Educ Psychol Rev (2012) 24:63–88
DOI 10.1007/s10648-011-9181-8

REVIEW

Reconstructing Readability: Recent Developments and Recommendations in the Analysis of Text Difficulty

Rebekah George Benjamin

Published online: 4 October 2011

Abstract Largely due to technological advances, methods for analyzing readability have increased significantly in recent years. While past researchers designed hundreds of formulas to estimate the difficulty of texts for readers, controversy has surrounded their use for decades, with criticism stemming largely from their application in creating new texts as well as their utilization of surface-level indicators as proxies for complex cognitive processes that take place when reading a text. This review focuses on examining developments in the field of readability during the past two decades with the goal of informing both current and future research and providing recommendations for present use. The fields of education, linguistics, cognitive science, psychology, discourse processing, and computer science have all made recent strides in developing new methods for predicting the difficulty of texts for various populations. However, there is a need for further development of these methods if they are to become widely available.

Keywords Readability · Text difficulty · Reading · Text analysis

A century of reading research paralleled a century of research into what makes one text more or less difficult to read and comprehend than another. Some estimate that by the 1980s, over 200 readability formulas had already been developed (DuBay 2004), and since the 1980s, the area has exploded in fields like discourse processing and computer science. The question that remains in the minds of educators and researchers alike is "what is the best way of determining the difficulty of a particular text?" Several reviews and summaries are available of the older, more traditional methods of assessing readability (e.g., Bormuth 1966; DuBay 2004; Klare 1974), and most researchers in the field discuss these classic methods by way of introduction to their own research. Controversy, however, has surrounded these older formulas, and new methods are constantly being developed and tested. The purposes of this review were to (a) examine recent developments in the field of

R. G. Benjamin (✉)
Department of Educational Psychology, University of Georgia, Athens, GA 30602, USA
e-mail: bben81@gmail.com

readability, (b) provide direction for future research, and (c) recommend appropriate readability analysis tools for educational and research settings.

The standard test of a text difficulty analysis method is how well its prediction of readability matches up with the actual reading comprehension scores of readers using existing texts. This very practical criterion has a simple and logical face value validity to it. However, the common problem with classic readability formulas is that once a formula is determined to be a good general predictor of a text's difficulty level, educational publishers and writers may then use the formula as a tool in the writing process in order to "scientifically" create texts at specific levels. The classic study by Davison and Kantor (1982) clearly demonstrates the shortcomings of this practice when they conducted a case study of four texts. Attempting to modify the texts via readability formulas was unsuccessful, and when the team made successful changes to make the texts easier to read, these changes actually ran counter to what the formulas would suggest. As Bormuth (1966) suggested, then, readability formulas are useful for measuring the factors *reflecting* a text's readability. When adapting texts, it was more important that adaptors were paying attention to the discourse structures of sentences, paragraphs, and the whole text rather than simply trying to shorten words and sentences.

Other critics have also complained that the practice of "writing to the formulas" does more harm than good (Schriver 2000) and that the assumptions underlying traditional readability formulas are far too simplistic to account for the varied linguistic and textual factors that can differ widely, especially in adult texts (Bailin and Grafstein 2001; Schriver 2000). Bailin and Grafstein (2001) argue that there is no single method for determining the readability of a text, and factors like style, vocabulary, background knowledge, grammar, textual coherence, and how well a reader can "fix" errors in a text likely interact with one another in making a text more or less difficult for readers.

The Need for Description and Evaluation of Text Difficulty Methods

Regardless of the problems with and criticisms surrounding traditional readability formulas, they are still popular today and continue to be used in reading research. However, in the past two decades, major technological advances have made way for the streamlining and automation of traditional readability formulas and have also sparked the development of more complex methods for measuring text difficulty. Researchers, administrators, and policy decision makers in the field of education, however, may need guidance on which methods might be most useful in research studies and in classrooms. Because of the rapid influx of newly designed procedures for analyzing text difficulty and matching readers and texts, educational researchers may either feel overwhelmed with the numerous options or may simply be tempted to stick with the classic methods they and their predecessors have used for decades. However, selecting appropriate texts for a population of readers requires some understanding of both the reader and the text, and different methods may be more or less appropriate for different types of texts and different populations of readers.

