Hiebert, E.H. (in press). Standards, assessment, and text difficulty. In A.E. Farstrup & S.J. Samuels (Eds.). What research has to say about reading instruction (3rd Ed.). Newark, DE: International Reading Association.

Standards, Assessments, and Text Difficulty

Elfrieda H. Hiebert
University of Michigan

> Example 1: Every morning, when the farmer woke up, the first thing he saw was the roof of his little house. Every morning for breakfast he ate two flat cakes of ground corn. His wife had made them the night before. (Hill of Fire, Lewis, 1971)
>
> Example 2: “Absolute silence,” Papa whispered as the four of them slipped out to the street. The family walked and walked, past all the familiar houses and buildings in their neighborhood, being careful to hide in whatever shadows the creeping sunrise left undisturbed. (Night Crossing, Ackerman, 1994)

These excerpts come from texts that two different policy groups have identified as grade-appropriate for third graders. The time of day in each of the examples is a critical part of the text’s context. In Example 1, most third graders are likely to understand the farmer’s early morning routine from the repetition of the phrase “every morning.” A significant portion of third graders may not grasp the time of day in Example 2 implied by the phrase “shadows the creeping sunrise left undisturbed.” Their cognitive resources may be spent in figuring out words and phrases such as “absolute silence,” “whispered,” “shadows,” and “undisturbed,” leaving few resources to establish that the family in the second example is fleeing their home and that the rise of the sun presents hazards.

These two texts illustrate the dilemma confronting many teachers and their students. Standards and the assessments designed to ascertain students' attainment of these standards have proliferated over the past decade (AERA 2001); this movement is likely to gain momentum with new federal initiatives (The New York Times, January 23, 2001). In federal and state literacy initiatives, an often-stated goal is for American schoolchildren to be proficiently reading grade-appropriate text by the end of third grade (U.S. Department of Education, 1997). The frequent and consequential use of the phrase *grade-appropriate text* in national and state policies suggests well-researched views of the proficiencies that distinguish between those students who read, for example, second-grade texts and those who read third-grade texts. This maelstrom of activity around standards and accompanying assessments follows a dramatic shift in the types of texts used in reading instruction. As a result of the dissemination of research on readability from the early 1980s (Anderson, Hiebert, Scott, & Wilkinson, 1985), readability formulas are no longer used in the creation of textbook programs. Educators welcomed the demise of readability because of its constricting effect on school texts (Green & Davison, 1988). But descriptions of the systems or criteria that have replaced readability formulas have been few. The range in what has been identified as grade-appropriate is evident in the preceding two examples.

This chapter examines current definitions of text difficulty and how they are expressed in standards and assessments. Research on and designations of text difficulty in standards and assessments are examined with reference to third grade, in particular, because of the focus in national and state contexts on this grade level. This attention to definitions of grade-appropriate text at the third-grade level does not exclude interest in other levels. The emphasis that third-grade attainment has received from policymakers suggests that, if grade-appropriate levels have been identified at any grade level, the process should be expressed most clearly with third-grade text.

This examination of the bases for grade-appropriate text begins with a review of empirical and theoretical work on text difficulty. Historical work on readability, as well as its current manifestation in the form of Lexiles, is reviewed. Next, the archival literature of the past decade is reviewed to identify the models of text difficulty that have replaced readability formulas.

The next focus is on the treatment of text difficulty in recent national and state-level standards documents. In their model of curriculum, instruction, and assessment, Hiebert and Calfee (1992) identified the *curriculum,* or the goals of literacy instruction, as the engine for instruction and assessment. The standards and curriculum frameworks of states and national agencies would be expected to be the source for defining grade-appropriate reading proficiency. In the event that standards documents do not detail the nature of text difficulty, the textbook programs adopted by states provide the next source for definitions of text difficulty. Textbook programs continue to be a driving force within American reading instruction (Bauman, Hoffman, Duffy-Hester, & Ro, 2000).

The interpretation of text difficulty in assessments also merits examination. Ultimately, some type of assessment will be used to determine whether standards have been achieved. The difficulty of the texts on assessments, then, figures prominently into evaluations of standards and curriculum. Is there a match between the texts identified by state and federal agencies as providing the standards to be met and the texts of assessments? When the reviews of standards and assessments produced few descriptions of text difficulty, the texts of textbook programs that states use to support their students' attainment of standards and the assessments that are used to determine students' attainment of standards were studied. The models of text difficulty that were uncovered in the review of literature and the scheme that has been used for much of the 20th century—readability formulas—were used to describe current manifestations of text difficulty.

Perspectives on Text Difficulty

During most of the 20th century, readability formulas influenced evaluations of text difficulty and the creation of texts. This review begins with readability formulas and the most recent manifestation of this model—Lexiles. Amid calls for authentic text in reading instruction during the late 1980s, data on readability formulas were no longer mandated as part of state or district textbook adoptions. The archival literature of the past decade was the basis for a review of literature on alternatives to readability formulas as the basis for establishing grade-appropriateness of texts.

Readability Formulas

Since their inception in the 1920s (Lively & Pressey, 1923), readability formulas have established text difficulty on the basis of syntactic and semantic complexity. Semantic complexity is measured by either word familiarity as compared to a particular list of words or by word difficulty as measured by the number of syllables per word. The number of words per sentence determines syntactic complexity. From the 1920s through the 1980s, readability formulas were the guide for the creation of texts, not simply the evaluation of texts (Green & Davison, 1988). School texts, whatever the content area, were required to comply with semantic and syntactic parameters of readability formulas.

During this period, text readability was also a central topic of research, as evidenced by the first volume of the Handbook of Reading Research (Pearson, Barr, Kamil, & Mosenthal, 1984). Klare's (1984) chapter on readability was one of the longest in that handbook, accounting for 7% of the volume. Klare (1984) predicted that readability would continue to be a central area of research and practice. But a study of the subsequent two volumes of the Handbook of Reading Research (Barr, Kamil, Mosenthal, & Pearson, 1991; Kamil, Mosenthal, Barr, & Pearson, 2000) produced only one reference to readability. This reference occurred in the second volume, where the discussion of readability accounted for 3 pages or .3% of the total volume. By the third volume of the Handbook of Reading Research in 2000, readability was not listed in the index.

While this research was not summarized in either the second or third volumes of the Handbook of Reading Research, a line of research that paralleled Klare's (1984) comprehensive review was identifying problems with readability formulas. In the 1980s, researchers discovered that roadblocks to comprehension were inadvertently created by strict compliance to readability formulas (Bruce, 1984; Green & Davison, 1988). By substituting high-frequency words for more infrequent and often more descriptive words, meanings were changed or made obscure. In making sentences shorter, conjunctions were eliminated, often obscuring the causal connections that had helped make the ideas comprehensible (Green & Davison, 1988). Others argued that bland or missing descriptive language created “primerese” (Amsterdam, Ammon, & Simons, 1990), which could be unfamiliar and even nonsensical for beginning readers.

Others compared students’ comprehension of well- and poorly structured texts (Beck, McKeown, Omanson, & Pople, 1984; Brennan, Bridge, & Winograd, 1986). Students’ superior performances on the well-structured texts were taken as evidence that readability formulas were detrimental for effective comprehension. In none of these studies of restructured texts, however, was the focus on children at the very beginning stages of reading. The nature of scaffolds for beginning readers had not been addressed when these critiques of school texts, including the role of readability formulas in creating or manipulating these texts, were communicated to practitioners through Becoming a Nation of Readers (Anderson et al., 1985). The message that readability formulas can make text more difficult for readers struck a chord with teachers. In 1987, California’s Framework for English/Language Arts (California English/Language Arts Committee, 1987) stated that texts needed to consist of authentic literature to be on the state-approved list for purchase by California school districts. Texts that had been manipulated to comply with readability formulas could not be purchased with state-allocated textbook funds. Several years later, the Texas Education Agency (1990) followed suit. An analysis of the textbook series that were adopted in Texas in 1993 confirmed that contrived text had been replaced with literature (Hoffman et al., 1994).

