READING RESEARCH REPORT

#13.01
August 2013

The State of the Field: Qualitative Analyses of Text Complexity

P. David Pearson
University of California, Berkeley

Elfrieda H. Hiebert
TextProject, Inc.
University of California, Santa Cruz

TextProject

textproject.org

RRR-13.01 V.1.00 AUGUST 2013

The State of the Field: Qualitative Analyses of Text Complexity

P. David Pearson
University of California, Berkeley

Elfrieda H. Hiebert
TextProject, Inc.
University of California, Santa Cruz

Abstract

The purpose of this review is to examine the function, logic, and impact of qualitative systems, with a focus on understanding their benefits and imperfections. We identify two primary functions for their use: (a) to match texts to reader ability so that readers read books that are within their grasp and (b) to unearth, and then scaffold, those features of specific texts that are likely to present challenges for readers of differing abilities. We examine three approaches to qualitative text analysis (text leveling systems, rubric and exemplar systems, and text mapping systems) relative to these functions. We conclude by strongly advocating the use of qualitative systems, if for no other reason than to prevent the unchecked use of quantitative approaches from promoting invalid applications of text complexity. At the same time, we raise a set of vexing issues that the field must address if these approaches are to be used with even a modicum of confidence.

READING RESEARCH REPORT #13.01
August 2013

TextProject, Inc.
SANTA CRUZ, CALIFORNIA

Contents

The State of the Field: Qualitative Analyses of Text Complexity

Analyses of text complexity have been with us for as long as human beings have tried to communicate effectively and efficiently with one another in writing. Worrying about whether what we write, or say for that matter, can be understood by our intended audience is inherent to human communication. But it was not until the late 19th century that literary analysts (e.g., Sherman, 1893) and linguists (e.g., Rubakin, 1889, in Choldin, 1979) began the systematic study of text complexity and, in the process, developed tools for analyzing, predicting, and intentionally controlling text complexity in the study of written communication. These early analyses of text complexity were exclusively qualitative, in the sense that they focused on rich descriptions of the types of text features that would likely impact the comprehensibility or readability of texts (e.g., sentence length, obscure vocabulary, rare syntax). Sherman, for example, noted that literary texts of his late 1800s generation had average sentence lengths (23 words) less than half as long as those in the Elizabethan era (50 words). Choldin (1979) reports that the Russian scholar Rubakin found that strange words and long sentences were the two greatest blocks to clear communication in the late 1880s.

The year 1923 marks the appearance of the first published readability formula. Created by Lively and Pressey, its purpose was to predict how likely students of varying levels of reading development (i.e., the sharpness of their skills, processes, and knowledge) would be able to successfully read and understand texts of increasing difficulty. The Lively and Pressey work is noteworthy for the almost exclusive reliance on word-level factors to predict students' ability to understand texts.

It would take another decade for more comprehensive readability investigations to emerge: Gray and Leary (1935) performed the "classic" earlier readability study. They began with 82 potential formal factors that might conceivably predict a text's readability. They included mundane factors such as sentence length and word frequency, but also added some more sophisticated indices, such as the ratio of prepositions (*of*, *for*, *on*) to conjunctive adverbs (*because*, *since*, *although*, *if*, *unless*). They concluded that 44 of the factors, both mundane and sophisticated, were significantly related to reading difficulty, as measured by comprehension questions tied to a set of graded passages (McCall & Crabbs,

1926/1979). In an even more exhaustive investigation of formal linguistic factors using cloze tests (fill in a blank left for every 5th word) as a dependent measure, Bormuth (1966) found over 60 structural indices that were useful in predicting comprehension difficulty. In the 30 years between the comprehensive efforts of Gray and Leary and Bormuth, nearly 100 different readability formulas came and went, with about 20 attracting enough of a following to be part of the readability toolbox available to classroom teachers and text researchers (see Klare, 1963, 1984).

When Klare reviewed readability research in 1984, he made a clear distinction between using readability formulas to forecast comprehension and using them as a guide to editing text to shape comprehension, usually by rewriting passages to give them a lower readability score. He was careful to point out that readability scores were, in general, correlated with comprehension scores, but that lower scores did not necessarily cause better comprehension. He also emphasized that readability formulas had only been validated for purposes of prediction. As such, they could give a teacher a general guide about the likelihood that a given book would be suitable for a given student or class. But whether the book would be a match would turn on a host of other factors, including knowledge, interest, and purpose in reading the book in the first place. We mention these cautionary notes because they turn up again and again in the history of measuring text complexity: (a) Does linguistic complexity produce barriers to comprehension or simply reflect the complexity of the ideas the language represents? (b) How precisely can a readability score predict a reader-book match for a given student or class?

In conjunction with the cognitive turn in psychology (Gardner, 1987), attention turned away from predicting reading difficulty and toward understanding the roles particular text features played in readers' cognitive processing of information. During this era, various analyses of text structure at both the micro (sentence level) and macro (paragraphs and rhetorical structures) were invoked to explain text comprehension. In general (see Pearson & Camparell, 1981 for a review), both narrative (story grammar elements) and expository (e.g., rhetorical structures such as conflict-resolution or problem-solution frames) features influenced comprehension; moreover teaching students how to exploit text structure during reading had a generally positive influence on text comprehension—a finding that has endured for over 30 years (Shanahan, Callison, Carriere, Duke, Pearson, Schatschneider, & Torgesen, 2010).

But the big ideas in the 1970s and 1980s were situated in the reader, not the text. Knowledge of ideas, represented as schemata in long-term memory, along with executive control processes (Brown, Armbruster, & Baker, 1986), dominated research on comprehension. The emphasis in comprehension studies was on synthesis (how humans integrate separate inputs into a coherent whole) rather than on analysis (how we deconstruct units to examine their infrastructure). This was not a comfortable context for a construct such as readability, with its

insistence that long words and long sentences were predictive of or perhaps even shaped, comprehension performance.

Linguists (e.g., Davison & Kantor, 1982) conducted analyses to demonstrate how, when writers of instructional materials for children try to simplify prose by breaking longer, grammatically more complex sentences into shorter, grammatically simpler ones, they often burden the reader with extra inferential tasks. Contrast these examples:

> ***Original:*** If given a chance before another fire comes, the tree will heal its own wounds by growing new bark over the burned part.
>
> ***Adapted:*** If given a chance before another fire comes, the tree will heal its own wounds. It will grow new bark over the burned part. (Davison & Kantor, 1982, p. 192)

Note that in the adapted version, which would contribute to a lower readability score, the reader has to infer that growing new bark over the burned part is the causal mechanism for healing: What was explicit in the original is implicit in the adaptation. Now there is a trade-off: In the original version, the reader confronts a greater short-term memory load in processing a single long and complex set of clauses. So the question is which processing load—short-term memory or inferential reasoning—trumps the other? After systematically reviewing several attempts to use readability formulas to adapt adult texts for younger readers, Davison and Kantor concluded that "… the most successful changes in the text often run directly counter to what readability formulas would suggest, and that the most unsuccessful changes are those motivated by the strictures of the readability formulas." (p. 191)

Other scholars in this period (e.g., Blau, 1982; Pearson, 1974–75) conducted empirical research that confirmed the negative impact of readability formula driven adaptations on comprehension. The combination of linguistic critique and empirical evidence brought readability formulas into question. One plausible explanation, which echoes Klare's (1984) concern, put forward by several of these scholars, was that readability formulas reflected rather than defined readability (e.g., Davison & Kantor, 1982; Pearson, 1974–75). The fundamental premise of this position is that it is the complexity of the ideas they are trying to communicate that drives writers to craft prose that ends up being grammatically and semantically more complex—that is, expressed in longer, more complex, and likely more obscure, words and sentences. In short, there is a reason for complexity—the ideas are difficult!

By 1984, these critiques had reached a point where, in the highly influential National Academy of Education report, *Becoming a Nation of Readers*, Anderson, Hiebert, Scott, and Wilkinson took the position that readability formulas, while they give educators a general guide to text placement, need to be supplemented with more qualitative analyses of text:

> In summary, readability formulas are useful as a first check on the difficulty and appropriateness of books. However, no formula gauges the clarity, coherence, organization, interest, literary quality, or subject matter adequacy of books. Inevitably, overreliance on readability formulas by the schools and their misuse by the publishing industry has contributed to bad writing in school books. The Commission urges those who buy books and those who write and edit them to supplement analyses using readability formulas with analyses of the deeper factors that are essential for quality. (p, 65).

Partially in response to these questions about the limitations of readability formulas and partially because experience taught them that readability formulas are especially inaccurate at the earliest stages of reading development, other scholars (e.g., Carver, 1976; Clay, 1991; Singer, 1975), began to adopt alternative approaches to scale books, particularly for use in instruction. The fundamental procedure of these approaches was to assemble expert teachers to examine and then judge the developmental level at which texts could be used. One particular approach—text leveling—grew in popularity in the 1980s and 1990s, until it had become the driving force influencing most major publishers of books for young readers.

Thus the linguistic critique of readability, accompanied by some empirical evidence that contradicted the predictions of difficulty emanating from readability formulas, and several forays into systems that rely on human judgment to scale difficulty formed the basis of a qualitative turn in the analysis of text complexity. This is the topic to which we now turn.

Qualitative Systems

If the essence of a qualitative system is the use of human judgment (National Governors Association (NGA), Center for Best Practices & Council of Chief State School Officers (CCSSO), Appendix A, 2010), then qualitative systems are not new as tools to help educators determine appropriate texts for use in instruction. The use of templates to control our subjectivity in making judgments is a fundamental tool of decision-making in many human endeavors. It is no different when it comes to matching books to readers. Clearly teachers, librarians, parents, teacher educators—as well as children—have used templates or prototypes for choosing books for generations. Writers such as Trelease (2006) identify recommendations for particular grade levels, as does Hirsch (2005a, 2005b). Lists of recommendations for different grade levels can be found at a variety of websites (Fountas & Pinnell, 2009; http://www.fountasandpinnell-leveledbooks.com/, http://www.scholastic.com/bookwizard/).

