 2014	239.44	<0.001 ***	<0.001 ***	0.11
		1974 vs. 1995	61.5	<0.001 ***	<0.001 ***	0.06
		1974 vs. 2014	22.59	<0.001 ***	<0.001 ***	0.03
		1995 vs. 2014	9.49	0.002 **	0.012 *	0.02
3	First	1957 vs. 1974	0.01	0.032 *	ns	0.00
		1957 vs. 1995	47.11	<0.001 ***	<0.001 ***	0.05
		1957 vs. 2014	0.27	ns	ns	0.00
		1974 vs. 1995	45.67	<0.001 ***	<0.001 ***	0.05
		1974 vs. 2014	0.17	ns	ns	0.00
		1995 vs. 2014	40.05	<0.001 ***	<0.001 ***	0.04
4	First	1957 vs. 1974	0.39	ns	ns	0.00
		1957 vs. 1995	65.68	<0.001 ***	<0.001 ***	0.06
		1957 vs. 2014	60.44	<0.001 ***	<0.001 ***	0.05
		1974 vs. 1995	76.13	<0.001 ***	<0.001 ***	0.06
		1974 vs. 2014	70.56	<0.001 ***	<0.001 ***	0.06
		1995 vs. 2014	0.13	ns	ns	0.00
1	Second	1957 vs. 1974	90.74	<0.001 ***	<0.001 ***	0.07
		1957 vs. 1995	286.13	<0.001 ***	<0.001 ***	0.12
		1957 vs. 2014	176.66	<0.001 ***	<0.001 ***	0.09
		1974 vs. 1995	61.69	<0.001 ***	<0.001 ***	0.06
		1974 vs. 2014	15.61	<0.001 ***	<0.001 ***	0.03
		1995 vs. 2014	15.3	<0.001 ***	<0.001 ***	0.03
2	Second	1957 vs. 1974	48.75	<0.001 ***	<0.001 ***	0.05
		1957 vs. 1995	56.95	<0.001 ***	<0.001 ***	0.05
		1957 vs. 2014	103	<0.001 ***	<0.001 ***	0.07
		1974 vs. 1995	0.29	ns	ns	0.00
		1974 vs. 2014	10.42	<0.001 ***	0.007 **	0.02
		1995 vs. 2014	7.09	0.008 **	0.047 *	0.02

Table 4. Cont.

Grade	Half	Comparison	$\chi^2$	<i>p</i>	Adjusted <i>p</i>	$\phi$
3	Second	1957 vs. 1974	49.43	<0.001 ***	<0.001 ***	0.05
		1957 vs. 1995	258.81	<0.001 ***	<0.001 ***	0.11
		1957 vs. 2014	64.52	<0.001 ***	<0.001 ***	0.06
		1974 vs. 1995	85.8	<0.001 ***	0.0048 **	0.07
		1974 vs. 2014	0.98	ns	ns	0.01
		1995 vs. 2014	68.21	<0.001 ***	<0.001 ***	0.06
4	Second	1957 vs. 1974	3.28	ns	ns	0.01
		1957 vs. 1995	80.16	<0.001 ***	<0.001 ***	0.06
		1957 vs. 2014	21.8	<0.001 ***	<0.001 ***	0.03
		1974 vs. 1995	51.25	<0.001 ***	<0.001 ***	0.05
		1974 vs. 2014	8.17	0.004 **	0.026 *	0.02
		1995 vs. 2014	17.77	<0.001 ***	<0.001 ***	0.03

Notes: *df* = 1 for all comparisons. Interpretation of  $\phi$  as effect size:  $0.1 \leq \phi < 0.3$  (small). ns = non-significant; \*\*\* denotes  $p < 0.001$ ; \*\* denotes  $p < 0.01$ ; \* denotes  $p < 0.05$ .

5.3. Summary of Findings for Research Question 1

Text complexity progression within grades exhibited markedly different patterns across the 4 years. Earlier years (1957, 1974) demonstrated systematic within-grade complexity increases, particularly in grades 1 and 2, suggesting deliberate scaffolding to support student growth during the academic year. In contrast, later years (1995, 2014) showed minimal within-grade variation, indicating a shift toward maintaining consistent complexity levels throughout each grade.

Another pattern was the disappearance of within-grade progression by fourth grade across all years. Additionally, whereas a dramatic complexity increase occurred between the 1974 and 1995 programs—with complex vocabulary more than doubling in first grade—the magnitude of within-grade changes remained negligible in recent decades despite statistical significance.

**RQ2.** What is the nature of differences in lexical density, percentage of complex words, and percentage of single-appearing words across the texts for grades 1–4 at four points in time: 1957, 1974, 1995, and 2014?

5.4. Type-Token Ratio Across Four Copyrights

TTRs for each year across grades 1–4, as well as the percent changes from grade to grade, are presented in Table 5 and depicted in Figure 4. In 1957, first graders encountered texts with highly repetitive vocabulary, then experienced a 140% jump in lexical diversity when moving to second grade. The climb in lexical diversity continues through the grades.

Table 5. Changes in percentages of type-token ratios across grades by year.

Copyright Year	Grade	Type-Token Ratio	Comparison Grade	Comparison Type-Token Ratio	Change (%)
1957	1	0.02	2	0.05	147.54
	2	0.05	3	0.09	81.35
	3	0.09	4	0.12	32.11
1974	1	0.04	2	0.09	109.81
	2	0.09	3	0.13	43.08
	3	0.13	4	0.14	13.62

Table 5. Cont.