Since the Hoffman et al. (1994) analysis, Texas and California have issued new guidelines for their 2000 (Texas Education Agency, 1997) and 2002 (California English/Language Arts Committee, 1999) textbook adoptions. Both states have called for texts at the beginning stages to have high percentages of decodable words—a guideline that has been followed in the textbooks adopted for use in Texas (Hiebert, 2000a). But the pace at which words with different letter-sound relationships are introduced and the repetition of these words presumably influences the difficulty of a text for beginning readers. Both the Texas and California guidelines were silent on this issue. For example, Hiebert (2000a) found that 40% of the unique words appeared only once in the first 10 texts of four of the largest reading/language arts textbook programs.

Such features in text for beginning readers have redirected educators' attention to readability formulas. The developers of the Lexile system—the most popular readability formula at this time—claim that it is not a readability formula (Smith, Stenner, Horabin, & Smith, 1989). Lexiles, however, are derived from the same two measures that are used to compute readability formulas: semantic difficulty, as measured by the presence of the texts' words on a word list, and syntactic difficulty, as measured by sentence length. According to the Lexile Framework for Reading (MetaMetrics, 2000), there are six ranges of Lexiles that cover the elementary grades:

First Grade: 200–370; Second Grade: 340–500; Third Grade: 480–670; Fourth Grade: 620–820; Fifth Grade 770–910; and Sixth Grade: 870–1000. The scale claims to measure texts through college-level with Lexiles through 1600.

To demonstrate the data provided by Lexile analyses, consider the most recent Harry Potter offering: Harry Potter and the Goblet of Fire (Rowling, 2000). This book is given a Lexile of 880. Another popular children's book, Charlotte's Web (White, 1952) has a Lexile of 680, which is the same Lexile rating as Grisham's (1990) The Firm. This comparison illustrates the ambiguity of data presented in a scale from 200 through 1600 that has been disassociated from its semantic and syntactic criteria. For third-grade level, for example, the range of 480 to 670 provides little indication of the corpus of words students need to automatically recognize. The developers describe sentence length and the presence of words from a list of words according to frequency of appearance in written English (Carroll, Richman, & Davies, 1971) as the basis for establishing Lexiles. Neither the framework nor the individual Lexile given for a text provides even a hint of precisely which corpus of high-frequency words relate to which level.

While legitimate criticisms can be leveled against readability formulas, including Lexiles, the disjuncture between the work of researchers and the needs of school-based educators in securing appropriate texts for students is apparent in the popularity that the Lexile system is enjoying. Six states currently have adopted the Lexile framework as a means for selecting textbooks (MetaMetrics, 2000). One of these, North Carolina, has linked end-of-grade tests to Lexiles and has mandated the use of Lexiles in textbooks that are purchased with state funds (North

Carolina Public Schools, 2000). The Reading Excellence Act also identifies Lexiles as one means of assessing performance. Harcourt Brace, a major publisher of an elementary textbook and test program, promises its customers that its materials comply with the Lexile framework. Scholastic Inc., too, lists Lexiles on covers of its books for the educational market.

Recent Perspectives on Text Difficulty

An extensive review of literature was conducted to determine the methods of establishing text difficulty that have replaced readability formulas in research and practice. The review began with the issues from 1991–2000 (the period since the last publication of this volume) of six archival journals: Reading Research Quarterly, Journal of Literacy Research, Journal of Educational Psychology, American Educational Research Journal, Journal of Educational Measurement, and Scientific Study of Reading. Chapters on text and policy were also analyzed from the most recent Handbook of Reading Research (Kamil et al., 2000). When this search produced only one scheme that had been used in a research study, the review was extended to include books and journals aimed at practitioners as well as journals oriented toward special education. This process produced three additional systems for analyzing texts; summaries of all four follow.

Engagingness, predictability, and decodability. Hoffman and his colleagues (1994) wanted to capture in as many ways as possible the differences between the first-grade texts of textbook programs published in 1986/87 and 1993. With this in mind, they analyzed texts on a variety of measures, particularly focusing on three measures that they developed: predictability, decodability, and engagingness. Hoffman et al. reported that the 1993 texts were more engaging than the 1986/87 texts when raters considered the content of texts, the sophistication of language in texts, and the design of the texts, as indicated by holistic ratings of 3.2 (on a five point scale) for the 1993 texts and 2 on the 1986/87 texts.

Predictability was established through an analysis of nine characteristics that make texts more predictable to children, including repeated patterns, familiar concepts, rhyme, rhythm, and cumulative pattern. The analyses of features as well as a holistic rating showed the early level of the 1993 texts to be substantially more predictable than the same level of the 1986/87 texts. A decodability scale was also applied with 1 representing a low level of decodability and 5 indicating the presence of common, easily decodable words The 1986/87 texts were evaluated as more decodable at both the beginning and end of the year than the 1993 texts: the 1986/87 texts were given ratings of 4.5 and 3.9, respectively, whereas the 1993 texts were rated 3.2 and 2.8. These descriptions were not placed within a theoretical construct that describes the role of these variables at different points in children's reading development or appropriate levels of the variables of engagingness, predictability, and decodability for different points in reading acquisition.

Potential for accuracy. Stein, Johnson, and Gutlohn (1999) evaluated the correlation between the elements of words in student materials and the guidance on teaching these words in the teachers' guides at the initial stages of reading. Stein et al. were interested in the potential that children had accurately decoding a text as a function of instruction. Potential for accuracy was established in the following manner. If the letter/sound correspondences for initial m, /m/ and b, the vowel a /a/, and the final n, d, t, and p have been introduced in previous lessons, words such as *mad man,* mat, *map, bad,* and *bat* have in a target text meet the criterion of “potential for accuracy” in decoding. Stein et al. reported considerable variation in the 1995 copyrights in the potential for accuracy. The range in potential for accuracy went from the 2% given to Open Court texts to 56% awarded to Scholastic texts.

There are a number of issues related to decodability in learning to read that Stein and her colleagues do not describe. For example, Stein et al.’s “potential for accuracy” measure assumes that children should be able to apply phonics elements after an instructional lesson—even if that lesson has attended to a number of different letter-sound relationships and high-frequency words as is the case with first-grade reading programs (Hiebert, Menon, Martin, & Huxley, 2000; Stein, Johnson, Boutry, & Bortleson, 2000).

<u>Task analyses of text</u>. The goal of the Text Elements by Task (TExT) work of Hiebert and her colleagues (Hiebert, 1999, 2000a, 2000b; Martin & Hiebert, 2001; Menon & Hiebert,1999) is to identify variables that influence the difficulty of the task posed by texts. For example, what capabilities would third graders need to read Examples 1 and 2 automatically, permitting full attention to comprehension? While many different factors influence readers’ interpretations of texts, students need to be able to recognize words automatically to comprehend texts (Adams, 1990; Fuchs, Fuchs, & Maxwell, 1988). The TExT index builds on work by Carroll et al. (1971) and Hayes, Wolfer, and Wolfe (1996) describing the relative frequency of words in texts. Using newspapers as the baseline, Hayes et al. have established the “LEX” scores of texts. Texts with (+) LEX scores are more difficult than newspapers, while texts with (-) LEX

scores are less difficult than newspapers. The larger the numerical score, the harder (or easier) the text. Hayes et al. reported that a 1956 Scott, Foresman preprimer had a LEX of –80.5, whereas children's books for ages 9 through 12 had a LEX of –29.6 and those for preschoolers had a LEX of –37.8. From a corpus of 5 million words from textbooks across subject areas used in grades 3 through 9, Carroll et al. produced an index similar to the LEX for individual words. Neither index is theoretically grounded in a model of reading acquisition or development. Is a LEX of –80.5 more appropriate than a LEX of –29.6? At what point should a word with a frequency rating of 500 be read?