Whether texts designated by these means provide too much challenge or not enough is uncertain. To our knowledge, no one has conducted a direct validation of any of these leveling systems to determine whether the texts assigned to a level provide just the right challenge for students judged (or more likely assumed) to be reading at that level. In large measure, those who create and implement these systems are more likely to use anecdotal classroom reports of

their success in matching students to books than any sort of careful analysis of student reading performance (e.g., decoding accuracy, fluency, or comprehension).

To operationalize any system of human judgment that aspires to match books to students, two estimates are needed: (a) an estimate of the level (often operationalized as a grade level) at which a given student can read, and (b) an estimate, hopefully on the same scale as the student scores, of the level of difficulty of a large number of books. Find the level of the readers and let them select from books judged to be at their independent reading levels. That's the logic.

Given that our topic is systems for scaling books, we will not dwell on systems for determining the level at which students can read a given book, either on their own (independent level) or with teacher and peer support (instructional level), except to say that there is a long and complicated literature on the topic, mostly conducted in the spirit of validating and rationalizing informal reading inventories (Betts, 1946; Pikulski & Shanahan, 1982) and their commercial counterparts (Beaver, 2003; Johns, 2012; Leslie & Caldwell, 2010). Our focus is on qualitative approaches for scaling texts, most often to allow teachers to match them to their students' current reading capacities, but also to provide teachers with insights that might help them in teaching particular texts.

The three qualitative systems we review in this section distinguish themselves from the informal approaches to text leveling that have emerged from grass roots efforts (e.g., Rog & Burton, 2001) in that they are more systematic in describing, and/or analyzing, and/or validating their criteria and procedures. All three approaches have been described in published documents, although—as we indicated—precious little is known about the validity of the text assignments in relation to measures of student reading (e.g., accuracy, fluency or comprehension) or to teachers' efficacy in providing appropriate instruction, with the notable exception of the work by Hoffman, Roser, Patterson, Salas, and Pennington (2001), which we will look at shortly. The first approach—text leveling (TL)—is used extensively in school contexts and is described in the pedagogical literature (e.g., Fountas & Pinnell, 1996, 2009; Peterson, 1991). The second approach—rubrics plus exemplars (R+E)—is the one promoted within the CCSS and used in several prior efforts (e.g., ACT, 2006). A third approach—text maps (TM)—is used by the National Assessment of Educational Progress (NAEP; AIR, 2008)—and was in use in several state assessments in the 1980s and 1990s to determine the critical content of a text (Valencia, Pearson, Peters, & Wixson, 1989; Wixson, Peters, Weber, & Roeber, 1987).

TL and R+E are similar in that both rely on two key elements: (a) the use of criteria for describing and rating text complexity and (b) the use of exemplar texts to "anchor" what it would mean to read at different levels within the system. The aims of the two systems are sufficiently unique, however, that we treat them separately. In TL systems, the primary goal is to provide teachers with a vetted level for a text that corresponds to students' reading levels. The major aim of the R+E systems, which are prominently represented in today's world of

the CCSS (Appendix A, 2010), is to involve teachers in identifying text features that can promote (or impede) their students' capacities to read more complex text, rather than on assigning a specific level to a text.

TL systems tell us who ought to be able to read a particular book, either on their own (independent level) or with help (instructional level); R+E systems tell us what adjustments (scaffolds and supports) a teacher might have to make in a given classroom to help a range of students work their way through the text, with or without teacher guidance. Another way to characterize the distinction is that the TL systems are more text-centric, while the R+E systems are more reader-centric in their end goals. It will be important to keep these distinct purposes in mind as we review these qualitative systems.

Text Leveling

The leveling of texts by expert judges is not a recent phenomenon (see, e.g., Carver, 1976; Singer, 1975). However, this procedure was not prominent until readability formulas were downplayed as a criterion for textbook selection in America's largest states (California English/Language Arts Committee, 1987; Texas Education Agency, 1990). The Reading Recovery levels (Peterson, 1991) that have evolved into guided reading levels (Fountas & Pinnell, 1996, 2001) were a response to this change in focus.

Reading Recovery and guided reading levels. The first systematic attempt at implementing a wide-scale text-leveling scheme was Peterson's (1988) dissertation research at Ohio State University on Reading Recovery books. Peterson started with the books that were in use as exemplars of Reading Recovery levels in New Zealand at that time. The degree to which Peterson's work resulted in changes in assigned RR levels is not certain, but she did identify four criteria that distinguished among books judged to be at different levels: (1) book and print features; (2) content, themes, and ideas; (3) text structure; and (4) language and literary elements. Descriptions were written to show how the features differed from level to level, but the features themselves were not analyzed as separate components. Sample texts that exemplify particular levels were provided, but the details of how and why these texts illustrate particular features at particular levels were not specified.

A decade after Peterson's (1988) work on Reading Recovery levels, Fountas and Pinnell (1999, 2001) applied the leveling system within the context of their approach to guided reading. Their system was similar to Reading Recovery levels, but Fountas and Pinnell used a 26-level (as compared to 20-level) scale that extended to sixth grade. Their criteria for scaling books include the same content foci (book and print features, content themes and ideas, text structure, and language and literary elements), but they divide the categories in a different way. Content is comprised of two separate scales: concepts and theme and ideas. Language and literary elements, a composite category in Reading Recovery, is divided into two categories: vocabulary and sentence complexity.

The Fountas and Pinnell process of evaluation, however, is the same as in Reading Recovery. A rater uses the descriptions of levels to assign a book to a level, under the untested assumption that the steps between levels for any of the key traits changed by roughly the same amount from level to level. In essence it operates like a holistic writing rubric; that is, a judge might examine a text on several different dimensions, but then amalgamate all of that feature-by-feature information to reach a judgment that the text should be assigned to a particular level. Scores or levels are not reported for individual categories (e.g., content, text structure); instead, the different categories or scales inform the holistic rating.

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 or different levels of readers. For example, print features might be expected to weigh more heavily at the very early levels such as A through E, but a variable such as referential cohesion or syntactic simplicity might dominate at levels V through Z. Indeed, in a recent revision of what they call the F & P Text Gradient, Fountas and Pinnell (2012) have moved away from providing any description of the form of a trait (or factor, in their presentation) at any level. Ten factors are now given—genre, text structure, content, themes and ideas, language and literary features, sentence complexity, vocabulary, words, illustrations, and book and print features. For each factor, a statement such as the following for genre is given: "The genre is the type of text and refers to a system by which fiction and nonfiction texts are defined. Each genre has characteristic features." (Fountas & Pinnell, 2012, p. 4).

The role of individual variables, it would seem, have been subsumed into a holistic rating. Holistic scoring may obscure between-criterion variability; it would not be hard to imagine a text that was at level T on vocabulary but only at level M on structural elements. Graesser, McNamara, & Kulokovich (2011), using more quantitative analyses for five separate linguistic elements, found that texts judged to reside at a particular level of readability can vary widely on an array of specific elements of text complexity.

Publishers and educators have applied the text leveling of Reading Recovery and guided reading to literally thousands of texts. Despite its widespread use, we were unable to find any reports of reliability across coders in leveling texts for either scheme. Further, while proponents of this form of leveling present it as an alternative to readability formulas, one of the only studies of its validity (Hatcher, 2000) has reported a strong correlation (r=.82) between text levels within Reading Recovery and the principal factors that make up traditional readability formulas (word frequency and sentence length).

We could find no studies that examined how instruction with texts ordered according to either Reading Recovery or guided reading levels influenced reading acquisition. We located a single study (Hoffman et al., 2001, reviewed next) that examined student performance on texts at different Reading Recovery and guided reading levels

Scale for Text Accessibility and Support (STAS-1). Similar to guided reading levels, the STAS-1 (Hoffman et al., 2001) uses expert judgment in the form of ratings on two separate five-point scales to evaluate text complexity. Unlike the holistic scores of guided reading levels, levels on the STAS-1 are a product of independent ratings of several raters on two criteria—decodability and predictability. Hoffman and his associates used a methodology (Carver, 1975; Singer, 1981) in which experts use anchor passages that have been previously ordered according to specific criteria. For example, on the decodability criteria, texts rated as highly decodable (1 on the scale) contain words with Consonant-Vowel-Consonant (CVC) patterns, single syllables, and short high frequency words, while minimally decodable texts (rated as 5) contain irregularly spelled words and a variety of vowel patterns. In between these end points are three interim points: (2) very decodable, (3) decodable, and (4) somewhat decodable. A comparable five-point scale used four predictable features (picture support, repetition, rhyming elements, and familiar events/concepts) to guide raters. Hoffman et al. reported that, on the basis of 21 texts (three texts from each of seven earliest of Reading Recovery levels), the average correlation between ratings of different judges correlated at .78.

Hoffman et al. (2001) examined how well the STAS-1, Reading Recovery levels, and guided reading levels of texts were able to predict student accuracy, fluency, and rate across three instructional conditions (text preview, word preview, and no preview). All three TL systems yielded modest but statistically significant correlations with accuracy, fluency, and rate metrics in the .2 to .4 range, with the consistent predictive advantage going to STAS-1 over the two leveling systems. Significant but unsurprising effects were found for reader ability. More interesting effects were found for the three conditions of support; those students who received adult modeling in the form of a text preview or a sight word preview achieved higher levels of performance on fluency and accuracy indices than students in the "no preview" condition, implying that teacher scaffolding exerts positive influences on typical early reading tasks.

The work of Hoffman et al. (2001) illustrates that particular dimensions of texts can be defined and that raters, when given clear criteria, can sort a group of texts reliably on a recognized trait of beginning reading texts, such as decodability or predictability. A side effect of the Hoffman work was that it also validated the leveling system of Fountas and Pinnell and Reading Recovery, which turned out to be almost, but not quite, as predictive of reading performance on common reading tasks (accuracy and fluency) as the STAS-1 system. Apparently, there is something in the combination of the two scales of decodability and predictability that is not captured by either readability formulas or impressionistic professional judgment.