Copyright Year	Grade	Type-Token Ratio	Comparison Grade	Comparison Type-Token Ratio	Change (%)
1997	1	0.12	2	0.15	25.49
	2	0.15	3	0.15	−0.55
	3	0.15	4	0.17	18.52
2014	1	0.12	2	0.15	16.9
	2	0.15	3	0.16	9.81
	3	0.16	4	0.18	14.32

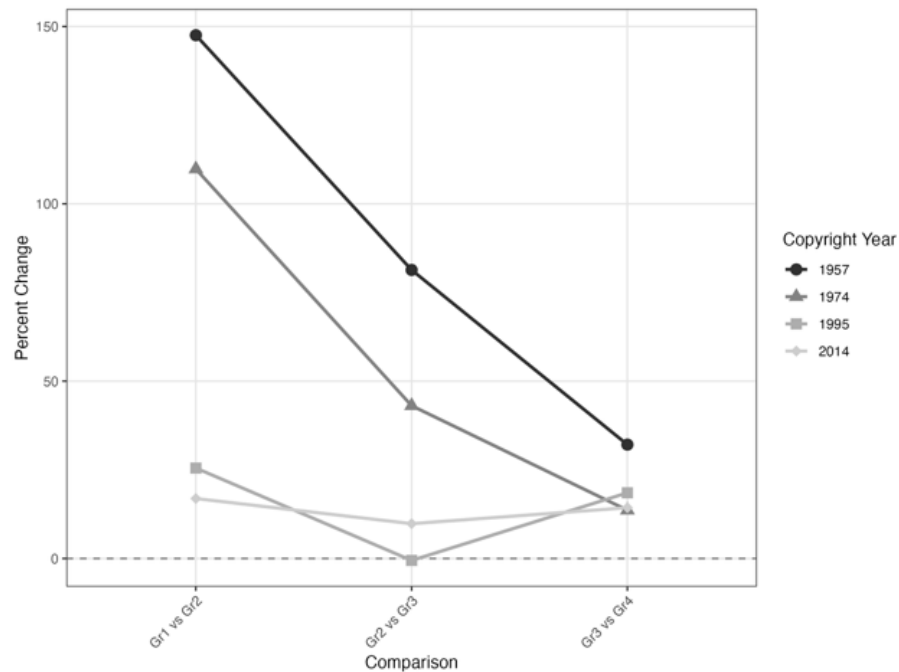


Figure 4. TTR differences by grade (Percentage change).

By 2014, this developmental trajectory had completely flattened. First graders started with the lexical diversity that once characterized much higher grades. The grade-to-grade increases dropped from greater than 100% to less than 10%—essentially eliminating vocabulary variety as a developmental progression tool.

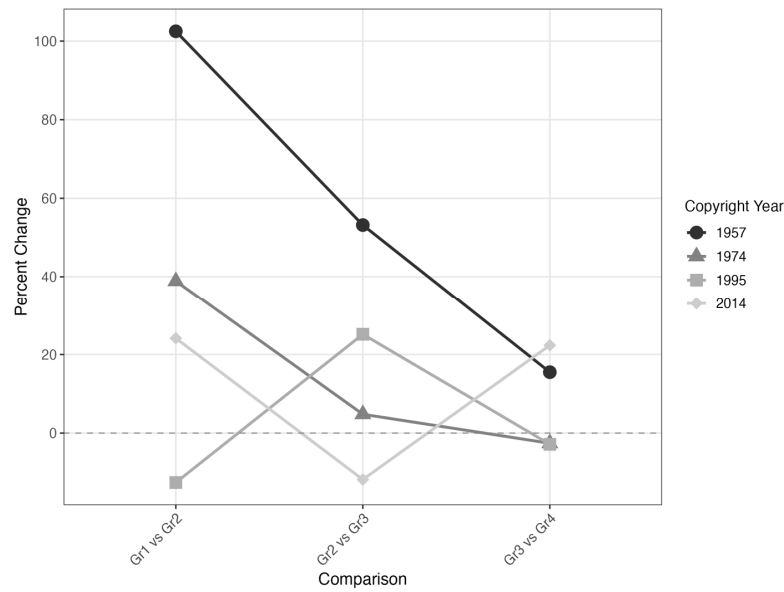
5.5. Complex Vocabulary Across Four Copyrights

As shown in Table 6 and Figure 5, 1957 began with a limited vocabulary but vocabulary complexity increased substantially (100%) by second grade. By 1974, this systematic progression weakened, showing a more moderate 40% increase from first to second grade with minimal growth in upper grades.

Table 6. Changes in percentages of complex tokens across grades by year.

Copy-Right Year	$\chi^2$	<i>p</i>	Adjusted <i>p</i>	<i>V</i>	Gr1 (%)	Gr2 (%)	Change (%)	Gr3 (%)	Change (%)	Gr4 (%)	Change (%)
1957	559.23	<0.001 ***	<0.001 ***	0.08	1.63	3.31	103.07	5.07	53.17	5.86	15.58
1974	92.54	<0.001 ***	<0.001 ***	0.03	4.27	5.93	38.88	6.21	4.72	6.05	−2.58
1997	52.39	<0.001 ***	<0.001 ***	0.03	8.46	7.40	−12.53	9.26	25.14	9.00	−2.81
2014	74.09	<0.001 ***	<0.001 ***	0.03	5.90	7.33	24.24	6.46	−11.87	7.91	22.45

Notes: *df* = 3 for all comparisons. Interpretation of Cramer’s *V* as effect size:  $0.06 \leq V < 0.17$  (small). \*\*\* denotes *p* < 0.001.



**Figure 5.** Percentage change in complex tokens by year and grade.

Starting with 1995, patterns of complex words became erratic, sometimes with negative changes between grades. The orderly developmental sequence disappeared, replaced by inconsistent patterns in which second graders might encounter fewer complex words than first graders.

Results of  $\chi^2$  goodness-of-fit tests for complex tokens across grades by year are shown in Table 7. Differences for 1957 and 1974 were all significant for pairwise comparisons, but effect sizes were negligible. All but one pair-wise comparison for the 1995 and 2014 programs was significant, but only one effect size had any heft—the one for 2014 grades 1–4. Importantly, grade 3–4 comparisons attained 0.12 in 1957 but, subsequently, was in the range 0.05–0.06. That is, the percentage of words that were complex was relatively the same at these two grade levels.