A model of text difficulty needs to explain how, when, and why the relative frequency of words matters. In beginning to build such a model, Hiebert (2000b) argues that highly decodable words should be treated differently than less decodable words. Even though words such as *guide* and *fence* are more difficult to decode than *win* and *pat,* words in both sets are counted as "difficult" within a readability formula such as the Dale-Chall (Chall & Dale, 1995) or the (Spache, 1981). In the TExT perspective, words with simple vowel patterns such *win* and *pat* are counted as recognizable vocabulary at an early stage, while single-syllable words with complex patterns such as *guide* or *fence* are not.

After studying hundreds of texts and students' reading of these texts, Hiebert (2000b) has proposed the complex word factor (CWF) as an indicator of the task demands for recognizing words in primary-level texts. The CWF indicates the number of words that will be difficult in a text when measured against a curriculum. Extended samples from texts are used to establish the number of unique words per 100. The unique words are further analyzed to establish the number that lie outside the high-frequency curriculum and, among this latter group of words, those with phonic or syllabic patterns that are beyond that level's curriculum. When the longer texts from which Examples 1 and 2 are drawn are analyzed against a third-grade curriculum (proficiency with the 1,000 most frequent words and all vowel patterns in single-syllable words), the CWF for Hill of Fire is 4 and for Night Crossing, 14. On average, a 100-word sample in Night Crossing will have 14 unique words that fall outside the third-grade curriculum—words such as *absolute, silence, whispered, neighborhood, shadows,* and *undisturbed*. Hill of Fire will have an average of 10 fewer unique words per 100 that fall outside the third-grade curriculum. From the TExT perspective, texts with more rare and multisyllabic words will pose a greater challenge for students who can The same text can have a number of different CWFs, depending on the underlying curriculum against which it is assessed. After examining many sources, Hiebert (in press) could find few curricula for word recognition across the primary grades. In lieu of established curricula, Hiebert identified "assumed" curricula at different grade levels. That is, what core group of words needs to be known to successfully read approximately 90% of a text? The percentage of 90 was chosen because it is typically described as characterizing a minimum for meaningful or instructional reading (Betts, 1946; Clay, 1985). Hiebert (2000b) tested various curricula against 500 texts drawn from textbook programs and assessments for the primary grades. The assumed curriculum of third-grade assessments, Hiebert concluded, was automaticity with the 1,000 most frequent words (Carroll et al., 1971); words related to the 1,000 most frequent words by inflected endings such as *ed, ing,* and *s/es;* and the ability to recognize all vowel patterns within single-syllable words. Hiebert (2000b) has reported data indicating that the CWF is distinct from the a commonly used readability formula, that of Fry (1968) and Lexile ratings.

Text leveling. The assignment of texts as benchmarks for particular grade levels by experts in children's literature or in reading processes has precedence in projects such as the National Assessment of Educational Progress (National Assessment Governing Board, 1994), where committees of experts select texts. Textbook publishers also use experts to select the literature for particular grade-level anthologies. The term "text leveling," however, refers to a particular process that has been disseminated through Reading Recovery (Peterson, 1991) and its classroom-based extension, guided reading (Fountas & Pinnell, 1999, 2001). The 20 levels of Reading Recovery (Peterson, 1991) and the 26 levels of guided reading (Fountas & Pinnell, 1999, 2001) are differentiated along four dimensions: (a) book and print features; (b) content, themes, and ideas; (c) text structure; and (d) language and literary elements. The nature of the definitions provided in Fountas and Pinnell (1999) for the features of the three levels that are end-of-grade accomplishments for first, second, and third grades appear in Table 1. [i]

Insert Table 1 about here

The scoring to date has been holistic; in other words, a text is assigned a single level. For example, the two texts that were given, respectively, Lexiles of 880 and 680—Harry Potter and the Goblet of Fire and Charlotte's Web—are given levels of "T" and "Q." Scores are not reported for individual categories (e.g., content, text structure) in sources discussing the ratings, such as Fountas and Pinnell (1999, 2001). Further, no research studies have reported on the relative weight given to different dimensions in these holistic ratings or whether the dominant factors vary for different types of texts. For example, print features would be expected to weigh in more heavily at the very early levels such as A through E but not at levels V through Z. Text leveling, however, is enjoying considerable popularity; numerous sets of little books advertise that they have been leveled according to this scheme.

Summary of current text difficulty schemes. Once the state and national standards and the texts of textbook and test programs were reviewed, the lack of measurement of text difficulty became readily apparent. In most instances, descriptions of how text difficulty was perceived or had been established were not explicit in the standards documents or program documentation. Further, there were no data on what characteristics had been considered in designating particular texts as grade-appropriate for third grade. Texts were simply presented as appropriate third-grade texts.

Shared constructs were needed in order to describe the common traits of appropriate third-grade texts. In lieu of analyses by publishers of texts and tests or standards committees, such analyses were conducted for this chapter. While existing indices of text difficulty may be less than perfect, those that bring appropriate data to bear were used. Although the meaning of readability scores and Lexiles is difficult to ascertain, such constructs are used in the marketplace. Consequently, when texts were proposed as exemplifying particular grade-appropriate levels, readability and Lexile levels were obtained. Because of the outdatedness of the word lists used in other formulas (e.g., Chall & Dale, 1995; Spache, 1981), a readability formula that relies on assessment of syllables per word for the measure of semantic complexity was used. The readability formula that was chosen was Fry's (1968) measure, which assesses sentence length and the number of syllables per word.

Of the four systems that were just described, only the text leveling system enjoys application and popularity in the marketplace. Further, it was the only text difficulty system cited in national and state standards documents, textbook programs, or tests. Since Fountas and Pinnell (1999) advise teachers that they can become experts in text leveling through study of the dimensions in their book, it seemed plausible to form a cadre of evaluators who could level the texts on tests as well as those in textbook programs that do not report these levels. When Table 1 was given to a group of experienced teachers (all with Master's degrees), there was simply too much ambiguity across the descriptions of an element to obtain interrater agreement for a set of 27 books selected from Fountas and Pinnell's (1999) list of 7,500 leveled books. These 27 texts represented 3 titles for each of the levels from H through P. Further efforts were devoted to identifying "anchors" or books that typified the characteristics of each of the three target levels, a technique that is frequently used in holistic scoring schemes (Calfee & Hiebert, 1991). The meaning of differences between "difficult content" (Level I), "higher level of conceptual understanding" (Level L), and sophisticated themes that "require interpretation and understanding"

(Level O) was difficult to ascertain when comparing a Level O book such as Class Clown (Hurwitz, 1987) and a Level I book such as Eat Up, Gemma (Hayes, 1988). Nor was it possible to establish commonalities in the conceptual load of a group of books at the same level such as the selected texts for Level L: Josefina Story Quilt (Coerr, 1986), Cam Jansen and the Mystery of the Television Dog (Adler, 1981), and Amelia Bedelia (Parish, 1963).