Rubrics and Exemplars (R+E)

Over a period of nearly half a century, professionals have used two fundamental tools—rubrics and anchors—to score student writing (DiPardo, Storms, &

Selland, 2011). In the rubric/anchor system, human judges identify a set of traits that characterize effective products, usually by examining artifacts that vary widely in holistic/impressionistic judgments of quality. These traits are placed on a continuum where less mature forms of the trait anchor one end and more sophisticated forms, the other. Each trait and its manifestations across the continuum are described as a rubric. The anchor metaphor is significant: Examples of student work that typify key levels or points along the continuum are often referred to as "anchor papers." The logic of the system (rubrics, along with anchoring exemplars) has been applied to a host of phenomena in which human judgment is involved in scoring or ranking performance other than in writing: debates, speeches, athletic events, and even university applications.

In adapting the logic of the writing rubric model to text analysis, the operative term has been exemplars rather than anchors. But procedures have been similar: identify important traits, develop descriptions that position levels of those traits along a set of continua, and locate anchor texts that typify points along those continua, or, in the case of holistic rubrics, a continuum.

Using the writing scoring systems as models, several research groups have worked with teachers in using rubrics or sets of exemplars for choosing books. For example, part of a staff development project on teacher-based assessment conducted through the National Center for Research on Evaluation, Standards, and Student Testing (CRESST) used an exemplar system in which teachers sorted books to identify a shared set that were then used to determine students' placement for instruction (Shepard, Flexer, Hiebert, Marion, Mayfield, & Weston, 1996).

Similarly, Hess and Biggam (2004) developed a fairly complex text analysis scheme that has been used into engage teachers in systematic text analysis during professional development. Their scheme consisted of seven features (word difficulty and language structure, text structure and discourse style, features of genre/text type, background knowledge and/or degree of familiarity, level of reasoning, format and layout, and length of text). In revising the system on the basis of feedback during its use, Hess and Hervey (2010) provided separate rubrics for narrative and informational texts at the same time as they reduced the system to five, rather than seven traits, with each trait presented along a continuum of simple, somewhat complex, complex, and very complex.

The Qualitative Assessment of Text Difficulty (QATD). The first qualitative system available through an academic publisher was the QATD (Chall, Bissex, Conard, & Harris-Sharples, 1999). Chall and her colleagues built four scales—each specific to one of the four major "content areas" in the system—literature, popular fiction, science, and social studies. Each scale describes not the features of the texts at different levels of complexity but rather the processes in which readers need to engage to be successful at successive levels of text difficulty. Each scale has four to five primary traits, as presented in Table 1. Each trait is unpacked for particular grade levels, as illustrated in Table 2. Each scale is anchored by a set of benchmark texts, one for each developmental level on the

TABLE 1

Comparison of Rubrics: CCSS, ACT, & QATD

CCSS		ACT	QATD			
Narrative	**Informational**		**Popular Fiction**	**Literature**	**Social Studies**	**Science**
Levels of meaning		Relationships		Skill in literary analysis	Levels of reasoning Cognitive density	Levels of reasoning Density of ideas
	Purpose	Purpose				
Structure	Structure	Structure	Familiarity with sentence structure	Familiarity with sentence structure	Familiarity with sentence structure	Familiarity with sentence structure
Language conventionality & clarity	Language conventionality & clarity	Vocabulary & style	Knowledge of vocabulary	Knowledge of vocabulary	Knowledge of vocabulary	Knowledge of vocabulary
Knowledge demands: life experiences; cultural/literary knowledge	Content/ discipline knowledge	Richness	Cultural & literary knowledge Depth & breadth of experiences	Cultural & literary knowledge Depth & breadth of experiences	Subject-related & general knowledge	Subject-related & general knowledge

scale. The science and the social studies scales are each represented by two sets of anchor texts, reflecting some concessions to both sub-discipline (life science and physical science for the sciences) and genre (narrative and expository accounts for history/social studies). These distinctions reveal the concern that Chall and her colleagues held for conceptual content, not simply linguistic features. Further, the distinction in the QATD between literature and popular fiction suggests the recognition of Chall and her colleagues that content is a primary source of complexity (i.e., literature has complex content; popular fiction has less complicated and easier content).

The relationship between the rubrics and the selection of the exemplars is not explicit in the approach of Chall et al (1999). That is, the ways in which the exemplars manifest the traits/rubrics is neither described nor transparent. Instead of asking raters to make their decisions about levels based solely on traits, they rely more on the aptness of the exemplars. Chall and her colleagues drew on what they described as a large number of texts to identify the exemplars. These exemplars (but not the rubrics) were validated through inter-rater agreement procedures among the researchers, several groups of teachers and reading specialists, and readability formulas.

The major contribution of the Chall et al. approach to qualitative analysis is to remind us that not all texts demand the same sort of cognitive and linguistic processing—that subject matter demands (science versus literature versus history) necessarily shape the ways in which readers engage the text, which connects her system intimately with the disciplinary grounding of the CCSS.

Illustrations of Traits in the Three Rubric/Exemplar Systems: Language Conventionality & Clarity/Vocabulary

TABLE 2
QATD's Knowledge of Vocabulary for Literature

Reading Level	
1–2	Mainly familiar words, often repeated
3–6	More varied, but generally familiar; some figurative language
7–12	Increasing number of uncommon words; nonliteral meanings
13–15 (College)	Wide vocabulary and range of meaning levels

TABLE 3
ACT's Vocabulary

Type of Text	
Uncomplicated	Familiar
More Challenging	Some difficult, context-dependent words
Complex	Demanding, highly context dependent

TABLE 4
CCSS's Language Conventionality & Clarity

Literal	Figurative
Clear	Ambiguous or purposefully misleading
Contemporary, familiar	Archaic or otherwise unknown
Conversational	General academic and domain-specific

ACT: Reading between the Lines. In commenting on college readiness (or the lack thereof) of high school students, ACT (2006) concluded that it was not the level of questions they were asked but the complexity of the text they were required to read that sorted students into levels of preparedness for college; in short, text mattered more than task, at least insofar as they adequately measured task performances. They identified three kinds of texts—uncomplicated, more challenging, and complex. These three levels of texts were differentiated on the basis of five traits that ACT writers described with the mnemonic RSVP: R: Relationships, Richness; S: Structure, Style; V: Vocabulary; and P: Purpose. In contrast to Chall's group, ACT scholars did not develop separate scales for disciplines, genres, or even broad categories, like expository versus narrative texts.

The process of rubric development is not described within the ACT report, nor does the report give any information on the scoring/sorting—who did the scoring or what the particular ratings were on the traits that make up the rubric. Nor does it provide evidence about the weighting that was assigned to particular elements of the rubric in determining the complexity of particular passages. Even so, it remains, in our estimation, the most interesting of all the qualitative systems, largely because of its commitment to close analysis of the particular features that render particular texts more or less accessible (see Table 3).

The ACT report does not provide exemplars per se but instead offers annotated versions of texts that represent two of the three levels (complex and more challenging texts). Each annotation begins with the selection's theme—a summary of the critical content of the passage—and goes on to describe the features of the text that account for the challenges students may experience when reading and answering questions about it. A portion of an annotation illustrating

FIGURE 1

Annotation of a Complex Text Representing Prose Fiction: ACT (2006, p. 18)

This text describes two complex, well-developed characters, Sunday and Delta, and their strained yet loving relationship. One factor that contributes to the complexity of the text is its structure: the third-person narrator presents the two sisters both as they see themselves and how each sees the other.

PROSE FICTION: This passage is adapted form the novel *Night Water* by Helen Elaine Lee (© 1996 by Helen Elaine Lee).

There had been no words for naming when she was born. She was "Girl Owens" on the stamped paper that certified her birthday, and at home, she had just been "Sister," that was all. When asked to decide at six, what she would be called, she had chosen "Sunday," the time of voices, lifted in praise.

That was one piece of the story, but other parts had gone unspoken, and some had been buried, but were not at rest. She was headed back to claim them, as she had taken her name.

VOCABULARY: Beginning with the opening sentence—"There had been no words for naming when she was born"—the text uses fairly sophisticated syntax.

a complex text within the prose fiction category is presented in Figure 1. Even with only a small portion of what is a small excerpt of a novel, *Night Water* (Lee, 1999), the material presented in Figure 1 shows that the goal of the annotation is to convey information about the ways in which text features influence readers' meaning-making, rather than descriptions of the text features per se. For a teacher making instructional decisions about readers and tasks, the annotation provides information about which features of the text may create obstacles for students and which could be the focus of instruction that develops student capacity with a particular type of complex text. Of all the systems, this one shows the most potential to provide direction for teachers about how to scaffold texts that challenge to students.

Common Core State Standards (CCSS) and its extensions. The qualitative system within the CCSS (NGA Center for Best Practices & CCSSO, Appendix A, 2010) is a hybrid of the qualitative systems described thus far, but it relies, by its own admission, more on the ACT system than any of the others. For each grade band within the Standards themselves, five texts are presented for literature (stories, drama, poetry) and five for informational texts (literary nonfiction and historical, scientific, and technical texts) as "illustrating the complexity, quality, and range of student reading" (p. 32) for a grade band. The list of exemplars is expanded in Appendix A to approximately a dozen per type (i.e., literary, informational), and each type is broken into additional categories. Literature is divided into genres (e.g., stories, poetry, drama), beginning in the grade 4–5 band, and informational texts are organize by content area (i.e., English/Language Arts, History/Social Studies, Science, Mathematics, and

Technical Subjects), beginning with the grade 6–8 band. The entire group of texts for a grade-band/category group is presented without differentiation as appropriateness for positions within the band; thus for the 2–3 band for stories, there is no indication of which stories might be used early in grade 2 versus late in grade 3. Given this lack of within-band differentiation, it is not surprising to learn that within, for example, the grade 2–3 band, the exemplar texts differ considerably from their placement in basal reading programs. *Poppleton in Winter* (Rylant, 2008) appears in the grade 2–3 list as does *Sarah, Plain and Tall* (MacLachan, 1987). The former currently appears in a first-grade core reading programs (Beck, Farr, Strickland, Ada, Hudson, McKeown, Scarcella, & Washington, 2008), while the latter has appeared on lists for fourth—and even fifth—graders (Hollingsworth, 1991).