**Table 7.** Pairwise comparisons of percentages of complex tokens across grades by year.

Copyright Year	Comparison	$\chi^2$	<i>p</i>	Adjusted <i>p</i>	$\phi$
1957	Gr1 vs. Gr2	115.63	<0.001 ***	<0.001 ***	0.05
	Gr1 vs. Gr3	363.33	<0.001 ***	<0.001 ***	0.10
	Gr1 vs. Gr4	492.93	<0.001 ***	<0.001 ***	0.11
	Gr2 vs. Gr3	76.79	<0.001 ***	<0.001 ***	0.04
	Gr2 vs. Gr4	147.58	<0.001 ***	<0.001 ***	0.06
	Gr3 vs. Gr4	11.82	<0.001 ***	0.004 **	0.02
1974	Gr1 vs. Gr2	56.56	<0.001 ***	<0.001 ***	0.04
	Gr1 vs. Gr3	75.53	<0.001 ***	<0.001 ***	0.04
	Gr1 vs. Gr4	64.45	<0.001 ***	<0.001 ***	0.04
	Gr2 vs. Gr3	1.37	ns	ns	0.01
	Gr2 vs. Gr4	0.25	ns	ns	0.00
	Gr3 vs. Gr4	0.42	ns	ns	0.00
1997	Gr1 vs. Gr2	15.46	<0.001 ***	<0.001 ***	0.02
	Gr1 vs. Gr3	7.76	0.005 **	0.032 *	0.01
	Gr1 vs. Gr4	3.51	ns	ns	0.01
	Gr2 vs. Gr3	45.26	<0.001 ***	<0.001 ***	0.03
	Gr2 vs. Gr4	33.97	<0.001 ***	<0.001 ***	0.03
	Gr3 vs. Gr4	0.82	ns	ns	0.01

Table 7. Cont.

Copyright Year	Comparison	$\chi^2$	$p$	Adjusted $p$	$\phi$
2014	Gr1 vs. Gr2	32.63	<0.001 ***	<0.001 ***	0.03
	Gr1 vs. Gr3	5.28	0.022 *	ns	0.01
	Gr1 vs. Gr4	62.02	<0.001 ***	<0.001 ***	0.04
	Gr2 vs. Gr3	11.58	0.001 **	0.004 **	0.02
	Gr2 vs. Gr4	4.67	0.031 *	ns	0.01
	Gr3 vs. Gr4	31.09	<0.001 ***	<0.001 ***	0.03

Notes:  $df = 1$  for all comparisons. Interpretation of  $\phi$  as effect size:  $0.1 \leq \phi < 0.3$  (small). ns = non-significant; \*\*\* denotes  $p < 0.001$ ; \*\* denotes  $p < 0.01$ ; \* denotes  $p < 0.05$ .

5.6. Singletons Across Four Copyrights

As shown in Table 8 and Figure 6, there were modest increases in percentages of singletons from grade 1 to grade 2 for the 1995 and 2014 samples but they became substantial in 1957 and 1974. The same is true for the transition from grade 2 to grade 3, except for a slight decrease in percentage for 1995. From grade 3 to grade 4, samples from 1974, 1995, and 2014 showed similar increases of less than 14%, whereas the increase for 1957 was nearly 40%.

Table 8. Changes in percentages of single-appearing words Across grades by year.

Copyright Year	$\chi^2$	$p$	Adjusted $p$	$V$	Gr1 (%)	Gr2 (%)	Change (%)	Gr3 (%)	Change (%)	Gr4 (%)	Change (%)
1957	254.25	<0.001 ***	<0.001 ***	0.21	13.19	19.19	45.49	30.21	57.43	41.67	37.93
1974	278.65	<0.001 ***	<0.001 ***	0.19	19.55	33.41	70.90	43.07	28.91	48.73	13.14
1997	65.71	<0.001 ***	<0.001 ***	0.07	40.81	46.04	12.82	45.34	-1.52	51.37	13.30
2014	87.7	<0.001 ***	<0.001 ***	0.08	41.13	44.59	8.41	47.52	6.57	52.61	10.71

Notes:  $df = 3$  for all comparisons. Interpretation of Cramer’s  $V$  as effect size:  $0.06 \leq V < 0.17$  (small);  $0.17 < V < 0.29$  (medium). \*\*\* denotes  $p < 0.001$ .

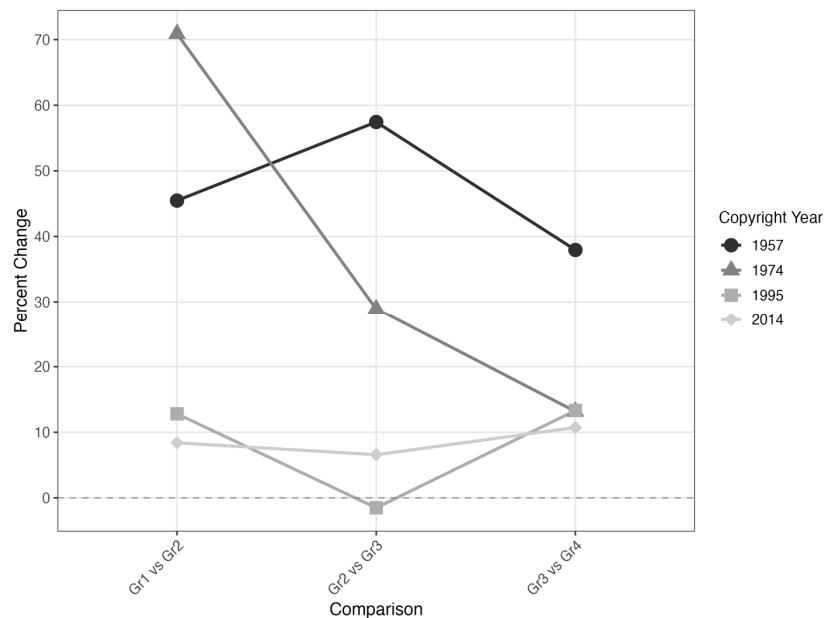


Figure 6. Percentage Change in Singletons By Year & Grade.

Table 8 also shows the results of  $\chi^2$  goodness-of-fit testing of percentages of singletons across grades by year. All comparisons were significant with medium effects for the 1957 and 1974 comparisons and small effects for the 1995 and 2014 comparisons. More than 80% of pairwise comparisons were significant, as presented in Table 9, with small effect sizes

for half of comparisons (including nonsignificant comparisons) and negligible effects in other comparisons.