The texts that have been proposed by the New Standards as end-of-third-grade, using the Reading Recovery levels, will be reviewed, and the texts of the Developmental Reading Inventory (DRA)(Beaver, 1997), which is used in many Reading Recovery efforts, will be included for analyses with the texts of other assessments later in this chapter. However, without refinement of the leveling criteria, the text levels of Fountas and Pinnell (1999, 2001) and of Peterson (1991) cannot be applied with fidelity.

The Hoffman et al. (1994) and Stein et al. (1996, 1999) systems describe critical dimensions of beginning reading texts but do not propose how to relate these features to texts at a second or third-grade level. For example, how predictable should texts be at the beginning of first grade? Research suggests that predictable texts could interfere with students' attention to individual words at particular stages of reading development (Johnston, 2000). The Stein et al. (1999) measure of decodability begs the question of how many phonics elements beginning readers—especially those who may be vulnerable in learning to read—can acquire as a result of a single lesson. The adaptation of the LEX system's word frequency in the TExT method (Martin & Hiebert, 2001) and its augmentation with highly decodable words as well as ratios of unique to total words makes it possible to compare the features of text at third grade with those for second and fourth grade, thus providing a point of reference.

Three indices, then, were used in analyzing texts that are proposed as part of state and national standards and those in published textbook programs and tests: (a) a readability formula (Fry, 1968); (b) Lexile analysis; and (c) the CWF index of the TExT system (Martin & Hiebert, 2001). When the text levels (Fountas & Pinnell, 1999, 2001) were provided, these, too, are reported.

State and National Standards

There are two sources for standards: reports from national committees and organizations and state frameworks. For this review, three documents from national organizations were examined, as were three documents from state agencies.

National Standards

In that education is an area over which states have jurisdiction in the United States, the number of national committees, agencies, and organizations that have produced standards documents is limited. Even so, all of the standards documents that have been produced over the past decade, including the NAEP framework (National Assessment Governing Board, 1994) and the standards of the International Reading Association and the National Council of Teachers of English (1996) were not analyzed. Rather, two national-level documents that have been in the public eye over the past several years were examined: Preventing Reading Difficulties (Snow, Burns, & Griffin, 1998) and Teaching Children to Read (National Reading Panel, 2000). In addition, the standards document from the New Standards Project (New Standards Primary Literacy Committee, 1999) was examined. The intent of the former two documents was to summarize critical research, not to achieve consensus on national standards. Neither was intended as a standards document. However, in their efforts to report on critical current research, these reports provide the best available indicators of perspectives on text difficulty in students' reading development.

The most recent of these documents, Teaching Children to Read, was searched with four terms—text difficulty, text leveling or levels, readability, and Lexiles. The word "readability" was found twice in the report. The first reference occurs in a description of an intervention where teachers estimated readability levels of text; the second was in a description of a research study where the readability level and the number of total words in texts were reported (Pressley et al., 1992). While a conference paper by the developers of Lexiles is referenced (Smith et al., 1989), the term Lexile does not appear in the body of the National Reading Panel's report.

The National Reading Panel's emphasis on replicable outcomes means that a preponderance of the conclusions about student reading accomplishments was based on norm-referenced tests. Despite an emphasis on evidence, there was no discussion of the contents of these tests or of the texts that were used in treatments. Students' performances on the special study of the NAEP (Pinnell et al., 1995) were summarized as "45% dysfluent" on "grade-level stories that the students had read under supportive testing conditions." (National Reading Panel, 2000, p. 3–5). But there is no information about what constitutes grade-level texts or whose definition of "grade-level" the NAEP stories represented.

In the report that preceded the National Reading Panel's report, Preventing Reading Difficulties, Snow et al. (1998) described the need for third-graders to have the skills, habits, and learning strategies required for fourth-grade success. However, descriptions of the content of these skills and learning strategies are limited. On a table listing grade-level accomplishments, the only reference to the characteristics of texts that exiting third-graders should be able to read is the following: "Reads aloud with fluency and comprehension any text that is appropriately designed for grade level." (Snow et al., 1998, Table 2–2, p. 83).

While Preventing Reading Difficulties is more explicit about the need for attainment of a particular text level than Teaching Children to Read, neither document sheds any light on the demands of texts. Does automaticity need to extend beyond the 200 to 300 most frequent words that one of the mainstream textbook publishers cites as the standard for proficient third-grade reading in its program (Afflerbach et al., 2000)? What types of metaphors do exiting third-graders need to be able to figure out? The flurry of activity on national reports provides no insight into what skills and strategies are needed to be successful with grade-level texts.

The primary-grade standards document produced by a committee of national experts for the New Standards Project was also examined. This document involved participation from scholars representing a range of perspectives. Unlike the authors of the other two national documents, the New Standards Committee was explicit in identifying the levels of appropriate reading for different grade levels. For each of primary grades from K–3, these standards describe text levels and give lists of book titles that exemplify the kinds of books students are expected to read. For the end of third grade, ability to read Level O texts is cited as the benchmark. While Fountas and Pinnell (1999) are not cited directly, the New Standards levels correspond to their scheme. Ten titles are given as exemplars of Level O texts that students should be able to read at the end of grade three. Two of these 10 titles were not accessible through major booksellers. The remaining eight titles are listed in Table 2. Also included in Table 2 is data on text difficulty according to the three indices of conventional readability (Fry, 1968), Lexiles, and the word recognition component of the TExT system.

Insert Table 2 about here

The texts identified as Level O according to Reading Recovery vary considerably in their ratings on other text difficulty schemes. On the Fry readability formula, the ratings range from the end of first grade through the middle of sixth grade. According to the Lexile framework, the texts have a similar range: from mid–second grade through sixth grade. The CWFs for the texts vary from 7 to 14 complex words per 100. On all of these indices, the variation in text difficulty is considerable for books that are reported to be on a similar level. The New Standards project gives teachers and children little concrete information on what appropriate third-grade text is. Unlike the two national reports, however, the New Standards project does recognize that text difficulty is a factor in describing goals for students' reading.

State Standards

Following the Goals 2000: Educate America Act (U.S. Department of Education, 1998) which supported states in formulating their standards, standards and framework documents in language arts/reading have proliferated. This analysis considered the standards documents of the three states that (a) rank in the top four states with regard to population and (b) identify textbooks as a state. These three states—California, Texas, and Florida—provide guidelines that textbook publishers need to meet for school districts to purchase their programs with state funds. With 25% of the American population residing in California, Texas, and Florida (U.S. Census Bureau, 2000), these three states' guidelines influence the opportunities for children in smaller states and in states where individual school districts or schools choose textbooks.

California's state standards documents are the only ones that cite texts exemplifying the content standards. The California standards (California English/Language Arts Committee, 1999) refer to the Recommended Readings in Literature (California Department of Education's Language Arts and Foreign Languages Unit, 1996) for text recommendations. Titles that were identified as appropriate for third grade were obtained from the six categories of literature presented in the document. For three categories, only a single third-grade title was in print; for one category, no in-print books were available. When there were several titles that were in-print for a category, 2 titles were randomly chosen. Ten titles were analyzed in all. The difficulty ratings of these 10 titles according to the three indices are provided in Table 2.