The CCSS recommends a tri-partite system of assessing text complexity (quantitative, qualitative, and reader/task considerations), including explicit quantitative guidelines in both Appendix A (CCSS, 2010) and its supplement (CCSS, 2012). Whereas Chall et al. (1999) used quantitative indices to ensure that the exemplars were assigned to the appropriate grade levels, CCSS developers did not report quantitative information on the exemplars in the original presentation of the Standards (though they did in the supplement to Appendix A). Thus, a text for the K–1 level (*A Weed is a Flower*) has a measured readability score approximately 2.5 grade levels beyond the assigned level. And several texts with readability scores that place them within the 2–3 grade band (e.g., Steinbeck's *Travels with Charley*) truly belong in middle or high school because of their themes and genre demands.

In a study reported as a CCSS supplement, Nelson, Perfetti, Liben, and Liben (2012) provide evidence on the relative efficacy of Lexiles and five additional quantitative indicators of difficulty to predict the grade band at which a text was placed in the CCSS exemplars. The correlations between readability scores and narrative texts' grade-band placement across the six text analysis systems averaged r=.47, with a substantial range—a low of .29 (Lexiles) to a high of .62 (Source-Rater). Correlations were higher for the informational exemplars: r=.66, with a somewhat smaller range—.53 (DRP) to .8 (Source-Rater).

The rationale behind text classifications in particular grade bands, it would be expected, can be explained with elaborate annotations (as was done with the ACT system) of the how the features of the rubrics either do or don't apply to particular texts. CCSS developers did not take this route. Rather, they provided examples of evaluations using the three assessment systems (i.e., quantitative, qualitative, and reader-task) for three of the 168 exemplar texts—two from the grade 9–10 band and one for the grade 6–8 band. The evaluation of one text—*Grapes of Wrath*—is reprinted in Figure 2. The purpose of this evaluation appears to be to justify the placement of a text in a particular grade band, rather than to provide information that might aid teachers designing lessons that might increase students' capacity to read increasingly more complex texts.

FIGURE 2

Annotation for *The Grapes of Wrath* (CCSS, 2010)

QUALITATIVE MEASURES	QUANTITATIVE MEASURES
Levels of Meaning There are multiple and often implicit levels of meaning within the excerpt and the novel as a whole. The surface level focuses on the literal journey of the Joads, but the novel also works on metaphorical and philosophical levels. ***Structure*** The text is relatively simple, explicit, and conventional in form. Events are largely related in chronological order. ***Language Conventionality & Clarity*** Although the language used is generally familiar, clear, and conversational, the dialect of the characters may pose a challenge for some readers. Steinbeck also puts a great deal of weight on certain less familiar words, such as faltering. In various portions of the novel not fully rep- resented in the excerpt, the author combines rich, vivid, and detailed description with an economy of words that requires heavy inferencing. ***Knowledge Demands*** The themes are sophisticated. The experiences and per- spective conveyed will be different from those of many students. Knowledge of the Great Depression, the "Okie Migration" to California, and the religion and music of the migrants is helpful, but the author himself provides much of the context needed for comprehension.	The quantitative assessment of *The Grapes of Wrath* demonstrates the difficulty many currently existing readability measures have in capturing adequately the richness of sophisticated works of literature, as various ratings suggest a placement within the grades 2–3 text complexity band. A Coh-Metrix analysis also tends to suggest the text is an easy one since the syntax is uncomplicated and the author uses a conventional story structure and only a moderate number of abstract words. (The analysis does indicate, however, that a great deal of inferencing will be required to interpret and connect the text's words, sentences, and central ideas.) **READER-TASK CONSIDERATIONS** These are to be determined locally with reference to such variables as a student's motivation, knowledge, and experiences as well as purpose and the complexity of the task assigned and the questions posed. **RECOMMENDED PLACEMENT** Though considered extremely easy by many quantitative measures, *The Grapes of Wrath* has a sophistication of theme and content that makes it more suitable for early high school (grades 9–10), which is where the Standards have placed it. In this case, qualitative measures have overruled the quantitative measures.

Kansas system. The text complexity model proposed within the CCSS has been adapted, extended, and applied in the Kansas Qualitative Measures Resources (Copeland, Lakin, & Shaw, 2012). The qualitative assessment is embedded within a four-step process that integrates quantitative and reader-task analyses in a deliberately hybridized approach. The first step of the process involves obtaining quantitative measures of the text. The second step is to ana-

TABLE 5

Qualitative Measures Resources (Kansas Adaptation of Common Core Qualitative Rubric)

High	Middle High	Middle Low	Low
Implicit or inferred meaning, heavy use of figurative or ironic language, may be purposefully ambiguous or misleading at times	Some implicit or inferred meaning, use of figurative or ironic language	Largely explicit and literal meaning, subtle use of figurative or ironic language	Explicit and literal meaning, little or no use of figurative or ironic language

lyze the qualitative measures of the text. The qualitative rubric has four rather than two levels of each of the basic traits from the CCSS, which are labeled as high, middle high, middle low, and low. Language conventionality and clarity has been broken into two sub-traits: meaning and register, with no specific attention to vocabulary. The content of the meaning sub-trait is provided in Table 5. Finally, the suggestion is made that reviewers attend to reader and task considerations (third step). All of this information is combined in the fourth step, which culminates in a recommendation for placement in the appropriate text complexity band of the CCSS.

This model appears to have generated considerable interest. For example, the Model Content Frameworks developed by the Partnership for Assessment of Readiness for College and Careers (PARCC; PARCC, 2012) assessment consortium has suggested a similar procedure. As another example, the state of Georgia (Georgia Department of Education, 2012) has used the Kansas rubric but added a quantitative overlay. They embed a 10-point scale into the Kansas categories of low, moderate, and high: 1–3 points for low, 4–6 for moderate, 7–10 for high. These adaptations, however, are variations on the CCSS qualitative theme; what remains consistent across the adaptions is the basic four-part rubric and reliance on human judgment in applying the rubric to texts to achieve both a grade-band placement and, where appropriate, an account of the peculiar difficulties that particular texts present.

Achieve the core qualitative rubric. A second adaptation of the qualitative rubric of the CCSS has been added to the website at http://achievethecore.com (Student Achievement Partners, 2012), the resource site for Student Achievement Partners, the agency that held the contract for the writing of the Common Core State Standards. The rubric itself is similar to the one in Appendix A (p. 6). The presentation, however, has been modified to include a place for reviewers to identify which trait trumped all the others in a judge's decision to place the text in a given grade band. Reviewers are also asked to assign an instructional and an independent level to the text. The emphasis is on the placement to ensure designation of a single level, not on the content to be taught or the peculiar difficulties of a given text.

Text Maps

Text maps depart radically from both text leveling (TL) and rubric plus exemplars (R+E) systems. In text maps (TM), the focus is on the conceptual structure of the text; for either narratives or informational texts, the result of text mapping is a diagram of the text. For stories, it most often resembles a flow chart of the sort popular in the story mapping (e.g., Pearson, 1984) and story grammar analyses (e.g., Stein & Glenn, 1979) popular at the height of the cognitive revolution in the late 1970s and early 1980s. For informational texts, the diagrams tend to be more elaborate and complex, with multiple nodes and branches representing the various semantic networks of ideas within the text.

Text mapping has been used within the NAEP since the late 1980s to (a) ensure that texts have sufficient conceptual grist for inclusion in the assessment and (b) to ensure that the items developed for NAEP passages assess all of the important content and focus on the higher level nodes in these elaborate semantic networks. As we have found for so many of these the qualitative systems described in this review, this procedure has not been examined in enough detail and with enough scrutiny to have yielded analyses that have found their way into archival sources.

The specifications for the procedure, however, are extensive (American Institutes for Research (AIR), 2008). Internal documentation of the procedure by contractors is presumably extensive as well, although we could not obtain such documentation for this review. The NAEP appears to have used text maps in item creation since the 1992 NAEP (National Assessment Governing Board, 1991), after reports of the successful experiences of two states—Illinois and Michigan—in using these maps for their state assessments (Valencia et al., 1989; Wixson et al., 1987).

The essential move in text mapping is examining the ideational structure of the text by focusing on the key ideas and displaying them visually in some sort of diagram that highlights the relationships among those key ideas. Protocols for literary and informational texts are different because of differences between the two text types. Evaluators discuss their maps with one another at key points to ensure fidelity in representing key ideas and relationships. Discussion occurs before item development as well as after to ensure fidelity between the maps and the items and the scoring (rubrics) procedures for short constructed response and extended constructed response items.

Whether mapping narrative or non-narrative texts, mapping begins with a thorough reading of the text, followed by summarizing the theme (narrative) or the purpose (non-narrative) of the selection. After the shared processes of reading the text and writing a concise but comprehensive summary of theme/purpose, the protocols for narrative and non-narrative take different forms, reflecting the different content of the two text types.