As research in the field of text difficulty analysis expands, researchers, administrators, publishers, and policy makers in education and other fields are increasingly responsible for defending their use or support of a particular readability tool. Likewise, so many developments have been made in the past two decades across disciplines that researchers in the text difficulty field need to guide their research with a wide variety of previous research in mind. A need, then, has arisen for both a description and evaluation of methods developed in the more classic style as well as emerging methods of assessing text difficulty and matching readers to texts. This need for exposure to and awareness of the rapidly expanding field of text difficulty analysis serves as the

motivation for this review. There are several goals for this paper. The first goal is to describe the development and evaluate the effectiveness of readability analysis methods which have emerged in the past two decades. This goal is met by describing the basic unifying elements within "types" of methods, pointing out their key differences, and identifying their strengths and weaknesses. The second goal builds upon the first: provide recommendations for practical use by evaluating each of these methods based on the following criteria:

- Is the method intended for/tested on a particular population?
- Is the method intended for/tested with particular texts?
- Has the method demonstrated predictive validity—especially in predicting readers' comprehension of texts?
- How widely available for immediate use is the method?
- How much special training is required to use the method?

The third and final goal is to consider these two decades of development in the field and make recommendations about what research needs to be done to advance the field of text difficulty analysis. For example, even though certain recent methods may not be ready now for wide-scale use, do the theoretical foundation for the method and preliminary tests of the method point to promising future applicability?

The following sections discuss the types of methods that have been developed in the past two decades and conclude by providing recommendations for using these methods in educational practice and research. The bulk of this paper describes and evaluates currently available tools and methods in readability research. These tools and methods are divided into three genres or types: (1) traditional methods, (2) methods inspired by cognitive science, and (3) methods based on the use of statistical language modeling tools. All methods discussed in this review are quantitative for the purpose of simplifying the comparison and evaluation of methods. The discussion of each readability genre concludes with some recommendations for future research directions. Following the discussion of methods developed in the past two decades is a section providing recommendations for current wide-scale use of existing readability methods. This final section is mainly intended for educators, administrators, and educational researchers.

Developments in the past two decades

Formulas based on traditional readability features

Although they have been on the receiving end of much criticism (e.g., Davison and Kantor 1982), several traditional readability formulas have remained popular and largely valid over the 90 years that they have existed as measures of text difficulty (Klare 1974). In the past two decades, technological advances have allowed researchers to automate the application of formulas and to test new variables for developing new formulas. Online corpora have automated word frequency analyses, and even the most basic word processing software can instantly report sentence length and word length statistics. In fact, readability formulas that utilize traditional variables like sentence length, percentage of familiar words, and word length are still being developed and have remained popular over the past two decades.

In their essential components, all traditional methods for computing readability are similar. They tend to incorporate some combination of easily measured units like sentence length, word length, and word frequency. Passages that contain shorter sentences, shorter

words, and more frequent words would be considered more readable or less difficult than passages with longer sentences, longer words, and rare words. The validity of these formulas for various readers is typically established by correlating reading comprehension scores with the formula's predicted readability of the texts. Of course, this technique can only result in a rough estimation of difficulty, and its weakness is that the formula might judge even a nonsense passage as quite readable if the text's jumbled words are frequent, short, and organized into brief sentences (see Davison and Kantor 1982 for a criticism). However, because of their wide use, simplicity, popularity, and some modern creative adaptations of traditional features like word frequency, several of the more recently developed methods are discussed below. These methods include the New Dale–Chall Readability Formula (Chall and Dale 1995), the Lexile framework (Smith *et al*. 1989), the Advantage-TASA Open Standard for Readability (ATOS) formula (School Renaissance Inst., Inc. 2000). A lesser known and less studied method is also briefly discussed to demonstrate the direction in which these traditional methods may be heading: the Read-X tool (Miltsakaki and Troutt 2007, 2008).