The standards documents from Florida and Texas contain many descriptions of what third-graders should be able to do. In Florida, the two standards that are devoted exclusively to reading attend to the effective use of the reading process and to the construction of meaning from a wide range of texts. The only mention made of word recognition in the Florida standards is one of the descriptors of Reading Standard 1—Uses the Reading Process Effectively. One of the ways a student shows evidence of attainment of this standard is "selects from a variety of simple strategies, including the use of phonics, word structure, context clues, self-questioning, confirming simple predictions, retelling, and using visual cues, to identify words and construct meaning from various texts, illustrations, graphics, and charts" (Sunshine Standard, Language Arts, Grade 3-5) (Florida Department of Education, 1999). Other points under this standard, the other Reading standard, and the two Literature standards refer to ways in which students will perform with texts, such as identifying personal preferences, identifying authors' purposes, and so on. No mention is made of the level of text with which students should be able to perform these processes.

In both Florida and Texas, the task of designating text levels falls to the textbook publishers and test developers. Since Florida will not adopt new textbooks until 2002 and Texas has just completed an adoption of new textbooks, Texas's current textbook offerings have been analyzed for text difficulty. Passages from two of the four mainstream programs that received a majority of Texas's business were analyzed. Ten passages from each program were analyzed, five representing the beginning of the third-grade texts and five representing the end of the third-grade texts. An equivalent number of words were analyzed for each passage. Summaries of these analyses are included in Table 2.

On all three measures of text difficulty, the range across the texts from the Recommended Readings in Literature (California Department of Education's Language Arts and Foreign Languages Unit, 1996) and Textbook Program I is considerable. The range in readability level is 5 and 7 grade levels for the California texts and the texts of Text Program I, respectively. On Lexiles, the texts range from second-grade—Hill of Fire—to sixth-grade and beyond—The Ox-Cart Man, Sleeping Beauty by Brothers Grimm, and Cocoa. Similarly, the CWF for these two programs show considerable variation. The range from the easiest to the hardest text, as measured by CWF, is 7.

On all three measures of text difficulty, Textbook Program II showed the most consistency. But even for this program, the differences among texts were considerable. The range for grade levels, according to the Fry readability formula, is 3; the range on the Lexiles is 410, equivalent to almost 2 and a half grade levels; and the range for the CWFs is 4.

One criterion against which the textbook programs can be measured is the progression in text difficulty. When textbook programs advertise one volume of the third-grade program as "3.1" and the other as "3.2," some order in difficulty is implied. Easier passages as measured by the Fry and Lexile measures occur with the same frequency in the second half of the program as they do in the first half. The demands of word recognition, as measured by the CWF, do show a progression in Textbook Program II but not in Textbook Program I. In the latter program, the average CWF of the first three passages is lower than the average of the last three passages in the program.

What conclusions about text difficulty can be derived from these analyses? In general, the texts proposed for third grade, including those proposed for the first period of instruction, are difficult. The New Standards' and the California exemplars' average Lexiles are both within the fourth-grade range, and those for the two textbook programs are in the upper ranges of third grade. Two of the average figures for the readability formulas also fall into the fourth-grade range. The average CWF is the lowest for the California texts, where 50% have figures that would be expected from instructional materials—4 to 6. In the case of the other three sets presented in Table 2, the average CWFs of 8, 9, and 10 suggest that third graders are expected to be adept at reading multisyllabic words that are not highly frequent. Whereas the need to attend carefully to 4 to 6 words per 100 is within an instructional range, the need to attend to 8 to10 words per 100 is approaching what has frequently been regarded as frustration level for third graders.

Text Difficulty and Assessment

Since national and state standards generally fail to address the grade-appropriateness of text, the text levels on the tests that are given to establish whether students have attained standards are critical in shaping perceptions of whether students can read grade-appropriate texts. Large states such as Florida and Texas have tests specially designed by the large test publishers. The documentation that accompanies these tests, however, does not contain any information on the manner in which particular texts were assigned to particular test levels. To understand the difficulty of current assessments, it was necessary to conduct an analysis similar to the analysis of exemplars offered as part of standards documents and of texts presented in instructional programs. The results of such an analysis, it was determined, would be of particular interest relative to the patterns obtained from the analyses of the texts from standards documents and instructional programs.

The texts of six tests were analyzed. These six tests represented three types of assessments. The first type consisted of widely used, norm-referenced, silent reading tests. One of the two tests included in this group was the Stanford Achievement Test, 9th Edition (SAT-9). At the present time, student scores on this test provide the sole basis for the State Testing and Reporting (STAR) program in California. In 2000, the state of Florida also mandated that the SAT-9 be administered to all students in grades 3–10. A second norm-referenced, silent reading test, the Comprehensive Test of Basic Skills (CTBS), was the second entry in this category. While none of the three large states whose policy activities are influential in reading education use this test, the CTBS is among the top handful of frequently used, norm-referenced tests.

The second group of tests were those that have been devised for Florida and Texas. The Florida Comprehensive Assessment Test (F-CAT) consists of both multiple-choice and essay tasks. Students read two different passages of 1,000 or more words on two consecutive days. The Texas Assessment and Standards (TAAS) is similar in design to the SAT-9 and CTBS. The comprehension section consists of a series of relatively short passages (275 words on average), each followed by a group of multiple-choice questions. Like the F-CAT, however, the TAAS is criterion-referenced. Each state sets a standard that students need to reach to have met state expectations.

The third category of assessments consisted of oral reading assessments. The use of informal reading inventories has a long history (Gray, 1920). This group of assessments is particularly interesting in that informal reading inventories continue to present texts according to grade levels. Because of its popularity in the marketplace as well as the existence of a recent revision, the Qualitative Reading Inventory (QRI; Leslie & Caldwell, 2001) was chosen as one of the two assessments in this category. The second was the DRA (Beaver, 1997). This assessment is advertised as exemplifying the leveled system of Reading Recovery. This assessment has enjoyed widespread use among Reading Recovery and guided reading users.

In all cases, the texts in the analysis came from the forms or versions of tests intended for third-grade. A similar number of samples and of words per sample was analyzed for each assessment. Characteristics of the six assessments on the three measures of text difficulty (Lexile, Fry readability, and CWF) are given in Table 3.

Insert Table 3 about here

Within each of the three indices, the ratings for the six assessments are quite consistent. When compared to the data on texts summarized in Table 2, the consistency is substantial. Further, most of the ratings for the assessments are within a third-grade band, as identified by the developers of the three text-difficulty measures. The DRA could present more of a challenge than the other assessments, at least in its demands for recognition of complex words. Even a CWF of 6 however, is reasonable when compared to demands of the exemplars from the New Standards list where 10 was the average CWF.

One aspect of the norm-referenced tests and the two state tests that looms large for third graders is the role of reading speed. Speed of reading can be captured on the QRI and DRA, although the procedures for both of these instruments do not emphasize gathering this data to the same degree as word accuracy information. Ample evidence exists to indicate that the factor of speed with which students recognize words figures heavily into students' comprehension (Pinnell et al., 1995). On the special study of the 1994 NAEP, Pinnell et al. (1995) found that the below-basic readers on the NAEP (approximately 40%) did not differ significantly from above-basic and basic-level readers in accurately recognizing words. However, the rate at which below-basic readers read a text was significantly slower than that of the other two groups. This rate of oral reading was highly related to students' comprehension on a silent reading of the same text.

For the F-CAT, in particular, rate of reading is likely a factor. This test has the most words that fall outside the third-grade curriculum (at least when measured by the 1,000 most frequent words and vowel patterns in single-syllable words). Further, in order to answer a handful of questions, students need to be able to sustain a theme and supporting details over an eight-page text.