Narrative maps are used for literary texts with plots (i.e., some form of problem, conflict, resolution), including tales, mysteries, and realistic and historical

FIGURE 3

Portions of a Narrative Text Map for Eighth Grade: *Thank you, M'am* (from 2009 NAEP Reading Assessment & Item Specifications)

STORY LEVEL THEME: A woman's tough, but sympathetic, response to a teenage boy who tries to steal her purse causes the boy to change his behavior/attitude

ABSTRACT THEME: Kindness, trust, and generosity are used to teach a young boy a lesson about right and wrong

PLOT:
Problem: Roger attempts to steal Mrs. Jones' purse in hopes of getting money to buy a pair of shoes he cannot afford to purchase

Conflict: Will Roger run or will he let Mrs. Jones help him

Resolution: Roger reciprocates the trust and caring demonstrated by Mrs. Jones, and is given a chance to change his life

SETTING (and how it is connected to the themes and significant ideas in the text): Urban area and small apartment where everything is in view provide a woman with an opportunity to help a young boy to see the wrongness of his actions

CHARACTER/S* (traits that are connected to significant ideas in the text):

Mrs. Luella Bates Washington Jones/Woman

- Trusting—she leaves her purse where the boy could take it if he wanted to; provides him with a choice about going to the store with her money to buy food or eating what she has on hand
- Honest—she is straightforward with the boy and never tries to deceive him
- Caring—she does not turn him over to the police, gives him food and money

MAJOR EVENTS:**

1. Roger attempts to steal a purse of an older woman but is thwarted in his attempt by a woman who is not easily taken advantage of.
2. The woman quickly establishes her physical and emotional control over the boy.
3. She is able to judge the character of the boy and use her insights and experience to build trust between them.

AUTHOR'S CRAFT:
Tone: one of authority in the beginning changing to one of concern
Rhetorical Devices
Use of italics
Significance of the title and use of M'am throughout
Use of slang diction
Use of "run" image throughout

* One of two characters included; 3 of 7 traits are listed

** 3 of the 12 major events are shown in this illustration

fiction. The protocol for the narrative map captures the structure of fiction—themes, plot structure, setting characters, and author's craft. In that a typical map is lengthy and detailed, only the highlights of a narrative map—for a piece of literary-realistic fiction for eighth grade—are presented in Figure 3. The process begins with identifying themes at both the story level (specific events of the narrative) and abstract level (general concepts that run through the narrative). The interrelatedness of text features are emphasized, such as the manner in which setting or the roles of characters influence plot.

Non-narrative maps are used for texts such as speeches, exposition, descriptions, explanations, argumentative essays, and other documents. Non-narrative maps are supposed to capture the hierarchical organization of information, with multiple levels of ideas (central, major, and supporting). Where possible and appropriate the maps also identify the role of text features (e.g., subheadings, charts, and illustrations), and elements of author's craft (e.g., figurative language and rhetorical devices).

Because of space limitations, only a small portion of a sample non-narrative text map is provided in Figure 4. The map begins with the central idea and purpose and then maps out major and supporting ideas and role in text organization. An organizational element, such as comparison structure, might be highlighted, after which the major and supporting components (what is being compared and on what criteria) for the element are depicted hierarchically in a portion of the map.

Summary

We have reviewed three different types of qualitative systems and elaborated on the ways in which they serve one of two general purposes for their use: Text leveling systems are designed exclusively to enable a better match between students' abilities and the texts we ask them to read. Rubric + Exemplar systems and Text Maps highlight parts of texts that deserve special attention and/or instruction when we ask students to read and understand them. Additionally, several of the R+E systems also result in assigning a text to a level, namely, the CCSS system and its derivatives, such as the Kansas and Georgia systems. Most important to remember about these systems is that the research base documenting their efficacy for either of these purposes is very meager. Even so, we conclude that TM systems can be very useful in identifying features or segments of text that deserve special instructional treatment.

Lingering Issues

As important as it is to employ qualitative analyses as a ballast for or complement to quantitative indicators of text complexity, it is even more important to refine our qualitative indicators and analyses so that they will be able to instill enough confidence in potential users to earn equal status alongside quantitative indicators in making decisions of consequence about which texts to use

FIGURE 4

Portions of a Non-Narrative Text Map for Eighth Grade: *Ellis Island* (from 2009 NAEP Reading Assessment & Item Specifications) (AIR, 2008)

CENTRAL IDEA: To provide a historical account of immigrants told in the words of immigrants who can to the US through Ellis Island between 1892 and 1954

MAJOR IDEAS*:

Org. Element—Description/Introduction

Major Idea: Between 1892 and 1954, Ellis Island was the "doorway to America" for 17 million people

Supporting Idea/s:

- not everyone was welcome
- "land of the free" was not so free to everyone

Org. Element—Cause

Major Idea: Immigrants came from Europe to escape oppression/poverty and/or seek a better life

Supporting Idea/s: First-hand accounts from a woman escaping Turkish oppression in Armenia, and a man from the Ukraine seeking opportunities offered by U.S.

Org Element—Effect/Problem

Major Idea: Those who wanted to immigrate had to endure great hardship to travel to U.S.

Supporting Idea/s:

- They had to contend with border guards, thieves, dishonest immigration agents, and bad conditions on the ships they crossed on.
- Once they saw NY and Statue of Liberty—they felt it was worth it.

TEXT FEATURES:

- Subheadings, illustrations, use of italics to set off quotations from past immigrants
- Illustration of "cattle-pen-like" method of processing

AUTHOR'S CRAFT:

- Use of first hand accounts to illustrate the points about the immigrant experience in general and on Ellis Island
- Use of a doorway to America/doorway metaphor

* 3 of 7 major ideas shown in this illustration

with whom and how. If any of qualitative indicators are to achieve this status, we will, as a field, have to settle a number of lingering issues regarding their construct validity and implementation, among these are issues that relate to (a) purpose, (b) teacher professional development, (c) exemplars, and (d) developmental progression.

Staying True to the Purposes of Qualitative Text Analysis

In the final analysis, the question of interest about qualitative systems is, What are they good for? That is, how can they help us in ways that quantitative systems cannot? In this paper we have highlighted the two major purposes of such systems, matching students to texts and unearthing the "tricky parts" of particular texts for support during reading. In theory, if we do a good job of matching texts to students, they should be able to read and comprehend most texts without too much intervention from teachers. But if our goal is truly to up the ante in text complexity (a central tenet of the CCSS), then the second purpose of highlighting challenging features for instruction will be even more important than the matching function.

A third purpose of qualitative analysis, which we have not discussed thus far, may be even more important than the two avowed purposes: Qualitative analyses, both the R+E and TM systems, can serve to vet, validate, or/or adjust the recommendations of quantitative systems.

Double-checking quantitative estimates of difficulty. The measurement issues with quantitative systems that rely on syntax and vocabulary have long been documented (Anderson et al., 1985; Klare, 1984). As a general rule, quantitative systems tend to underestimate the complexity of narrative texts and overestimate the difficulty of informational texts. These patterns have to do with the unique features of narratives and informational texts. Specifically, narrative texts often contain long stretches of dialogue, expressed in common words and short sentences, that belie their surface simplicity, embodying as they often do subtle nuances of meaning, character development, and complex themes of human experience. For informational texts, it is the rare, often technical vocabulary that sends readability scores soaring, even though the evidence tells us that when these words are repeated—and explained thoroughly—readers can become sufficiently accustomed to them that they lose their vexing power.

The vagaries of quantitative systems are illustrated by the range of Lexiles reported for grade-level bands in Table 6. It is significant to note the discrepancy between the Lexile targets recommended in the CCSS Appendix and supple-

TABLE 6

Grade Bands: CCSS and Actual Lexile Ranges

Grade Band	CCSS Recommend Ranges	Actual Ranges of CCSS Exemplars
2–3	420–820	240–1100
4–5	740–1010	550–1190
6–8	925–1185	560–1430
9–10	1050–1335	600–1600
11–CCR	1185–1385	670–1750

ment and the actual measured Lexile scores of texts in every grade band: The range of actual Lexile scores exceeds the recommended ranges on both ends. Thus the grade 11–CCR band contains some texts with measured Lexile scores in the grade 2–3 band; conversely, some of the texts in the grade 2–3 band have measured Lexile scores designated for the 6–8 band.

Qualitative analyses will serve a critical function in ensuring that texts are assigned to appropriate levels. Qualitative analyses, for example, will prevent us from concluding that we can use *Grapes of Wrath* in grades 2–3 in spite of its measures Lexile level of 680. *Grapes of Wrath* has content that will challenge many of the grade 9–10 students who are expected to read it. Similarly, a qualitative analysis of *Let's Investigate Marvelously Meaningful Maps*, with a Lexile of 1070, will suggest to us that it is not really at the upper reaches of the grade 4–5 band (as the Lexile score would suggest). Many third graders should be able to read it because the rare words that resulted in the elevated Lexile score—words such as equator, latitude, and meridian—are repeated frequently and explained well. The function of the qualitative scheme of the Common Core and the various spin-offs (e.g., Kansas, Georgia, Achieve the Core) appears to be to provide a second sorting score. In all of these endeavors, qualitative analyses are used to validate a quantitative assignment.

Supporting instruction for challenging texts. When the purpose of qualitative systems is to support instruction, the focus on ensuring that texts are sorted into an appropriate "fifth grade" or "eighth grade" bin is less compelling than providing guidance for teachers in implementing lessons that provide students with scaffolds and skills for navigating texts that are just out of their reach. The annotations in Figures 1–4 show that the ACT annotation and the NAEP text maps provide precisely this sort of guidance. By contrast, the application of the R+E of the CCSS (Figure 1) provides little guidance for instruction. A missed opportunity is Steinbeck's use of mixed genres in *Grapes of Wrath*, in which he weaves the content from nonfiction articles on the conditions of farmworkers into the rich narrative of the Joad family.

Support for Teachers in Teaching and Selecting Texts

If qualitative indicators of complexity are to make any difference in improving students' comprehension of challenging text, then they will have to influence teacher beliefs, knowledge and, ultimately, practices. Teachers who don't know why some characteristics of text, some purposes for reading, some comprehension tasks are harder than others will not be in a position to select texts that are likely to "hit the just-right mark" for particular individuals or groups. And without this knowledge, they certainly won't be able to offer scaffolding that allows students to access the key ideas from text that is just beyond students' reach. This means that professional learning, and hence, professional development, is a key to increasingly the salience and influence of qualitative schemes for analyzing text complexity.