Recent traditional-style methods

The new Dale–Chall readability formula After publishing and observing the use of their popular Dale–Chall Readability Formula (Dale and Chall 1948), Edgar Dale and Jeanne Chall began to work on revising their formula in the 1970s. During these years of development, controversy regarding the use of traditional readability formulas came to a head, and several researchers began focusing on measures of assessing text difficulty that were based on theories in cognitive science (e.g., Kintsch and van Dijk 1978). Cognizant of the need to provide solid evidence of the validity of the new formula, when Chall published the revised formula, she included some methods for users to assess cognitive–structural elements of texts to better match readers to texts (Chall and Dale 1995). The new formula took sentence length and word familiarity into account, as the older formula did, but Dale and Chall improved upon the older version by updating and expanding their corpus of familiar words and validating the scale using cloze scores from the 32 passages used in Bormuth's (1971) extensive study of readability. Both the new and old formulas correlated highly with Bormuth's cloze mean scores (r=0.92 for both), and correlations with other criterion cloze scores were 0.85 or above (Caylor *et al*. 1973; Miller and Coleman 1967). Finally, the scale was also successfully validated by comparing predicted difficulty levels with various standardized reading tests. An additional improvement included expanding the range of grade levels that can be analyzed using the formula. While the old formula was recommended only for reading levels at grades 4 and above, the new formula can be used for assessing readability at grade 1 through college level.

The New Dale–Chall Readability Formula provides two readability measures for a text: a cloze score (the lower the score the more difficult the text) and a grade level. The authors recommend using the cloze score for research purposes as it allows for finer distinctions between texts. While this formula is subject to criticism for its failure to account for more complex structural relations within a text (e.g., lexical overlap, coherence, etc.), it has also received praise for potentially being the most valid of the popular traditional readability formulas (DuBay 2004). Indeed, interpretations of readability based on this formula may be at least moderately valid in a psycholinguistic sense as word frequency is an established indicator of the comprehensibility of texts (Just and Carpenter 1980).

The Lexile framework While the New Dale–Chall (1995) formula has gained appeal thanks to its user-friendly simplicity, another popular text-leveling method—developed in the late 1980s

but still flourishing in schools—is more complex in design: the Lexile scale (Smith *et al*. 1989). Creators of the Lexile scale attempted to design a scale that matched their construct definition of reading comprehension, thereby creating a measure with a high degree of construct validity in addition to the predictive validity sought by many traditional scale developers. Determining that reading comprehension depends upon "…the familiarity of the semantic units and…the complexity of the syntactic structures used in constructing the message" (Smith *et al*. 1989, p. 6), the creators devised a scale that included measures of word frequency (the semantic variable) and sentence length (a proxy for syntactic complexity). At this level, the scale is quite similar to the New Dale–Chall Readability Formula (Chall and Dale 1995). But the actual measures taken and the application of the Lexile scale differ from prior methods and cannot easily be computed manually as the Dale–Chall method can. The Lexile word frequency measure is the mean log word frequency from a 5 million word corpus (Carroll *et al*. 1971), and their sentence length measure is the log of the mean sentence length in the text.

Creators found that the Lexile scale is limited to use with continuous prose (as are most formulas), but correlated highly with item difficulties of numerous reading comprehension tests (r=0.84, 0.93 when corrected for range restriction and measurement error) [Smith *et al*. 1989]. The scale ratings also correlated highly with rank order difficulties of texts from 11 different basal series for grades 1–8 (r=0.83, 0.99 when adjusted for range restriction and measurement error) [Smith *et al*. 1989]. When compared with nine other readability formulas, however, the Lexile scale did not differ significantly in performance. The appeal of the scale for wide use seems to be largely based on its application: a person receives a Lexile score based on his or her ability to answer comprehension questions correctly; a text also receives a Lexile score. If the person and the text are matched, then the person has a 75% chance of answering a comprehension item correctly for that text. Teachers, then, can look at the Lexile score for a text and determine whether or not that text would be appropriate for a student based on the student's Lexile score.

In recent years, researchers—often the original developers of the scale—have written numerous defenses of the scale in addition to instructions for its use (Blackburn 2000; Stenner 1996, 1999; Stenner and Burdick 1997; Wright and Stenner 1998, 2000; Smith 2000a, b). The Lexile Framework for Reading is a commercial venture owned by Metametrics, Inc. Because the formula is complex and is performed on entire texts rather than samples, individual teachers and researchers are not able to perform analyses on particular texts without the assistance of the company. The regression equation and detailed explanations of the logit system and Lexile difficulty scale equation are available in the original NIH report (Smith *et al*. 1989) as well as a more recent MetaMetrics, Inc. publication (Stenner *et al*. 2007). A large body of literature, test items, and educational materials has been analyzed for Lexile scores, and the company is continually updating its library.