Next Steps

The aim of this chapter was to describe the definitions of text difficulty in standards and framework documents and the treatment of text difficulty in assessments. In both national and state policies, the goal of "proficient reading of grade-appropriate text" is frequently cited. Despite this rhetoric by politicians about the need for proficient reading of grade-appropriate text at third grade, the research literature and the standards espoused by states and national agencies are vague about what constitutes such text and the proficiencies required to read it. The summaries of research from two prestigious panels of scholars over the last two years provide no insight into what makes one text appropriate for a particular grade and another text appropriate for a subsequent grade. State frameworks do not make explicit the nature of texts with which students at different levels need to be proficient. When exemplar texts are identified by a national center (New Standards Primary Literacy Committee, 1999) and a state frameworks document (California English/Language Arts Committee, 1999), there is substantial variability across them. Neither of these two efforts specify the critical competencies that are required to be successful with these texts.

With state standards mute on how text difficulty figures into the goals for children's reading, publishers of tests and textbooks have been placed in the position of determining what is appropriate grade-level text. But publishers' interpretations of appropriate third-grade reading levels vary. Among the publishers of textbooks and committees that designate standards, the interpretation appears to be that students should be able to read a particular body of literature, whatever the difficulties it poses for word recognition. With one exception, standards committees do not identify the dimensions on which these titles have been chosen. The exception to this pattern is the identification of books according to genre in the California's Recommended Readings in Literature. Even in the choices of titles for different genres in California, answers to a host of questions are not evident: Do the titles deal with different treatments of critical themes that should be understood at this developmental level? Are there particular figurative language devices that are evident in these books and with which third graders should be facile? Are these simply "popular" books among teachers or children's literature experts? Textbook publishers are somewhat more explicit in placing particular books within themes. But even these themes are often sufficiently general that a variety of books could fit the bill.

When criteria that describe the word recognition demands of texts are applied to these books, there is little consistency either within a set of books or across the different sets. For the Fry and Lexile ratings, almost 50% of the texts fall into the instructional range for third grade (grade 2 through 3.4). Another 30% fell solidly into the range for sixth grade and higher (5.0 and beyond), and the remaining 20% was in the fourth- and fifth-grade range. This range in difficulty is considerable. Depending on the text that a school district might chose from the list provided by either the New Standards or California's Recommended Readings in Literature, children's representation as readers could be quite different.

While publishers of textbook programs and standards committees provide a varied picture of text difficulty, test publishers' interpretations of appropriate third-grade text has been more consistent. The characteristics of tests are similar on all three measures of text difficulty.

The texts of the state tests tend to be somewhat more difficult in their word recognition demands than the norm-referenced tests. Further, the leveled texts of the DRA are also at the higher end of the distribution in terms of word recognition demands. But as a group, the variation is considerably less among the tests than it is among the benchmark and instructional texts. On the Lexile measure, all of the tests except for the DRA fall at the lower end of the third-grade range, and all but the DRA fall into the third-grade range according to the Fry readability formula. The CWF data indicate that the use of infrequent and/or multisyllabic words is moderate—an average of 5 rather than the 9 of the texts offered by standards committees and textbook programs.

To illustrate the differences between the tasks of the tests and the textbooks, a portion of a passage from the TAAS is excerpted below in Version 1. It has a CWF of five, the average across the six assessments reviewed in this chapter. In Version 2, the TAAS passage has been modified to achieve a CWF of 9, the average for the exemplar texts of the New Standards and California Standards as well as the two textbook programs created for Texas.

Version 1: Buddy, the first guide dog in the United States

> Buddy stopped before she crossed the street. She looked in all directions. There was a lot of traffic. Cars and trucks zoomed by, but Buddy knew what she was doing. She walked carefully across the street. The people on the other side cheered. They could not believe what they had just seen. A dog had led a man across a busy street. It was an amazing sight. The man Buddy had led across the street was blind. The dog, Buddy, was the first guide dog in the United States.

> Version 2: Buddy, the first guide dog in the United States
>
> Buddy halted at the intersection. She searched the street in both directions. Cars and trucks zoomed by, but Buddy moved and acted with confidence. She waited for a gap in the traffic and then walked carefully across the street. The audience on the other side cheered. They were astounded by what they had just seen. A dog had guided a man across a busy street. It was an amazing sight. The man, Morris Frank, was completely blind. The dog, Buddy, was the first guide dog in the United States.

The number of words and ideas in the two versions are the same. Where the two passages differ is in the presence of infrequent words, particularly multisyllabic words. If students were to stumble on the challenging words in Version 1 on an informal reading inventory, they would be judged to be at their instructional levels according to typical criteria (Betts, 1946). Stumbling on the challenging words in Version 2 would place students at a frustration level (Betts, 1946).

The third-grade texts of instruction are closer to the standard for text difficulty for the fourth-grade assessment of the NAEP—at least the standard of passages that have been released (see, e.g., Pinnell et al., 1995)—rather than the standard for text difficulty in the third-grade assessments. One hypothesis is that the current design of textbook programs is intentional: Students should be taught with and practice on texts that are more difficult than the tests of assessments (Stahl, 2000). Previous work, however, indicates that students are most successful when they have frequent occasions to read texts on which they do not make substantial numbers of errors (Fisher, Filby, Marliave, Cahen, Dishaw, Moore, & Berliner,1978). This work needs to be revisited and follow-up studies need to be conducted.

To provide guidance to policymakers, teachers, and, ultimately, the students who are not gaining the necessary reading levels, researchers need to describe the nature of literacy tasks of the digital age and the proficiencies that are required to be successful with these tasks. A critical part of this description involves features of texts that challenge readers. In this chapter, three methods of describing text difficulty were reviewed. One is in its early stages of development—the CWF. A second—that of text leveling—provides insufficient information to guide a group of experts in leveling texts. The third method of conventional readability formulas and a recent manifestation of this method—Lexiles—fails to inform teachers and students about the critical processes that readers are lacking. Advice to a teacher to pick texts for particular students that have shorter sentences will hardly do the job.

Other methods of selecting texts have been suggested, such as using a standard of engagingness or quality of literature (Alvermann & Guthrie, 1993). Proof that children can read texts because they are engaging or profound needs to be obtained. A substantial amount of research needs to be directed to the issue of how different types of text influence the task for readers. Such models will eventually be complex and will include the medium of the text (e.g., electronic or printed form (Leu & Kinzer, 2000). But to ignore the features of texts and to assume that the quality of literature is all that matters is an inadequate response. Simply too many children are struggling with the tasks that are essential for full participation in the digital age to ignore the role that texts play in the development of high levels of literacy.

References

Adams, M.J. (1990). Beginning to read: Thinking about print. Cambridge: MIT Press.

Afflerbach P., Beers, J., Blachowicz, C., Boyd, C. D., Diffily, D., Gaunty-Porter, D., Harris, V., Leu, D., McClanahan, S., Monson, D., Perez, B., Sebesta, S., & Wixson, K.K. (2000). Scott Foresman Reading. Glenview, IL: Scott Foresman.

Alvermann, D., & Guthrie, J. (1993). Themes and directions of the National Reading Research Center (National Reading Research Center, Perspectives in Reading Research, No. 1). Athens, GA: University of Georgia, National Reading Research Center.

American Educational Research Association (2001). High-stakes testing in pre K-12 education. Washington, DC: Author.

Amsterdam, L., Ammon, P., & Simons, H. (1990). Children's elicited imitations of controlled and rewritten reading texts. Journal of Educational Psychology, 82, 486–490.

Anderson, R. C., Hiebert, E. H., Scott, J. A., & Wilkinson, I. A. G. (1985). Becoming a nation of readers: The report of the Commission on Reading. Champaign, IL: The Center for the Study of Reading; Washington, DC: National Institute of Education.

Barr, R., Kamil, M. L., Mosenthal, P., & Pearson, P. D. (Eds.). (1991). Handbook of reading research (Vol. 2). New York: Longman.