Surely the level of information required of teachers will differ as a function of age of the readers and a text's developmental complexity. A teacher working with a class of eighth or ninth graders on *The Book Thief* (Zusak, 2007) or a twelfth-grade teacher using *Their Eyes Were Watching God* (Hurston, 2006) will presumably need different information about text, task, or knowledge complexity than a second-grade teacher working with students on *The Treasure* (Shulevitz, 1986) or *Tops and Bottoms* (Stevens, 1995). To appropriately teach the latter, distinctions between parables (*The Treasure*) and trickster tales (*Tops and Bottoms*) are useful as is an understanding of critical concepts (e.g., inscription in *The Treasure* and particular vegetables in *Tops and Bottoms*). However, the level of information required to work with students on *The Book Thief*—especially students whose knowledge of the Holocaust is limited—will be extensive.

Especially critical is the question as to whether teachers need to do the analysis themselves or whether there are ways in which teacher collectives and/or publishers can provide some of the information. Even if publishers provide the information, teachers will need to engage in in-depth analyses of complex texts at particular levels in published anthologies to satisfy themselves that the authors of the teacher editions "got it right". As a practical concern, a question we will have to answer is whether a "coach" who works, say, at the school level, can give the kinds of supports required, especially if he or she is assigned from the outside.

Even at the beginning reading levels, it is doubtful that an overall designation of "uncomplicated" (ACT), an alphabetic letter on a scale of A through Z (Fountas & Pinnell, 2001), complex (CCSS), or grade-3 level reader (Chall et al.) will aid teachers in providing the instruction required for a text that is truly complex for a group of students. Presumably, a text that is truly complex for a reader requires the kind of scaffolded coaching that has been described as part of deliberate practice (Ericsson, 1996). That is—there is something for learners to learn; and the teacher, coach, or tutor, must do what is required to help them dig it out. Generic ratings (e.g., Level A, moderately complicated, requires grade 3 skills) will be inadequate to provide the kind of instruction that develops students' capacity to read progressively more complex texts across the grades—the essence and explicit goal of Standard 10 of the CCSS. In-depth information about the qualities of a text is required for the prior knowledge of the reader to be brought into the mix. Interpreting the qualitative (and quantitative) information in relation to readers and the task is the teacher's milieu. And when teachers operate in this space, they come close to turning the corner (the vertex of the text complexity triangle) where the qualitative leg joins the reader and task leg.

The Tyranny of the Exemplar

Exemplars (for text analysis and leveling) and anchor papers (for scoring writing) are a core element of qualitative analysis systems that rely on human judg-

ment to create scores or prescribe instruction. As we suggested earlier, they are the concrete realization of the phenomenon being judged, and they make abstract rubrics come alive so that judges know what that phenomenon looks like when they see it. Anchor papers do it for writing, and exemplars do it for systems that require students to make judgments about the level and/or complexity of text. Earlier we pointed out the key role that exemplars play in (a) the ACT system (ACT 2006), where they highlight the particular challenges that particular texts and genres present to readers and teachers at different levels of complexity; (b) Chall et al. QATD (1996), where they epitomize the cognitive moves that readers need to make to successfully understand texts at a given level; and (c) various text leveling systems (Fountas & Pinnell, 2009; Hoffman et al., 2001; Peterson, 1991), where exemplars stand as prototypes for a given level of challenge within the staircase of complexity used to match students to books. A common characteristic of these various ways of addressing complexity is that the creators of each system develop and implement some sort of vetting standards for determining where texts "fit" in their particular complexity continuum. The vetting is typically carried out by trained professionals, who use their deep experience with texts and readers along with specific criteria for selecting exemplars that they acquire in some sort of training procedure. And in some sense, the validity of each of these approaches depends on the credibility and reliability of that vetting process.

But the exemplars in the CCSS, both in the Standards themselves and also in Appendix A, present dilemmas that do not surface in the various text level/complexity systems. The basic difference is that exemplars in a policy document play a different role than in a technical document that describes a procedure for establishing text complexity.

Canonical texts. First, protestations to the contrary (e.g., these examples are meant to illustrate the range of types of text that might be used in a school reading program), exemplars often get interpreted as a canon. So instead of illustrating the sorts and range of texts that might be used, the exemplars become the entire population that educators use in a grade-level band. In short, the exemplars ***become*** the canon of texts that are taught. We have labeled this dilemma, the "tyranny of the exemplar," because it is hard for any of us to resist believing that, if a text is good enough to exemplify a level, then it ought to be taught at that level. And, indeed, some of the materials currently under development suggest that the exemplars provided in the CCSS are making their ways into curriculum packages (Engage NY, 2013). But this is a temptation that must be resisted lest we marginalize all attempts by educators to adapt the portfolio of texts used in specific district and school settings to the needs and interests of their students.

This aspect of the ELA standards—the subtle transformation from exemplars into canon—is most strident in terms of state autonomy. The standards promise in the introduction that states, districts, even teachers will have autonomy in curricular choices, but the dominance of the exemplars betrays such a promise;

the list will become the canon unless some dramatic pronouncement is made or some significant step is taken. Since the standards say that the exemplars are only illustrative of the range, the step must be bold. Of this we can be sure: the smaller the list, the more likely it will become a canon. So one useful step might be to expand the list so dramatically that no district or school could possibly cover all the exemplars. Another might be to require states and districts to develop their own lists, perhaps even contributing them to a national exemplar bank. A third might be to establish a commission that every year has the task of adding newly published works to the exemplar bank. Try as the standards might to deny their canonical role, it is the default role they will serve unless specific steps are taken to reign in that natural tendency.

Unwarranted assumptions of homogeneity. Second, at least through grade 5, the use of "bands" that are considered more or less homogeneous is problematic. While it might make sense to have an internally undifferentiated band that defines the range of texts that a typical high school junior or senior can read, it makes little sense to distinguish among the range of texts that fall into the grade 2–3 band. Here's the issue: Relative to one's starting point, the proportion of intellectual growth from the beginning of 2nd grade to the end of 3rd grade is much greater than the comparable proportion of growth from the beginning of 11th to the end of 12th grade. So in the earlier grades (and the earlier the grade level, the more problematic the practice), dumping a set of texts into a grade band without specifying where in the grade band students would be expected to read any given text leads to confusion and even unreasonable expectations for our youngest and most vulnerable readers. The broad conception of bands means that we are highly likely to find second graders reading *Charlotte's Web* (White, 1952)—a text conventionally assigned to 4th grade. Technically, a proficient second grader can read *Charlotte's Web* because it, like many conversationally written stories, has many common, high-frequency words in it. But just because words are easy, doesn't mean that ideas are. Second graders may be to read the words, and they may even be able to understand the basic premises of the text (especially if they've seen a movie version of it). But even advanced second graders may miss the nuances of the text, especially character development and word choice. By suggesting, perhaps even mandating that students who fall anywhere within a band (e.g., 2–3, 4–5, 6–7) should be able to read the most complex of texts within that band with guidance, we end up with unrealistic expectations for at least some of the students in the band.

The vetting problem. A final problem with the exemplar texts is that the CCSS document provides no account of how text band assignments were made. The document requires users to exercise blind faith in an undocumented process. With so much at stake, namely the well-being and academic progress of our children, scientific evidence rather than blind faith is a more appropriate standard for fixing the expected levels of difficulty of text.

Rethinking Developmental Progressions

As with any framework designed to promote, examine, and monitor student learning, the question of what develops over time and across grade levels is critical to the CCSS. Such theories of development are always implicit—but usually explicit—in documents that guide learning and teaching, and the CCSS document is no exception. Thus the first question for our consideration is what is the theory of development underlying the common core? The second—did we get it right? Right enough at least so that if we enact the CCSS, we will promote student learning and the ability to handle the range of texts that our schools and society require of each generation of citizens.

Implicit or explicit progressions? If one looks at the reading standards themselves, there is an attempt to build an explicit theory of ***task*** development—what we ask students to do from one level to the next. Unfortunately, the progressions offered are more ad-hoc than systematic, let alone theoretical, in delivery. As Pearson (2013) and Applebee (in press) have noted, the changes in focus (what the reader is asked to do in the name of the standard), scope (how much text the reader would have to consult to complete the task) and support (what sorts of scaffolds are present to help the reader carry out the task) vary considerably in what seem like random ways across the bands of grade level for which specific iterations of the standards play out. The net result is that one is baffled about why, for example, analogies and allusions first appear in Standard 4 (vocabulary usage) at grade 8 and are gone by grade 9? Are we to infer that grade 8 is the first point in the curriculum at which they can or should be addressed? Or that they should not be continued in grade 9? Similar discontinuities abound at every level of the standards (see Pearson, 2013, for more examples in grades K–5).

Other things besides tasks also develop in the CCSS; namely, both the structural and the conceptual complexity of the texts encountered. And these changes constitute an answer to the question, why does reading become more challenging as students move from one grade level to the next?

What changes occur in text features across the developmental progression? The rubrics of the CCSS, ACT, and QATD (see Tables 2–5) aim to answer this question. All three of the systems share the trait of vocabulary (illustrated in Table 4), structure (although the QATD focuses on the structure of sentences while the other two focus on text structure as well), and knowledge demands. There is somewhat more ambiguity in terms of levels of meaning or relationships among ideas and literary analysis (QATD) but presumably this trait represents the degree of inference required to construct meaning.

Excerpts from narrative exemplars from the beginning, middle, and end of the CCSS's staircase of complexity (Table 7) illustrate the challenge of ferreting out the implicit theory of text complexity development across levels. With respect to text structure, a surface-level examination suggests the texts are not substantially different from one another.