**Table 9.** Pairwise comparisons of single-appearing words (%) across grades by year.

Copyright Year	Comparison	$\chi^2$	$p$	Adjusted $p$	$\phi$
1957	Gr1 vs. Gr2	6.86	0.009 **	ns	0.07
	Gr1 vs. Gr3	48.77	<0.001 ***	<0.001 ***	0.15
	Gr1 vs. Gr4	121.69	<0.001 ***	<0.001 ***	0.21
	Gr2 vs. Gr3	41.04	<0.001 ***	<0.001 ***	0.12
	Gr2 vs. Gr4	160.38	<0.001 ***	<0.001 ***	0.21
	Gr3 vs. Gr4	59.75	<0.001 ***	<0.001 ***	0.12
1974	Gr1 vs. Gr2	52.62	<0.001 ***	<0.001 ***	0.14
	Gr1 vs. Gr3	148.64	<0.001 ***	<0.001 ***	0.21
	Gr1 vs. Gr4	226.6	<0.001 ***	<0.001 ***	0.25
	Gr2 vs. Gr3	40.55	<0.001 ***	<0.001 ***	0.10
	Gr2 vs. Gr4	104.91	<0.001 ***	<0.001 ***	0.15
	Gr3 vs. Gr4	17.23	<0.001 ***	<0.001 ***	0.06
1997	Gr1 vs. Gr2	14.29	<0.001 ***	0.001 **	0.05
	Gr1 vs. Gr3	10.65	0.001 **	0.007 **	0.05
	Gr1 vs. Gr4	62.41	<0.001 ***	<0.001 ***	0.10
	Gr2 vs. Gr3	0.26	ns	ns	0.01
	Gr2 vs. Gr4	18.02	<0.001 ***	<0.001 ***	0.05
	Gr3 vs. Gr4	22.95	<0.001 ***	<0.001 ***	0.06
2014	Gr1 vs. Gr2	6.38	0.012 *	ns	0.03
	Gr1 vs. Gr3	22.79	<0.001 ***	<0.001 ***	0.06
	Gr1 vs. Gr4	77.64	<0.001 ***	<0.001 ***	0.11
	Gr2 vs. Gr3	5.16	0.023 *	ns	0.03
	Gr2 vs. Gr4	41.51	<0.001 ***	<0.001 ***	0.08
	Gr3 vs. Gr4	17.52	<0.001 ***	<0.001 ***	0.05

**Notes:**  $df = 1$  for all comparisons. Interpretation of  $\phi$  as effect size:  $0.1 \leq \phi < 0.3$  (small). ns = nonsignificant; \*\*\* denotes  $p < 0.001$ ; \*\* denotes  $p < 0.01$ ; \* denotes  $p < 0.05$ .

### 5.7. Summary of Findings for Research Question 2

A transformation in developmental complexity occurred across the 4 years. The 1957 program exhibited steep, systematic progression across all complexity measures: lexical diversity increased 140% from first to second grade, complex vocabulary doubled, and singletons showed substantial grade-to-grade increases (nearly 40% from third to fourth grade).

This developmental trajectory progressively flattened over subsequent decades. By 2014, grade-to-grade increases in lexical diversity had collapsed from greater than 100% to less than 10%, essentially eliminating vocabulary variety as a developmental tool. Most critically, first graders in 2014 encountered the lexical diversity that once characterized much higher grades.

Patterns reveal three major shifts in reading textbooks for students from grades 1–4 across a 60-year period: a lessening of developmental shifts within a grade level; a decrease in across-grade complexity from grades 1–4; and increased expectations for what is defined as grade-level reading, most specifically of first grade.

## 6. Discussion

Analyses of text complexity patterns reveal three major shifts in reading textbooks for students from grades 1–4 across a 60-year period: a lessening of developmental shifts within a grade level; a decrease in across-grade complexity from grades 1–4; and increased expectations for what is defined as grade-level reading, most specifically of first grade.

### 6.1. Changes Within the Texts for a School Year

At the outset of this 60-year period, reading instructional materials exhibited a systematically scaffolded trajectory of increasing lexical complexity across the academic year. The 1957 first-grade curriculum exemplified this developmental architecture most distinctly: New vocabulary density doubled between first and second terms, whereas complex words—entirely absent from first-term materials—were introduced in second-term texts. This pattern of incremental intensification persisted through third grade, with fourth grade representing a transition point where lexical density stabilized across terms.

By 1974, this carefully calibrated progression had begun to erode, yielding inconsistent developmental patterns across grade levels. Whereas first and second grades maintained increased lexical demands in the second term, this within-year progression disappeared in grades 3 and 4. The 1995 program extended this flattening downward: Lexical demands remained relatively constant across both terms in all four grade levels. The 2014 program sustained this pattern, consolidating what had become the new pedagogical norm.

The shift from graduated to uniform difficulty levels throughout the school year represents a fundamental reconceptualization of reading development. The earlier graduated approach presumed students required incremental increases in difficulty, with each term building systematically on prior knowledge. The current flat progression suggests an assumption that students can engage with comparable lexical complexity throughout the year.

### 6.2. Changes in Text Complexity Across Grades Within a Year

The movement to similarity in the texts of the first and second terms of a grade is replicated in the complexity progression across grade levels within each copyright. The 1957 textbooks show the strongest grade-level differences with TTRs increasing fivefold from first to fourth grade and single-appearing words more than tripling. The 1974 materials continue this strong developmental approach, although differences are somewhat smaller than for the 1957 edition. However, TTRs still climb steadily across grades, and singleton percentages nearly double from first to fourth grade.

A significant turning point occurred with the 1995 materials. There is an increase in TTR from grade 1 to grade 2, but increases are relatively negligible in subsequent grades. For percentages of complex words, 1995 shows an erratic pattern in which second-grade texts contain fewer complex words than grade 1, breaking the expected developmental sequence. This flattening trend continued in the 2014 materials, with TTRs showing only modest differences across the 4-year span and percentages of complex words and single-appearing words displaying irregular, nonsequential patterns.

These findings reveal a dramatic historical shift in reading curriculum design: Older textbooks (1957, 1974) followed developmental progressions over the grades in which vocabulary complexity increased substantially across grade levels, but recent textbooks (1995 onward) have shown a flat progression with first- through fourth-grade texts becoming increasingly similar in complexity.

### 6.3. Grade-Level Distinctions over the Period

As striking as the shifts within the grade-level material across the texts of a specific copyright are the changes in grade-level expectations over this time period. Grade 1 has undergone the most radical changes in any level, with TTRs increasing 500% from 1957 to 2014 (0.02 to 0.12), complex words climbing from 1.64% in 1957 to 5.90% by 2014, and single-appearing words tripling from 19.2% to 41.0% over the decades. The texts given to first graders in 2014 bore little resemblance to their 1957 predecessors. Indeed, for the three measures of text complexity—TTR, percentage of complex words, and singletons—the texts

given to first graders in 2014 had almost precisely the same features as the texts given to fourth graders in 1957. Changes were also evident in second-grade texts but less dramatic than grade 1. Grade 3's changes were more modest, whereas Grade 4 showed the smallest changes in all.