Baumann, J.F., Hoffman, J.V., Duffy-Hester, A.M., & Ro, J.M. (2000). The First R yesterday and today: U.S. elementary reading instruction practices reported by teachers and administrators. Reading Research Quarterly, 35, 338-377.

Beaver, J. (1997). Developmental Reading Assessment. Parsippany, NJ: Celebration Press.

Beck, I. L., McKeown, M., Omanson, R., & Pople, M. (1984). Improving the comprehensibility of stories: The effects of revisions that improve coherence. Reading Research Quarterly, 19, 263–277.

Betts, E. (1946). Foundations of reading instruction. New York: American Book.

Brennan, A., Bridge, C., & Winograd, P. (1986). The effects of structural variation on children's recall of basal reader stories. Reading Research Quarterly, 21, 91–104.

Bruce, B.C. (1984). A new point of view on children's stories. In R.C. Anderson, J. Osborn, & R.J. Tierney (Eds.), Learning to reading in American schools: Basal readers and content texts (pp. 153-174). Hillsdale, NJ: Erlbaum.

Calfee, R.C., & Hiebert, E.H. (1991). Classroom assessment of literacy. In R. Barr, M. Kamil, P. Mosenthal, & P.D. Pearson (Eds.). Handbook of research on reading (2nd Ed., pp. 281-309). New York: Longman Publishers.

California Department of Education's Language Arts and Foreign Languages Unit. (1996). Recommended readings in literature (kindergarten through grade eight). Sacramento: California Department of Education.

California English/Language Arts Committee. (1987). English-Language Arts Framework for California Public Schools (Kindergarten Through Grade Twelve). Sacramento: California Department of Education.

California English/Language Arts Committee (1999). English-Language Arts Content Standards for California Public Schools (Kindergarten Through Grade Twelve). Sacramento: California Department of Education.

Carroll, J. B., Davies, P., & Richman, B. (1971). Word frequency book. Boston: Houghton-Mifflin.

Chall, J., & Dale, E. (1995). Readability revisited: The new Dale-Chall readability formula. Cambridge: Brookline.

Clay, M.M. (1985). The early detection of reading difficulties (3rd Ed.). Portsmouth, NH: Heinemann.

Fisher, C. W., Filby, N. N., Marliave, R., Cahen, L. S., Dishaw, M. M., Moore, J. E., & Berliner, D. C. (1978) Teaching behaviors, academic learning time, and student achievement: Final report of phase III-B, beginning teacher evaluation study. San Francisco: Far West Laboratory.

Florida Department of Education (1999). Instructional Materials Specifications: Reading Grades K-12 (2001–2002 Adoption). Tallahassee, FL: Author.

Fountas, I., & Pinnell, G. S. (1999). Matching books to readers: Using leveled books in guided reading, K–3. New York: Heinemann.

Fountas, I.C., & Pinnell, G.S. (2001). Guiding readers and writers: Grades 3-6. Heinemann: Portsmouth, NH.

Fry, E. B. (1968). A readability formula that saves time. Journal of Reading, 11, 513–516, 575–578.

Fuchs, L., Fuchs, D., & Maxwell, L. (1988). The validity of informal reading comprehension measures. Remedial and Special Education, 9 20-28.

Green, G., & Davison, A. (Eds.). (1988). Linguistic complexity and text comprehension: Readability issues reconsidered. Hillsdale, NJ: Erlbaum.

Grisham, J. (1990). The firm. New York: Island Books.

Gray, W.S. (1920. The value of informal tests of reading achievement. Journal of Educational Research, 1, 103-111.

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, 489-508.

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

Hiebert, E.H. (April 2000a). The task of the first-grade texts: Have state policies influenced the content? Paper presented at the annual meeting of the American Educational Research Association, New Orleans, LA.

Hiebert, E.H. (September 2000b). What is third-grade reading? Paper presented at the meeting of the Southeast literacy consortium, Athens, GA.

Hiebert, E. H., & Calfee, R. C. (1992). Assessment of literacy: From standardized tests to portfolios. In A. E. Farstrup & S. J. Samuels (Eds.), What research has to say about reading instruction (2nd ed., pp. 70–100). Newark, DE: International Reading Association.

Hiebert, E.H., Menon, S., Martin, L.A., & Huxley, A. (April 2000). Teacher's guides as a scaffold for teaching young children to read: Curriculum, instruction, and guidance. Paper presented at the annual meeting of the American Educational Research Association, New Orleans, LA.

Hoffman, J. V., McCarthey, S. J., Abbott, J., Christian, C., Corman, L., Dressman, M., Elliot, B., Matheme, D., & Stahle, D. (1994). So what's new in the "new" basals? A focus on first grade. Journal of Reading Behavior, 26, 47–73.

International Reading Association and National Council of Teachers of English (1996). Standards for the English Language Arts. Newark, DE: IRA/NCTE.

Johnston, F.R. (2000). Word learning in predictable text. Journal of Educational Psychology, 92, 248-255.

Kamil, M., Mosenthal, P., Barr, R., & Pearson, P. D. (Eds.). (2000). Handbook of reading research (Vol. 3). Mahwah, NJ: Erlbaum.

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

Leslie, L., & Caldwell, J. (2001). Qualitative reading inventory-3. New York: Longman.

Leu, D. J. Jr., & Kinzer, C. K. (2000). The convergence of literacy instruction and networked technologies for information and communication. Reading Research Quarterly, 35, 108–127.

Lively, B., & Pressey, S. (1923). A method for measuring the “vocabulary burden” of textbooks. Educational Administration and Supervision, 99, 389–398.

Martin, L.A., & Hiebert, E.H (2001). TExT (Task Elements by Task) software (2nd Ed.). Santa Cruz, CA: TextProject.

Menon, S., & Hiebert, E. H. (1999). Literature anthologies: The task for first-grade readers (CIERA Report #1-009). Ann Arbor, MI: Center for the Improvement of Early Reading Achievement.

MetaMetrics (2000). The Lexile framework for reading. Durham, NC: Author. [Online]. Available: http:\\lexile.com\about_meta\press\21098b.htm.

National Assessment Governing Board(1994). Reading framework for the 1992 and 1994 National Assessment of Educational Progress. Washington, DC: U.S. Government Printing Office.

National Reading Panel. (2000). The report of the National Reading Panel. Washington, DC: U.S. Government Printing Office.

New Standards Primary Literacy Committee. (1999). Reading and writing grade by grade (CD-ROM & guidebook). Washington, DC: National Center on Education and the Economy; Pittsburgh, PA: University of Pittsburgh.

The New York Times (January 23, 2001). Text: President Bush and Education Secretary Paige (www.nytimes.com).

North Carolina Public Schools (2000). Education initiatives. [On-line]. Available: htttp://www.dpi.state.nc.usa.gov.

Pearson, P.D., Barr, R., Kamil, M., & Mosenthal, P. (1984). Handbook of reading research. New York: Longman.

Peterson, B. (1991). Selecting books for beginning readers: Children's literature suitable for young readers. In D. E. DeFord, C. A. Lyons, & G. S. Pinnell (Eds.), Bridges to literacy: Learning from Reading Recovery (pp. 119–147). Portsmouth, NH: Heinemann.

Pinnell, G. S., Pikulski, J. J., Wixson, K. K., Campbell, J. R., Gough, P. B., & Beatty, A. S. (1995). Listening to children read aloud: Data from NAEP's integrated reading performance record (IRPR) at grade 4. Washington, DC: Office of Educational Research and Improvement, U.S. Department of Education; Princeton, NJ: Educational Testing Service.