TABLE 7

Excerpts from Exemplars of Narrative Texts for Grade Bands

Text	Grade Band	Excerpt
The Stories Julian Tells (Cameron, 1981)	2–3	Huey was the one who wanted the house of flowers the most. I wanted the giant corn. My father said he wasn't sure he wanted either giant corn or a flower house, and if we wanted them, we would have to take care of them all summer by pulling weeds.
Little Women (Alcott, 2008)	6–8	"Birds in their little nests agree," sang Beth, the peacemaker, with such a funny face that both sharp voices softened to a laugh, and the pecking ended for that time. "Really, girls, you are both to be blamed," said Meg, beginning to lecture in her elder-sisterly fashion.
Crime and Punishment (Dostoyevsky, 1996)	11–CCR	Nobody wears such a hat, it would be noticed a mile off, it would be remembered. What matters is that people would remember it, and that would give them a clue. For this business one should be as little conspicuous as possible. Trifles, trifles are what matter! Why, it's just such trifles that always ruin everything."

Knowledge demands. When it comes to knowledge demands, there are transparent differences. The overt decision-making of an individual to commit a crime in *Crime and Punishment* is likely more demanding than understanding the squabbling between siblings in *Little Women* or deciding what should be planted in *The Stories Julian Tells*. Reading about deliberately choosing to commit a crime is inappropriate for primary-level students. How much "harder" it is to understand planting, sibling squabbling, or details of a plan to commit a crime is less certain.

Sheer length. One feature of texts that no analysis has yet captured is their sheer length. The excerpt for grades 2–3 comes from a 1,200-word chapter of a book in which each of six chapters (i.e., a 7,200-word book) tells another story from Julian's life, each based on experiences of middle-class children. *The Little Women* excerpt is from an 88,000-word text where the persistent squabbling between Jo and Amy is a secondary theme that runs throughout the book. Similarly, the excerpt for grade 11–CCR, *Crime and Punishment*, is from a book with over 203,500 words. The character's contemplation of how trifles might thwart his success as a burglar is only a small part of the retrospective contemplation in which the character engages. But the length issue, along its implications for the attribute that some have labeled stamina (Greenleaf, Schoenbach, Cziko, & Mueller, 2001; Hiebert, Wilson, & Trainin, 2010; Valencia, Smith, Reece, Li, Wixson, & Newman, 2010) remains largely uninvestigated. One thing we do know is that with longer texts, both fluency (Valencia et al., 2010) and comprehension (Hiebert et al., 2010) decrease as students move through a longer text.

Differential importance of text features across grade level bands. A related (to the step size) issue is the question of whether different aspects of complexity do, could, or should play differentially important roles at different levels. For example, do issues of word decodability and predictability (remember the work of Hoffman et al., 2001) matter more than syntax at K–1, while syntax matters more in intermediate grades and yet another factor, such as levels of meaning or purpose, in middle-school texts? We suspect they do. As we learn more about the empirical development of students' capacity to cope with increasingly challenging texts, we will certainly develop insights about which facets of complexity matter most in different grade bands.

Disentangling natural co-variation among aspects of complexity. In the initial section of this paper, we raised the question of whether readability causes or merely reflects comprehension difficulty, pointing to quantitative and linguistic research that suggests that sometimes more complex words and syntax may simply reflect the communication of more complex ideas (Davison & Kantor, 1982; McNamara, Kintsch, Songer, & Kintsch, 1996; Pearson, 1974-75); in other words, for complicated ideas, there may be a lower limit on how simply they may be expressed. What this suggests, when it comes to reviewing complex texts for potential instructional scaffolding, is that teachers might be well advised to focus on the complexity of the content rather than the obscurity of the words or the syntax. Figuring out what explanations, analogies, and examples might help students negotiate tough content may be more productive than addressing rare syntax or rare words. One possible, even probable, approach would be to analyze how and why the choice of words and syntax made by an author were just right for communicating the ideas conveyed in the text. Of course, we haven't a shred of evidence to support this approach, but it is certainly worth exploring experimentally. And it might have the side benefit of preventing us from some very unproductive ventures into teaching syntactic complexity or drilling students on the meanings of rare words.

Mapping task complexity onto text complexity. One final perspective on developmental progressions—the role of task complexity. Task complexity is the one variable that has ***not*** been examined in any of the qualitative analyses of complexity we have reviewed. No one seems to have addressed the question of what students do to demonstrate their understanding of a text. In readability studies, it is assumed that the particular task used to scale comprehension for passages doesn't matter much; after all, looking across a wide range of readability studies, researchers seldom specify the outcome measure that serves as the criterion, implying that any one task is just as good as any other for validating readability formulas. However, a prima facie analysis suggests that task has to matter: Asking middle school students to identify the topic of a chapter out of a high school life science text is likely easier than asking them to critique E.B. White's use of symbolism in *Charlotte's Web*. Moreover, task must also vary at least partially independent of text; that is, as our examples illustrate, one can

construct a relatively simple task about a very difficult text or a relatively difficult task about a simpler text.

What is it then that makes most tasks difficult for complex texts and most easy for simple texts? Our claim is that it is the ideas themselves that drive complexity. Thus, for the most part, both the structural apparatus within which they are communicated (the sentence syntax and rhetorical frames) and the tasks we ask students to complete in demonstrating their comprehension (finding the main idea, inferring character motives, connecting ideas across paragraphs, creating a summary or a synopsis) are driven by those ideas.

The match between content and structure or content and tasks isn't perfect, and it's the imperfections that tell us that *Grapes of Wrath* is appropriate for the grade 2–3 band or that *Charlotte's Web* can be used early in grade 2 rather than its usual grade 4 placement. But in general, harder content will come packaged with bigger, less common words, longer, more complex sentences, and more intricate rhetorical frames (a problem-solution frame rather than a list). Moreover, finding the main idea or inferring character motives will, in general but not always, be harder for *Crime and Punishment* than for *The Stories Julian Tells*.

Notice, also, that if we are right about the centrality of content, if it is the ideas that matter most, then all of the tortured machinations about which version of a particular standard should prevail for narratives in third versus fourth versus fifth grade is unnecessary. We might be just as well off (perhaps better off) to accept the appropriateness and necessity of each of the nine anchor standards as representing the full range of tasks we'd like all students to engage in as they make their way through texts at each level from K–12; then we could figure out how to find ways to embed them in the texts we decide to use at different grade levels. In short, we should let the content—the ideas—drive our placement of texts and the tasks we generate to ensure and assess comprehension of those texts. Surely, we will attend also to the structures in which those ideas travel and to the tasks we use to engage students in conversations about the texts, but we will always start and end with the ideas as the object of our analyses.

A focus on content will impose two additional requirements on us. First, qualitative analysis will necessarily trump quantitative analyses of texts, and second, the analyses we engage in for structure and task will be focused on the goal of making texts more accessible to the broadest possible range of students. Such a focus will also put quantitative analyses in proper perspective, for we will recognize that the key elements of quantitative inquiry, long words and complex syntax, are nothing more than symptoms of challenging content. Armed with that knowledge, we will be better positioned to figure out how to help students manage that content, which is our most important job as teachers. And this brings us full circle to one of the central goals of the Common Core State Standards for English Language Arts, which is eloquently stated in the introduction to the Standards, when they assert that readers who meet these stan-

dards "actively seek the wide, deep, and thoughtful engagement with high-quality literary and informational texts that builds knowledge, enlarges experience, and broadens worldviews." (NGA Center for Best Practices & CCSSO, p. 3).

References

ACT (2006). *Reading between the lines: What the ACT reveals about college readiness in reading*. Iowa City, IA: Author.

American Institutes for Research (2008). *Reading assessment and item specifications for the 2009 National Assessment of Educational Progress*. Washington DC: Author.

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.

Applebee, A. (in press). *Common Core State Standards: The promise and the peril in a national palimpsest*. English Journal.

Beaver, J. (2003). *Developmental reading assessment*. Parsippany, NJ: Celebration Press.

Beck, I.L., Farr, R., Strickland, D.S., Ada, A.F., Hudson, R.F., McKeown, M.G., Scarcella, R.C., & Washington, J. (2008). *Storytown*. Orlando, FL: Harcourt, Inc.

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

Blau, E.K. (1982). The effect of syntax on readability for ESL students in Puerto Rico. *TESOL Quarterly, 16*, 517–528.

Bormuth, J.R. (1966). Readability: A new approach. *Reading Research Quarterly, 1*(3), 79–132.

Brown, A.L., Armbruster, B., & Baker, L. (1986). The role of metacognition in reading and studying. In J. Orasanu (Ed.), *Reading comprehension: From research to practice* (pp. 49–75). Hillsdale, NJ: Erlbaum.

California English/Language Arts Committee. (1987). *English-language arts framework for California public schools (kindergarten through grade twelve)*. Sacramento, CA: California Department of Education.

Carver, R.P. (1976). Measuring prose difficulty using the Rauding scale. *Reading Research Quarterly, 11*, 660–685.

Chall, J. S., Bissex G., Conard, S., Harris-Sharples, S. (1999). *Qualitative assessment of text difficulty*. Brookline, MA: Brookline Publishers.

Choldin, M.T. (1979), Rubakin, Nikolai Aleksandrovic, in Kent, Allen; Lancour, Harold; Nasri, William Z. et al., *Encyclopedia of library and information science, 26 (illustrated ed.)*, CRC Press, pp. 178–79, ISBN 9780824720261

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

Copeland, M., Lakin, J., & Shaw, K. (January 26, 2012). *Text complexity and the Kansas Common Core Standards for English Language Arts and Literacy in History/Social Studies, Science, and Technical Subjects*. Retrieved from http://www.ccsso.org/Resources/Digital_Resources/The_Common_Core_State_Standards_Supporting_Districts_and_Teachers_with_Text_Complexity.html

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), 187–208.

DiPardo, A., Storms, B.A., & Selland, M. (2011). Seeing voices: Assessing writerly stance in the NWP Analytic Writing Continuum. *Assessing Writing, 16*(3), 170–188.