A comment is also appropriate regarding the interpretations of text complexity over time. Methodological advances in corpus sampling have fundamentally altered our understanding of TTR. The inadequacies of earlier approaches become particularly evident when sample size is uncontrolled or unspecified. [Chall's \(1967\)](#) influential analysis, which failed to report sample sizes for grades 2 and 3, substantially underestimated lexical diversity in these texts. Where Chall identified merely 1.35 new words per 100 running words in third-grade materials, a systematic analysis of 20,000 contiguous words from the complete 1957 program revealed nine new words per 100—a nearly sevenfold discrepancy. This pattern of underestimation in partial-sample analyses extends beyond [Chall's work](#). [Gamson et al. \(2013\)](#) demonstrated that [Hayes et al.'s \(1996\)](#) findings similarly underestimated text complexity, a distortion directly attributable to their incomplete sampling methodology. These substantial disparities underscore a critical methodological principle: Valid TTR assessment requires both adequate sample size and transparent reporting of sampling procedures.

#### *6.4. Implications of the Findings*

The overall picture reveals a fundamental shift in the conceptualization of elementary literacy development: an elevation of the starting point and a more compressed and less differentiated learning experience. The findings lead to one observation and two questions. The observation: Third- and fourth-grade texts show remarkable stability over time with not much difference between the two grades and fairly stable across six decades. The first question is whether the dramatic acceleration in text complexity at first grade reflects genuine increases in children's cognitive capacity or merely elevates expectations imposed on young learners. The second question is whether these systematic shifts in textbooks have improved student reading achievement or inadvertently created new barriers to literacy development.

##### *6.4.1. The Stability of Third- to Fourth-Grade Benchmarks*

Contrary to claims that school reading texts have trended downward in difficulty ([NGA-CBP & CCSSO, 2010](#)), the evidence shows that text complexity in grades 3 and 4 has remained remarkably consistent across the entire 60-year period. Despite substantial shifts in reading instruction philosophy—from controlled vocabulary to whole language and then to systematic phonics—texts used in grades 3 and 4 show consistent patterns across lexical density, percentage of complex words, and frequency of single-appearing words over time.

This stability reflects something fundamental about the structure of elementary-level English prose rather than pedagogical choices. English follows predictable word frequency distributions ([Thorndike, 1927](#); [Zipf, 1935](#)). Once students have acquired approximately the 5000 most frequent words, any text written in ordinary English will naturally exhibit a consistent ratio of high- to low-frequency words, typically a ratio of approximately 70% high-frequency to 30% content-carrying words ([Zipf, 1935](#)). Authors writing about different topics for this age group (whether community helpers in 1957 or ecosystems in 2014) draw from the same basic vocabulary structure. In other words, the reading wars are fought primarily over the path to fundamental proficiency in reading English text, not over the linguistic characteristics of the destination.

The prevailing policy focus on third-grade interventions appears fundamentally misdirected. Evidence indicates that the most significant cognitive and developmental leap in

reading expectations has occurred at first grade, not third grade. The oft-cited distinction between “learning to read” until third grade and “reading to learn” beginning in fourth grade (Chall, 1983) proves artificial in that these grade levels demonstrate remarkably similar complexity profiles. Even in earlier educational eras, the distinction between third and fourth grades was far less substantial than that between grades 1 and 3. Intervention and retention policies that trigger at third or fourth grade not only arrive too late in the developmental sequence but also target a transition point that lacks the pedagogical significance assumed by policymakers.

#### 6.4.2. The Acceleration of Curriculum to Kindergarten and the Speed of First Grade

Although third- and fourth-grade texts remained relatively stable across years, first- and second-grade materials varied substantially, serving as direct artifacts of each era’s dominant instructional philosophy. Mid-century basal readers operated on strict vocabulary controls, producing texts with relatively low lexical density and few single-appearing words through engineered high rates of word repetition. The whole language movement rejected these constraints (Goodman, 1986), producing first-grade texts with lexical density nearly indistinguishable from texts designed for higher grades. Contemporary phonics-based programs employ the LTM system (Stein et al., 1999) focusing on a particular model of decodability rather than word repetition, which permits high lexical density because it focuses on orthographic connections to lessons rather than repeated exposure.

With the increasing demands on first grade has come a concurrent shift in reading instruction to kindergarten that has been described as the new grade one (Bassok et al., 2016). Before NCLB, kindergarten materials in core reading programs consisted of readiness workbooks or big books for teacher read-alouds (Hiebert & Papierz, 1990). After NCLB, core reading programs began including texts for kindergarten reading instruction. There were no kindergarten reading texts during the first 3 years (1957, 1974, and 1993) of the program and so were not analyzed. But indications are that the texts for kindergarten in 2014 exceed those of the earlier 1957 and 1974 grade 1 programs (Hiebert, 2015).

The central question is whether children’s capacity has kept pace with these elevated demands. As a result of shifts in kindergarten, children entering first grade in the 2010s showed shifts in literacy preparedness (D’Agostino & Rodgers, 2017). Low-income students showed higher performance on foundational skills (i.e., phonemic awareness, letter name recognition). However, middle- and upper-income students entered with stronger word recognition abilities, increasing the achievement gap between affluent and low-income students.

DIBELS word recognition fluency data (University of Oregon, 2022) demonstrates the widening gap. Whereas a small minority enters kindergarten already reading (95th percentile: 22 words), most children begin at or near zero. By kindergarten’s end, the spread ranges from 2.5 to 19 words across the 25–75th percentiles. This stratification intensifies through first grade (10–33 words) and second grade (22.5–62 words). Students at the 75th percentile ultimately perform nearly three times better than those at the 25th percentile, whereas those at the 55th percentile (47 words at end of second grade) more than doubles those at the lowest quartile’s performance, demonstrating how early differences in fluency compound into substantial disparities in reading automaticity. Three years of kindergarten through second-grade instruction leave substantial proportions of students without sufficient word recognition automaticity. Critically, the increased text complexity demands appear to be calibrated to the word recognition skills higher socioeconomic students possess, potentially making academic tasks more challenging for students who depend on school-based instruction for their literacy development.