Advantage-TASA open standard for readability Like Metametrics, Inc., two large companies sought to develop a readability formula that would be widely applicable and used in schools nationwide. In 1998, researchers began work developing the ATOS formula (School Renaissance Inst., Inc. 2000). Renaissance Learning, Inc. (formerly Advantage Learning Systems, Inc., developers of the Accelerated Reader software) and Touchstone Applied Science Associates, Inc. (TASA, developers of the Degrees of Reading Power program) used their massive book and reading assessment databases to create two formulas: the ATOS for Text Readability Formula and the ATOS for Books Readability Formula. Both formulas are based on the same traditional variables—word length, sentence length, and grade level of words—though the formula for books also takes book length into account, a factor that was found to significantly influence book difficulty.

The development of the ATOS method involved comparing the predictive performance of numerous variables with the reading achievement scores of over 3 million students using the Accelerated Reader and STAR Reading programs—programs developed by Renaissance Learning, Inc. Final development and refinement of the formula involved examining the predictive value of numerous text variables when compared with the Rasch difficulty values of several hundred reading comprehension items. The final ATOS for Text Readability Formula incorporated three variables: average characters per word, average words per complete sentence, and the average grade level of the words which could be found in the ATOS graded vocabulary list (excluding the 100 most frequent words in the corpora). Validation of the formula was conducted using several hundred test items as well as authentic texts that had been used to validate the New Dale–Chall formula (Chall and Dale 1995). To convert the formula values to grade-level equivalents, researchers examined the performance and reading lists of thousands of student who had participated in the Accelerated Reader program and used these data to develop a grade equivalency conversion formula. An additional formula was also developed for use with books as ATOS researchers found that when other factors were controlled, students scored lower on Accelerated Reader quizzes for longer books than for shorter books. Development of the ATOS for Books Readability Formula simply involved adding a weighted book length variable that is adjusted based on general book length categories, e.g., whether the book has fewer than 500 words, more than 500 words, more than 5,000 words, more than 50,000 words, etc.

While the variables in the ATOS for Text formula individually explained over 80% of the variance in text difficulty during the development phase of the formula (Milone 2009; School Renaissance Inst., Inc. 2000), the study reports do not provide evidence of the formula's performance as a whole. This, of course, makes it difficult to determine the predictive validity of the formula—a significant weakness in this method. Because of the unknown or publicly inaccessible predictive value of the formula as a whole, independent studies are still needed to test the effectiveness of the ATOS formulas. The method's strengths lie in its extensive development phase and the large databank of student reading performance that the company was able to use in its development. The comprehensive research conducted in the development of ATOS, in addition to its wide applicability—the formulas are appropriate for grades K–12, adjustments are available for non-fiction texts, and conversion scales have been developed for Reading Recovery levels, Degrees of Reading Power, and Lexiles—lends to its large-scale use in educational settings.

Read-X A promising new text–user matching tool called Read-X is in development (Miltsakaki and Troutt 2007, 2008). This software uses some traditional readability variables—number of sentences, number of words, number of "long words," and number of letters in the text—to analyze the readability of texts on the Web in real time so that a person can perform a web search and filter results by reading level. The uniqueness of this program lies specifically in its ability to categorize search results by theme (e.g., science, music, history, etc.), and future versions of Read-X should be able to take a user's existing topical knowledge into account and customize reading level filtering based on a reader's level of content knowledge about a particular topic (e.g., long words in a particular domain are not necessarily difficult for a reader with a lot of background knowledge in that domain). This customization is possible because the authors developed theme-based corpora and gathered word frequencies based on theme (e.g., *atom* and *cell* both occur frequently in science texts, but are not frequent in general). Thus, the program is similar to the methods discussed above in its use of formulas to determine the readability of texts, but it differs in that it customizes the formulas for various domains of information.

Additionally, its future versions should do much more to take user knowledge into account than a more basic traditional method could ever do.

This program is promising for its potential use in school computer labs and libraries, where students often need to find information about a particular topic but may be overwhelmed by the difficulty of texts that result from their search. Future empirical research will determine the predictive validity of Read-X and its usability in educational settings. Meanwhile, developers are planning to examine whether psycholinguistic and discourse processing factors like syntactic complexity, propositional density, and rhetorical structure may improve the program's readability analysis (Miltsakaki and Troutt 2008).