Pressley, M., El-Dinary, P. B., Gaskins, I., Schuder, T., Bergman, J., Almasi, J., & Brown, R. (1992). Beyond direct explanation: Transactional instruction of reading comprehension strategies. Elementary School Journal, 92, 513–555.

Smith, D., Stenner, A. J., Horabin, I., & Smith, M. (1989). The Lexile scale in theory and practice: Final report. Washington, DC: MetaMetrics. (ERIC document reproduction service number ED 307 577).

Snow, C., Burns, M. S, & Griffin, P. (1998). Preventing reading difficulties in young children. Washington, DC: National Research Council.

Spache, G.D. (1981). Diagnosing and correcting reading disabilities. Boston: Allyn & Bacon.

Stahl, S. (September 29, 2000). Participant comments in session What is third-grade reading? Conference of the Southeast literacy consortium, Athens, GA.

Stein, M., Johnson, B . , Boutry, S., & Borleson, C. (2000). An analysis of the decoding instruction in six first-grade reading programs. Tacoma, WA: University of Washington-Tacoma.

Stein, M.L., Johnson, B.J. & Gutlohn, L. (1999). Analyzing beginning reading programs: The relationship between decoding instruction and text. Remedial and Special Education, 20 (5), 275-287.

Texas Education Agency. (1990). Proclamation of the State Board of Education advertising for bids on textbooks. Austin, TX: Author.

Texas Education Agency. (1997). Proclamation of the State Board of Education advertising for bids on textbooks. Austin, TX: Author.

U.S. Census Bureau (2000). U.S. Census 2000. Washington, DC: Author (www.census.gov).

U.S. Department of Education (1997). Priorities of the President and Secretary of Education. Washington, DC: Author.

U. S. Department of Education. (1998) Goals 2000: Reforming education to improve student achievement. Washington, DC: Author.

Children's Books

Ackerman, K. (1994). The night crossing. New York: Alfred A. Knopf.

Adler, D.A. (1981). Cam Jansen and the mystery of the television dog. New York: Puffin Books.

Coerr, E. (1986). The Josefina story quilt. New York: Harper Trophy.

Hayes, S. (1988). Eat up, Gemma. New York: Mulberry Books.

Hurwitz, J. (1987). Class clown. New York: Scholastic.

Lewis, (1971). Hill of fire. New York: Harper Trophy.

Parish, P. (1963). Amelia Bedelia. New York: HarperCollins.

Rowling, J.K. (2000). Harry Potter and the goblet of fire. New York: Arthur A. Levine Books.

White, E.B. (1952). Charlotte's web. New York: Harper.

Table 1. Characteristics of Text Levels for End-of-Year Reading

Primary Category	Secondary Category	Level I (Grade 1)	Level L (Grade 2)	Level O (Grade 3)
Book and Print Features	Length	•From 30–40 pages; layout varies widely, including maps and charts	•From 70–80 pages, with chapters ranging from 5–15 pages	•Between 50–200 pages
	Illustrations	•Enhance meaning but provide little support for precise word solving and meaning	•Included but readers are less dependent on them	•Some black and white illustrations in the text
	Punctuation	{No description provided]	•Many conventions of text are introduced, including ellipses, italics, all capitals, indentations, and bold type	•Highly complex sentences with range of punctuation, which is often important to the meaning of text
	Layout & Font	•New sentences begin on left margin in some longer texts; others are signaled by clear spaces after a period or other ending punctuation within a line •Font size is smaller with more words per page	•Print size is varied and often small	[No description provided]
Content, Themes, and Ideas		•Content is more difficult	•Some stories have abstract or symbolic themes; books require a higher level of conceptual understanding	•Themes are sophisticated and require interpretation and understanding •Readers will experience the same themes through different genres •Previously read texts support interpretation of new texts.
Text Structure		•Mostly narrative, although shorter, informational books as well •Texts use dialogue, which is indicated by the identification of speakers and sometimes by spaces between speakers •One main plot with a solution; episodes or events are more highly elaborated and have multiple events; char-	•More sophisticated plots with characters that are developed throughout the texts and over longer periods of time •Events build on each other, requiring recall and tracking of information •More characters are speaking and dialogue is not always assigned •Readers are required to follow actions of several different	•Books have multiple characters who are developed through dialogue and actions, not simply through narrative

		acters and story events require interpretation	characters	
Language & Literary Features		•Complex word solving is required; texts have more multisyllabic words and these are embedded within longer sentences and paragraphs	•Vocabulary support may be required because of content specific, unfamiliar multisyllabic words	•Vocabulary is sophisticated and varied with many new, multisyllabic words; need to quickly analyze many new words, while focusing on meaning

Table 2. Characteristics of Exemplars

New Standards				California				Textbook I				Textbook II			
	Read	Lexile	CWF		Read	Lexile	CWF		Read	Lexile	CWF		Read	Lexile	CWF
Mieko & the Fifth treasure	4.1	680	13	Nana Upstairs & Nana Downstairs	3.5	640	5	Water Woman	2.8	430	9	How I Spent my Summer Vacation	2.0	450	6
The Patchwork Quilt	3.1	520	10	Chair for My Mother	2.3	640	8	Turtle Bay	2	430	6	Goldilocks	2.9	480	8
The Night Crossing	6.5	960	14	Sleeping Beauty by Brothers Grimm	6.8	1090	12	Balto, the Dog who saved Nome	3	490	8	Fly Traps	4.1	740	7
Class Clown	3.3	670	7	Courage of Sarah Noble	3.3	580	5	The Talent Show	2.3	560	10	Tornado Alert	4.2	640	9
Ramona Quimby, Age 8	5.5	860	13	The Ox-Cart Man	6.5	1220	12	Centerfield Ballhawk	2.3	490	11	Night of Pufflings	5.0	890	8
Henry & Beezus	6.4	730	10	Hill of Fire	1.5	350	4	Coyote Places the Stars	5.3	770	9	Flight	3.2	460	7
Beezus & Ramona	5.3	780	8	Fables	2.5	540	7	A Bookworm Who Hatched	5.9	800	13	More Than Anything Else	2.5	620	7
The Boxcar Children	1.8	440	5	Winnie the Pooh	5.5	760	5	Leah's Pony	2	560	10	Leah's Pony	2	560	10
				A Weed Is a Flower	3.4	610	9	Cocoa	6.3	1010	10	Mailing May	4.6	710	10
				Story of Johnny Appleseed	3.6	710	6	If You Made a Million	9.3	810	7	Floating Home	3.8	530	10
Mean	4.5	705	10	Mean	3.9	714	7	Mean	4.1	635	9	Mean	3.4	608	8

Table 3. Characteristics of Assessments[1]

		Complex Word Factor	Fry Readability	Lexile
Norm-Referenced Tests				
	CTBS	3	2.6	450
	SAT	4	2.8	630
State Tests				
	FCAT	7	2.0	550
	TAAS	5	2.5	560
Oral Reading Assessments				
	QRI	4	2.6	575
	DRA	6	3.5	650

[1] Based on third-grade curriculum of fluency with 1,000 most frequent words and single-syllable words.

[i] In their extension of the levels to sixth grade (Fountas & Pinnell, 2001), the category of content has been separated from theme and ideas and two aspects of language that were part of language and literary elements in Fountas and Pinnell (1999) now form their own categories: vocabulary and sentence complexity. For sentence complexity, the description cites length, embedded clauses, and punctuation as factors in text difficulty and vocabulary cites multisyllabic words and choice of words related to content. Anchors are not provided that illustrate the range of difficulty on any of these dimensions.