#### 6.4.3. Effects of Shifts in Textbook Perspectives on Students' Reading Achievement

A striking feature of recurring transformations in reading instruction is how rarely these shifts have been grounded in empirical evidence of their effects on student achievement, particularly for struggling readers. Each major pedagogical movement has been implemented at scale with limited or absent data demonstrating that proposed changes would improve outcomes for the full range of learners.

Tuinman et al. (1976) observed that controlled vocabulary texts dominant through the 1960s had produced satisfactory reading achievement. By available measures of that time, students were reading well. Chall (1967) criticized the texts for their low lexical density, a pattern that was underestimated at grades 2 and 3, as is evident in the current data, but never examined how students performed with these texts.

The whole language movement represented an even more dramatic departure from evidence-based practice. This transformation occurred without grounding research demonstrating how students, especially those in the bottom third, would respond to texts with no control of vocabulary. The theoretical foundation rested on observations of skilled reading and assumptions about how literacy naturally develops, not on experimental evidence showing that struggling first graders could successfully learn using high lexical-density texts.

The current LTTM model represents another fundamental assumption implemented without supporting evidence: that once a letter–sound correspondence has been taught in a lesson, students can successfully read any word containing that pattern. This “once taught, then learned” premise drives the high lexical density in contemporary first-grade texts. Again, this system was implemented without supportive research. Even after a decade of rhetoric related to the science of reading, research has not been conducted to demonstrate that students, particularly those in the lowest quartile, can develop automatic word recognition when encountering minimal repetitions of decodable words, or that all taught patterns are equivalently accessible regardless of vowel complexity, word frequency, or phonological load. Each transformation has been driven by theoretical commitments, critiques of previous approaches, and assumptions about reading acquisition—rarely by evidence demonstrating improved outcomes, especially for students in the lower half of the achievement distribution.

#### 6.5. *Toward More Thoughtful Text Design*

Although debates typically focus on phonics versus whole language approaches, the fundamental architecture of vocabulary progression itself requires serious reconsideration. The answer is not a return to the contrived texts that prevailed from the 1950s through the 1970s. In an era of artificial intelligence, and as literature on reading acquisition processes has increased voluminously in the past 60 years, there seems no reason why more carefully calibrated reading materials that reflect the learning processes of children, and the quasi-orthographic nature of English, cannot be created for students whose primary reading acquisition experiences occur in school.

Large language models and corpus linguistics research (Litman, 2016; Wang et al., 2024) combined with decades of empirical findings on reading acquisition (Snowling et al., 2022) offer unprecedented opportunities to create pedagogically optimized texts for beginning readers that address cognitive load more strategically than previous approaches. Unlike the creators of mid-20th-century reading programs such as Gray and Gates, contemporary educators have access to sophisticated data on the effects of numerous word features, including but not limited to word frequency, age of acquisition, and semantic concreteness on students' word recognition (Gao et al., 2022). By leveraging this evidence, educators can move beyond Gates and Russell's (1938) uniform repetition requirements

to design texts that strategically distribute cognitive load, providing appropriate levels of repetition for complex orthographic and less common orthographic patterns while allowing concrete, regular words to appear less frequently, thereby creating more engaging and developmentally appropriate reading experiences.

### 6.6. Limitations

It could be argued that this analysis fails to capture the authentic reading experiences of elementary students in that contemporary classrooms offer a plethora of reading materials (Center for Education Market Dynamics, 2025) and that evaluating the demands of students at the present time on the text complexity of anthologies of core reading programs fails to capture the full range of reading opportunities available to children. They might point to specialized sets of texts, such as decodable, content-area readers, and in some cases, leveled texts that accompany core reading programs. Further, students can be expected to encounter a variety of text types in their classrooms beyond the texts used for reading instruction. The widespread use of the Accelerated Reader program (Tischner et al., 2023) alone demonstrates the presence of trade books in educational settings. However, the argument that all text types were not included in this analysis overlooks two critical issues.

First, assessment frameworks like DIBELS follow identical complexity trajectories and embrace the same underlying philosophy, as is evident in the texts of the 2014 program. The percentage of complex words in the texts of both the assessments and the core reading programs are similar—approximately five to six words (Toyama et al., 2017).

Second, in analyses of decodable texts—which emphasize phonetic patterns and sound–letter relationships—and leveled texts—which are organized by perceived difficulty and reading level—the proportion of complex words found across both text types remained remarkably similar (Hiebert, 2024). First-grade students encounter comparable levels of vocabulary complexity across different text types and instructional philosophies. This consistency suggests that the present anthology-based analysis may be more representative of typical first-grade reading experiences than critics contend. The fundamental challenge of complex word exposure appears largely independent of the pedagogical framework used to organize and present texts, indicating that vocabulary demands remain stable across varying approaches to early literacy instruction.

### 6.7. Conclusions

The developmental staircase for elementary reading has indeed flattened but not in the ways commonly assumed. This analysis of lexical complexity progression across six decades reveals a fundamental transformation: The graduated steps that once characterized reading development have been replaced by an escalator that rises steeply at the start and then levels off. Contemporary first graders encounter the lexical complexity that characterized fourth-grade texts in 1957, yet the progression from first through fourth grade has compressed dramatically with grade-to-grade increases in lexical diversity declining from greater than 100% to less than 10%.

Rather than viewing these findings as an indictment of current practice, they represent an opportunity to synthesize the strengths of different eras. The data challenge two prevailing narratives: third- and fourth-grade texts have not declined in complexity but have remained remarkably stable across all six decades, whereas the primary locus of change has been first grade, not the upper elementary years targeted by most intervention policies.

The path forward lies in combining approaches: honoring students' capacity for early challenges while maintaining meaningful differentiation across grades. This shift reflects more than changing instructional philosophies—it represents a reconceptualization of literacy development itself. Students who successfully navigate first grade's accelerated

demands currently face limited vocabulary growth through fourth grade, whereas those who struggle initially encounter few opportunities to catch up through increasingly complex texts. Future text design should ensure that elementary years provide sustained intellectual growth with materials that challenge all learners while creating systematic pathways for those who depend on school-based instruction for literacy development.

**Funding:** This research received no external funding.

**Institutional Review Board Statement:** Not applicable.

**Informed Consent Statement:** Not applicable.

**Data Availability Statement:** The data supporting this study consist of copyrighted textbook excerpts that cannot be made publicly available due to copyright restrictions.