Evaluation and directions for research

The four methods discussed in this section are similar in their use of traditional text difficulty analysis features like word frequency and length of sentences, words, and paragraphs. While available data regarding the performance of these methods can only be found for the New Dale–Chall and Lexile methods, it is likely that ATOS performs similarly since similar variables and methods of development were used for all three methods. These three methods have been designed to determine the readability of books and articles. The widespread use of readability formulas in business, research, and education—in addition to the large-scale commercial endeavors—demonstrates the continued relevance of these methods for typical use in analyzing general connected text. These methods are simple to use and have been around long enough that there are many user-friendly software programs available to analyze texts' readability. However, the original criticisms of these methods still hold true: (1) a jumbled passage would be judged just as readable as a sensible passage containing the same words and sentence lengths and (2) the temptation to "write to the formula" in creating leveled texts is likely too strong for most publishers to resist. Thus, these methods can still be used, but must be used only with appropriate texts (usually defined as authentic books or articles containing at least 300 words). Their use, then, is limited.

The Read-X method addresses the increasing need for analyzing the readability of web texts. While it is not yet sophisticated enough to take web images, captions, and non-standard text (words or phrases that can be used as "Internet shorthand"), it does move in the direction of automating user- and text-specific text difficulty analysis. Due to the increasingly widespread use of social media, Internet news sites, and interactive educational web sites, methods that can instantly determine the readability of web pages in a way that takes the user's knowledge as well as the page's non-traditional content into account will likely be highly useful. However, while traditional variables might be used to accomplish this, it is doubtful that a single traditional "formula" would be successful since variables like word frequency would have to draw from numerous domain-based corpora.

Methods inspired by advances in cognitive theory

As connectionist (McClelland and Rumelhart 1981; Rumelhart and McClelland 1982), schema (e.g., Anderson and Pichert 1978), prototype (Rosch *et al*. 1976), and spreading activation (Anderson 1983) theories emerged to explain how humans store and retrieve information in long-term memory, some researchers who studied text processing began to hypothesize that text difficulty and readability were more related to coherence and the relationships between elements in a text rather than simply the sum or averages of individual surface features (Britton and Gülgöz 1991; Kintsch 1988; McNamara and Kintsch 1996). Walter Kintsch's work with proposition density (Kintsch and Keenan 1973),

as well as his construction–integration and computational models of comprehension (Kintsch 1988; Kintsch and van Dijk 1978), inspired much of the recent work in text difficulty and matching readers to appropriate texts. This is especially the case since computer software like Coh-Metrix (Graesser *et al*. 2004) and WordNet (Fellbaum 1998) have made such analyses easily accessible to researchers.

Since many researchers sought to analyze text difficulty based on cognitive theories, several new methods and variables were developed. Because of the complexity involved in trying to represent cognitive processes involved in reading text, these methods often require automation. Before discussing the empirical work conducted to test these cognitively inspired tools, a brief introduction is provided to define some terms and describe some of the tools used in this work.

Variables and tools used in cognitively inspired readability methods

Propositions and inferences Kintsch and van Dijk (1978) set the stage for incorporating analysis of propositions and inferences into text difficulty analysis through their theoretical framework for text comprehension. Simply put, sentences can be broken down into propositions, or brief meaningful units that do not take into account information like tense, voice, or aspect (Graesser *et al*. 1997). Propositions are units comprised a predicate and at least one argument, and propositions can also include other propositions. An argument serves a functional purpose within a proposition, indicating the relationships between meaningful words in the sentence. For example, in the sentence *The nurse placed the scalpel on the table and grabbed the sponge*, the propositional breakdown is as follows:

1. Place (AGENT = nurse; OBJECT = scalpel; LOCATION = on table)
2. Grab (AGENT = nurse; OBJECT = sponge)
3. And (PROP 1; PROP 2)

In order to carry on coherent discourse or write a coherent text, there must be some propositional or at least argument overlap among successive sentences. Likewise, at the macro level, there must be some propositional connections across the larger text or throughout a conversation if the text or conversation is supposed to address a particular topic.