EngageNY. (2013). *Lincoln Gettysburg Address*. Retrieved from http://www.docstoc.com/docs/101478560/Lincoln-Gettysburg-Address-EngageNY (March 15, 2013)

Ericsson, K.A. (1996). The acquisition of expert performance: An introduction to some of the issues. In K. A. Ericsson (Ed.), *The road to excellence: The acquisition of expert performance in the arts and sciences, sports and games* (pp. 1–50). Mahwah, NJ: Erlbaum.

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

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

Fountas, I.C., & Pinnell, G.S. (2009). *The Fountas & Pinnell leveled book list, K–8+: 2010–2012 Edition*, Print Version. Portsmouth, NH: Heinemann.

Fountas, I.C., & Pinnell, G.S. (2010). *The continuum of literacy learning, Grades PreK–8*. Portsmouth, NH: Heinemann.

Fountas, I.C., & Pinnell, G.S. (2012). *The F & P Text Level Gradient: Revision to Recommended Grade-Level Goals*. Portsmouth, NH: Heinemann. Retrieved from: http://www.heinemann.com/fountasandpinnell/pdfs/WhitePaperTextGrad.pdf

Gardner, H. (1987). *The mind's new science*. New York, NY: Basic Books.

Georgia Department of Eduction (2011). *Common core Georgia performance standards text complexity rubric*. Retrieved from https://www.georgiastandards.org/Common-Core/Documents/9%20-11%20rubrics.pdf

Graesser, A.C., McNamara, D.S., & Kulikowich, J. (2011). CohMetrix: Providing multilevel analyses of text characteristics. *Educational Researcher, 40*, 223–234.

Gray, W.S., & Leary, B.W. (1935). *What makes a book readable*. Chicago: University of Chicago Press.

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

Greenleaf, C.; Schoenbach, R.; Cziko, C.; & Mueller, F. (2001). Apprenticing adolescent readers to academic literacy. *Harvard Educational Review, 71*(1), 79–129.

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

Hess, K., & Biggam, S. (2004). *A discussion of "increasing text complexity"*. Published by the New Hampshire, Rhode Island, and Vermont departments of education as part of the New England Common Assessment Program (NECAP). Retrieved from http://www.nciea.org/publications/TextComplexity_KH05.pdf

Hess, K., & Hervey, S. (2010). *Local assessment toolkit: Tools for examining text complexity*. Dover, NH: The National Center for the Improvement of Educational Assessment, Inc.

Hiebert, E.H., Wilson, K.M. & Trainin, G. (2010). Are students really reading in independent reading contexts? An examination of comprehension-based silent reading rate. In E.H. Hiebert & D. Ray Reutzel (Eds.), *Revisiting Silent Reading: New Directions for Teachers and Researchers*. Newark, DE. IRA.

Hirsch, E.D., Jr. (Ed.) (2005a). *What your fourth grader needs to know: Fundamentals of A Good Fourth-Grade Education (Core Knowledge)*. New York, NY: Dell Publishing.

Hirsch, E.D., Jr. (Ed.) (2005b). *What your second grader needs to know: Fundamentals of A Good Second-Grade Education (Core Knowledge)*. New York, NY: Dell Publishing.

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

Hollingsworth, S. (1991). *Learning to teach literature in California: Challenging the rules for standardized instruction (Research Series #200)*. East Lansing, MI: The Institute for Research on Teaching, Michigan State University.

Johns, J. (2012). *Basic reading inventory (11th Ed.)*. Sunnyvale, CA: Kendall Hunt Publishing.

Klare, G.R. (1963). *The measurement of readability*. Ames, IA: The Iowa State University Press.

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

Koslin, B.L., Zeno, S., & Koslin, S. (1987). *The DRP: An effective measure in reading*. New York: College Entrance Examination Board.

Leslie, L., & Caldwell, J.S. (2010). *Qualitative reading inventory (5th Ed.)*. Boston, MA: Pearson.

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

McCall, W.A., & Crabbs Schroeder, L. (1979). *McCall-Crabbs Standard Test Lessons in Reading*. New York: Teachers College, Columbia University. [Originally published in 1926]

McNamara, D.S., Kintsch, E., Songer, N.B., & Kintsch, W. (1996). Are good texts always better? Interactions of text coherence, background knowledge, and levels of understanding in learning from text. *Cognition and Instruction, 14*(1), 1–43.

MetaMetrics (2000). *The Lexile framework for reading*. Durham, NC: Author. Retrieved from http://cdn.lexile.com/m/cms_page_media/135/The-Lexile-Framework-for-Reading.pdf

Milone, M. (2009). *The development of ATOS: The Renaissance readability formula*. Wisconsin Rapids, WI: Renaissance Learning.

National Assessment Governing Board (1991). *Reading Framework for the 1992 National Assessment of Educational Progress*. Washington, DC: US Government Printing Office.

National Governors Association Center for Best Practices & Council of Chief State School Officers. (2010). *Common Core State Standards for English language arts and literacy in history/social studies, science, and technical subjects with Appendices A–C*. Washington, DC: Authors.

Nelson, J., Perfetti, C., Liben, D., & Liben, M. (2012). *Measures of text difficulty: Testing their predictive value for grade levels and student performance*. New York: Student Achievement Partners.

Partnership for Assessment of Readiness for College and Careers (2012). *Model Content Frameworks*. Retrieved from http://www.parcconline.org/parcc-model-content-frameworks

Pearson, P.D. (1974–75). The effects of grammatical complexity on children's comprehension, recall and conception of semantic relations. *Reading Research Quarterly, 10*, 155–192.

Pearson, P.D. (1984). Asking questions about stories. In A.J. Harris & E.R. Sipay (Ed.). *Readings in Reading Instruction [3rd Ed.]*. New York: Longman.

Pearson, P.D. (2013). Research foundations of the Common Core State Standards in English language arts. In S. Neuman and L. Gambrell (Eds.), *Quality Reading Instruction in the Age of Common Core State Standards* (pp. 237–262). Newark, DE: International Reading Association.

Pearson, P.D., & Camparell, K. (1981). Comprehension of text structures. In J. Guthrie (Ed.), *Comprehension and Teaching* (pp. 27–54). Newark, DE: International Reading Association.

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

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.

Pikulski, J.J., & Shanahan, T. (Ed.) (1982). *Approaches to the informal evaluation of reading*. Newark, DE: International Reading Association.

Rog, L.J., & Burton, W. (2001). Matching texts and readers: Leveling early reading materials for assessment and instruction. *The Reading Teacher, 55*, 348–356.

Shanahan, T., Callison, K., Carriere, C., Duke, N.K., Pearson, P.D., Schatschneider, C., & Torgersen (2010). *Improving reading comprehension in kindergarten through 3rd grade: A practice guide (NCEE 2010-4038)*. Washington, DC: National Center for Education Evaluation and Regional Assistance.

Shepard, L.A., Flexer, R.J., Hiebert, E.H., Marion, S.F., Mayfield, V., & Weston, T. J. (1996). Effects of introducing classroom performance assessments on student learning. *Educational Measurement Issues and Practices, 15*, 7–18.

Sherman, L.A. (1893). *Analytics of literature: A manual for the objective study of English prose and poetry*. Boston: Ginn and Co.

Singer, H. (1975). The SEER technique: A non-computational procedure for quickly estimating readability level. *Journal of Reading Behavior, 7*, 255–267.

Stein, N. L., & Glenn, C. G. (1979). An analysis of story comprehension in elementary school children. In R. Freedle (Ed.), *Multidisciplinary Approaches to Discourse Comprehension*. Hillsdale, N.J.: Ablex, Inc.

Student Achievement Partners. (2012). *Qualitative dimensions of text complexity chart: 2nd–3rd grade band*. New York: Author. Retrieved from http://www.achievethecore.org/steal-these-tools/text-complexity/qualitative-measures (October 4, 2012)

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

Trelease, J. (2006). *The read-aloud handbook (6th Ed.)*. New York, NY: Penguin Books.

Valencia, S.W., Pearson, P.D., Peters, C.W., & Wixson, K. (1989). Theory and practice in statewide reading assessment: Closing the gap. *Educational Leadership, 46*, 57–63.

Valencia, S.W., Smith, A.T., Reece, A.M., Li, M., Wixson, K.K., & Newman, H. (2010). Oral reading fluency assessment: Issues of construct, criterion, and consequential validity. *Reading Research Quarterly, 45*(3), 270–291.

Wixson, K., Peters, C., Weber, E., & Roeber, E. (1987). New directions in statewide reading assessment. *The Reading Teacher, 40*, 749–754.

Literature

Alcott, L.M. (2008). *Little women*. New York, NY: Puffin Books.

Aliki. (1988). *A weed is a flower*. New York, NY: Simon & Schuster Books for Young Readers.

Cameron, A. (1981). *The stories Julian tells*. New York, NY: Random House Books for Young Readers.

Dostoyevsky, F. (1996). *Crime and punishment*. New York, NY: Bantam Classics.

Hurston, Z.N. (2006). *Their eyes were watching God*. New York, NY: Harper Perennial Modern Classics.

Lee, H. E. (1999). *Night water*. New York, NY: Scribner.

MacLachlan, P. (2004). *Sarah, plain and tall*. New York, NY: HarperCollins.

Rylant, C. (2008). *Poppleton in winter*. New York, NY: Cartwheel Books.

Shulevitz. U. (1986). *The treasure*. New York, NY: Square Fish.

Steinbeck, J. (2006). *The grapes of wrath*. New York, NY: Penguin Classics.

Stevens, J. (1995). *Tops and bottoms*. New York, NY: Harcourt Children's Books.

White, E.B. (1952). *Charlotte's web*. New York, NY: HarperCollins.

Zusak, M. (2007). *The book thief*. New York, NY: Alfred A. Knopf.

TextProject, Inc. is a non-profit public benefit corporation. Its aim is to bring beginning and struggling readers (of any age) to high levels of literacy through a variety of strategies and tools, particularly the texts used for reading instruction.

Find out more at **textproject.org**