**Conflicts of Interest:** Author Elfrieda H. Hiebert is the founder, president and CEO of TextProject. The author declares that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

## References

- ACT, Inc. (2006). *Reading between the lines: What the ACT reveals about college readiness in reading*. ACT, Inc.
- Alvermann, D. E., & Guthrie, J. T. (1993). *Themes and directions of the National Reading Research Center* (Perspectives in Reading Research, No. 1). National Reading Research Center (NRRRC).
- 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*. National Academy of Education.
- Anthropic. (2025). *Claude 4 Opus [Large language model]*. Available online: <https://www.anthropic.com/claude> (accessed on 9 September 2025).
- Aukerman, R. C. (1984). *Approaches to beginning reading* (2nd ed.). John Wiley & Sons.
- Bassok, D., Latham, S., & Rorem, A. (2016). Is kindergarten the new first grade? *AERA Open*, 2(1), 2332858415616358. [CrossRef]
- Beck, I. L., McKeown, M. G., Omanson, R. C., & Pople, M. T. (1984). Improving the comprehensibility of stories: The effects of revisions that improve coherence. *Reading Research Quarterly*, 19(3), 263–277. [CrossRef]
- California English/Language Arts Committee. (1987). *English-language arts framework for California public schools (Kindergarten through grade twelve)*. CA Department of Education.
- California English/Language Arts Committee. (1999). *English-language arts framework for California public schools (Kindergarten through grade twelve)*. CA Department of Education.
- Campbell, J. R., Donahue, P. L., Reese, C. M., & Phillips, G. W. (1996). *NAEP 1994 reading report card for the nation and the states: Findings from the national assessment of educational progress and trial state assessment*. National Center for Education Statistics (NCES).
- Chall, J. S. (1967). *Learning to read: The great debate*. McGraw-Hill.
- Chall, J. S. (1983). *Stages of reading development*. McGraw-Hill.
- Chall, J. S., Conard, S. S., & Harris-Sharples, S. (1991). *Should textbooks challenge students? The case for easier or harder textbooks*. Teachers College Press.
- Chall, J. S., Jacobs, V. A., & Baldwin, L. E. (1990). *The reading crisis: Why poor children fall behind*. Harvard University Press.
- Chall, J. S., & Squire, J. R. (1996). The publishing industry and textbooks. In R. Barr, M. L. Kamil, P. B. Mosenthal, & P. D. Pearson (Eds.), *Handbook of reading research* (vol. 2, pp. 120–146). Routledge.
- D’Agostino, J. V., & Rodgers, E. (2017). Literacy achievement trends at entry to first grade. *Educational Researcher*, 46(2), 78–89. [CrossRef]
- Davison, A., & Kantor, R. N. (1982). On the failure of readability formulas to define readable texts: A case study from adaptations. *Reading Research Quarterly*, 17(2), 187–209. [CrossRef]
- Doan, S. Y., & Kaufman, J. (2024). What role do states play in selecting K-12 textbooks? A network of states move the needle on quality without usurping local control. *State Education Standard*, 24(1). Available online: <https://www.nasbe.org/what-role-do-states-play-in-selecting-k-12-textbooks/> (accessed on 8 November 2025).
- Dolch, E. W. (1928). Vocabulary burden. *The Journal of Educational Research*, 17(3), 170–183. [CrossRef]
- Elson, W. H., & Gray, W. S. (1931). *Elson-gray basic readers*. Scott Foresman.
- Fitzgerald, J., Elmore, J., Relyea, J. E., Hiebert, E. H., & Stenner, A. J. (2016). Has first-grade core reading program text complexity changed across six decades? *Reading Research Quarterly*, 51(1), 7–28.

- Foorman, B. R., Francis, D. J., Davidson, K. C., Harm, M. W., & Griffin, J. (2004). Variability in text features in six grade 1 basal reading programs. *Scientific Studies of Reading*, 8(2), 167–197. [CrossRef]
- Gamson, D. A., Lu, X., & Eckert, S. A. (2013). Challenging the research base of the Common Core State Standards: A historical reanalysis of text complexity. *Educational Researcher*, 42(7), 381–391. [CrossRef]
- Gao, C., Shinkareva, S. V., & Desai, R. H. (2022). SCOPE: The South Carolina psycholinguistic metabase. *Behavior Research*, 55, 2853–2884. [CrossRef]
- Gates, A. I., & Huber, M. B. (1931). *The work-play books*. Macmillan.
- Gates, A. I., Huber, M. B., & Salisburg, F. S. (1957). *Tuffy and boots* (The Macmillan Readers). Macmillan.
- Gates, A. I., & Russell, D. H. (1938). Types of materials, vocabulary burden word analysis, and other factors in beginning reading. I. *The Elementary School Journal*, 39(1), 27–35. [CrossRef]
- Ginn. (1961). *The ginn basic readers*. Ginn.
- Goodman, K. S. (1986). *What's whole in whole language?* Heinemann.
- Graesser, A. C., McNamara, D. S., & Louwerse, M. M. (2003). What do readers need to learn in order to process coherence relations in narrative and expository text. In A. P. Sweet, & C. E. Snow (Eds.), *Rethinking reading comprehension* (pp. 82–98). Guilford.
- Gray, W. S., & Leary, B. E. (1935). *What makes a book readable*. University Chicago Press.
- Gray, W. S., Monroe, M., Artley, A. S., Arbuthnot, M. H., & Gray, L. (1956). *The new fun with dick and jane*. Scott Foresman.
- Hayes, D. P., Wolfer, L. T., & Wolfe, M. F. (1996). Schoolbook simplification and its relation to the decline in SAT-verbal scores. *American Educational Research Journal*, 33(2), 489–508. [CrossRef]
- Heitz, C. A. (1979). *Vocabulary load and control of first grade basal readers published in the late 1970s* [Doctoral dissertation, University of Iowa]. ProQuest Dissertations and Theses Global. Available online: <https://www.proquest.com/7924482> (accessed on 22 October 2025).
- Hiebert, E. H. (2005). State reform policies and the task textbooks pose for first-grade readers. *The Elementary School Journal*, 105(3), 245–266. [CrossRef]
- Hiebert, E. H. (2015). Changing readers, changing texts: Beginning reading texts from 1960 to 2010. *Journal of Education*, 195(3), 1–13. [CrossRef]
- Hiebert, E. H. (2024). Enhancing opportunities for decoding and knowledge building through beginning texts. *The Reading Teacher*, 77(6), 965–974. [CrossRef]
- Hiebert, E. H., Goodwin, A. P., & Cervetti, G. N. (2018). Core vocabulary: Its morphological content and presence in exemplar texts. *Reading Research Quarterly*, 53(1), 29–49. [CrossRef]
- Hiebert, E. H., & Papierz, J. M. (1990). The emergent literacy construct and kindergarten and readiness books of basal reading series. *Early Childhood Research Quarterly*, 5(3), 317–334. [CrossRef]
- Kearns, D. M., & Hiebert, E. H. (2022). The word complexity of primary-level texts: Differences between first and third grade in widely used curricula. *Reading Research Quarterly*, 57(1), 255–285. [CrossRef]
- Kurtz, H., Lloyd, S., Harwin, A., Chen, V., & Furuya, Y. (2020). *Early reading instruction: Results of a national survey of K-2 and elementary special education teachers and postsecondary instructors*. Editorial Projects in Education, EdWeek Research Center. Available online: <https://epe.brightspotcdn.com/1b/80/706eba6246599174b0199ac1f3b5/ed-week-reading-instruction-survey-report-final-1.24.20.pdf> (accessed on 8 November 2025).
- Lane, H. B., Contesse, V. A., Gage, N. A., & Burns, M. K. (2025). Effect of an instructional program in foundational reading skills on early literacy development of students in kindergarten and first grade. *Reading Research Quarterly*, 60(1), e607. [CrossRef]
- Levine, A. (1996). America's reading crisis: Why the whole language approach to teaching reading has failed millions of children. *Parents*, 16, 63–65, 68.
- Litman, D. J. (2016). Natural language processing for enhancing teaching and learning. *Proceedings of the AAAI Conference on Artificial Intelligence*, 30(1), 4170–4176. [CrossRef]
- Malvern, D., Richards, B., Chipere, N., & Durán, P. (2004). Comparing the diversity of lexical categories: The type-type ratio and related measures. In D. D. Malvern, B. J. Richards, N. Chipere, & P. Durán (Eds.), *Lexical diversity and language development: Quantification and assessment* (pp. 121–151). Palgrave Macmillan.
- Nagy, W. E., & Herman, P. A. (1985). Incidental vs. instructional approaches to increasing reading vocabulary. *Educational Perspectives*, 23(1), 16–21.
- National Governors Association Center for Best Practices (NGA-CBP) & Council of Chief State School Officers (CCSSO). (2010). *Common Core State Standards for English language arts & literacy in history/social studies, science, and technical subjects*. Available online: <https://learning.ccso.org/wp-content/uploads/2022/11/ADA-Compliant-ELA-Standards.pdf> (accessed on 8 November 2025).
- Office of Educational Research & Improvement, U.S. Department of Education, Center for Education Market Dynamics. (2025, July 23). *ELA decision trends: What district choices reveal about curriculum, coherence, and quality*. Available online: <https://www.cemd.org/ela-decision-trends-what-district-choices-reveal-about-curriculum-coherence-and-quality/> (accessed on 8 November 2025).

- Schwartz, S. (2019). The most popular reading programs aren't backed by science. *Education Week*, 39(15), 19–22. Available online: <https://www.edweek.org/teaching-learning/the-most-popular-reading-programs-arent-backed-by-science/2019/12> (accessed on 8 November 2025).
- Schwartz, S. (2023, May 11). 4 more states pass 'science of reading' mandates. *Education Week*. Available online: <https://www.edweek.org/teaching-learning/4-more-states-pass-science-of-reading-mandates/2023/05> (accessed on 8 November 2025).
- Schwartz, S. (2025, August 4). Districts using 'high-quality' reading curricula still supplement with other materials. Why? *Education Week*. Available online: <https://www.edweek.org/teaching-learning/districts-using-high-quality-reading-curricula-still-supplement-with-other-materials-why/2025/08> (accessed on 8 November 2025).
- Scott Foresman. (1956). *The new basic readers*. Scott Foresman.
- Snowling, M. J., Hulme, C., & Nation, K. (Eds.). (2022). *The science of reading: A handbook*. John Wiley & Sons.
- Stein, M., Johnson, B., & Gutlohn, L. (1999). Analyzing beginning reading programs: The relationship between decoding instruction and text. *Remedial and Special Education*, 20(5), 275–287. [CrossRef]
- Texas Education Agency. (1997). *Proclamation of the state board of education advertising for bids on textbooks*. Texas Education Agency.
- Texas Educational Agency. (1990). *Proclamation of the state board of education advertising for bids on textbooks*. Texas Education Agency.
- Thorndike, E. L. (1927). *The teacher's word book*. Teachers College, Columbia University.
- Tischner, C. M., Ebner, S. E., Aspiranti, K. B., Klingbeil, D. A., & Fedewa, A. L. (2023). Effectiveness of accelerated reader on children's reading outcomes: A meta-analytic review. *Dyslexia*, 29(1), 22–39. [CrossRef]
- Toyama, Y., Hiebert, E. H., & Pearson, P. D. (2017). An analysis of the text complexity of leveled passages in four popular classroom reading assessments. *Educational Assessment*, 22(3), 139–170. [CrossRef]
- Tuinman, J., Rowls, M., & Farr, R. (1976). Reading achievement in the United States: Then and now. *Journal of Reading*, 19(6), 455–463.
- University of Oregon. (2022). *DIBELS 8th edition 2021–2022 percentiles* (Technical Report 2201). University of Oregon.
- U.S. Congress. (2002). *No child left behind act of 2001, 20 U.S.C. § 6301 et seq.*. U.S. Congress.
- Wang, X., Zhao, Y., Petzold, L., Chandrasekar, R., Wang, S., Chen, M., & Feng, Y. (2024). Large language models for education: A survey and outlook. *arXiv*, arXiv:2403.18105v1.
- Williamson, G. L. (2008). A text readability continuum for postsecondary readiness. *Journal of Advanced Academics*, 19(4), 602–663. [CrossRef]
- Zeno, S. M., Ivens, S. H., Millard, R. T., & Duvvuri, R. (1995). *The educator's word frequency book*. TASA.
- Zipf, G. K. (1935). *The psychobiology of language*. Houghton-Mifflin.

**Disclaimer/Publisher's Note:** The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.