When there are few or no gaps in overlap across sentences, then a text is seamlessly moving from one point of information to another while giving the reader all the help he or she needs to build new knowledge. This type of text is a highly cohesive text: inferences are explicit and the reader does not have to fill many gaps using his or her own knowledge about the topic. However, in a text where less propositional overlap exists, the reader will be required to fill gaps of information with his or her own knowledge. This type of text has low cohesion: inferences are implicit and require more work on the part of a novice reader. Kintsch and van Dijk (1978) describe these latter texts as more difficult texts because a novice reader may not have the schema in place to make the necessary inferences to comprehend the text. Thus, in cognitively oriented text difficulty analysis, propositions and inferences play an important role, and the analysis and manipulation of these variables can yield significant differences in reader comprehension (Britton and Gülgöz 1991; Britton *et al*. 1993).

Latent semantic analysis Latent semantic analysis (LSA; defined and described in Landauer *et al*. 1998) is an automated tool that represents text content (e.g., an individual word and all the contexts in which it appears, for example) as a vector in semantic space. LSA has been used for many purposes including analyzing interview data (Dam and Kaufmann 2008), assessing

reading comprehension (see Millis *et al*. 2007), and scoring essays (Miller 2003). LSA analyzes the semantic relatedness either between texts or among segments of text in a more expanded way than simple measures of word overlap. The researchers can train the system on a large corpus of topical text so that it begins to develop a "knowledge" of which words tend to appear in particular contexts. For example, the word *music* probably appears frequently in the same contexts as *guitar*. Thus, *music* and *guitar* would have a strong semantic relationship and *music* would be considered to be an important word in texts containing the word *guitar*. This is a simplification of the numerous connections that LSA makes between words and contexts, but it describes the general principle.

Not only are these direct relationships analyzed, but LSA examines the indirect relationships among words in contexts as well. For example, in a text about trumpets where the word *music* exists but *guitar* does not exist, the word *guitar* will still have a positive semantic relationship to the word *trumpet* even though they may have never appeared in the same context. This happens because through *guitar*'s relationship to *music* and other words, the system understands that *guitar* is indirectly semantically related to *trumpet*. Mathematically, this is accomplished by using matrices to capture a word and its relatedness to all the other words in a text, paragraph, or sentence. Each word is a vector in multidimensional semantic space with rows in the vector being the contexts (lines, paragraphs, etc.) in which the word appears. These vectors are linked by words appearing in proximity to other words. The cosine between two vectors provides a numerical value of the relationship between two vectors.

Complete texts can be represented as a vector as well, with the text vector being the average of the vectors of the words within the text. The system, then, can get an idea of which words are most likely to appear in which contexts, and it can determine the importance of particular words in particular contexts based on the strength of the relationships among words via the training corpus. Because an entire text, sentence, or paragraph can be represented as a vector as well, the relatedness of these various units can be compared. A more technical explanation of how this all works is too involved for this review, but LSA is discussed in greater detail by Landauer *et al*. (1998) and Foltz *et al*. (1998).

As a readability tool, LSA has been used to match readers to appropriately difficult texts (Wolfe *et al*. 1998) as well as to gauge the cohesiveness of a given text (Foltz *et al*. 1998). To match readers to texts, LSA can be used to compare the semantic relatedness of several texts about global warming, for example, with a student's composition about global warming. The belief is that the student is most likely to learn more from a text that more closely relates to the student's prior knowledge. LSA provides a measure of this relationship by creating vectors for each text, allowing researchers to compare cosine between text A and the student composition with the cosine between text B and the student composition. The higher the cosine, the more closely related the texts.

To estimate the difficulty of a text, LSA can report on the cohesiveness of the text: LSA is used to compare the semantic relatedness of adjoining sentences, for example, to determine how closely they connect. Texts in which there is a high degree of cohesion tend to be easier for non-expert readers to read than texts in which more connections have to be made by the reader (McNamara and Kintsch 1996; McNamara *et al*. 1996).

Empirical studies of cognitively based text difficulty analysis

The following studies illustrate the use of cognitively based variables and tools for text difficulty analysis. They are discussed in a roughly historical sequence, beginning with studies examining propositions, inferences, and text cohesion. Following these studies

which are based largely on early work by Kintsch and colleagues, studies using a program called Coh-Metrix (Graesser *et al*. 2004) are examined. Coh-Metrix can conduct LSA as well as analyze text cohesion and dozens of other variables. As an alternative to methods based largely on Kintsch's text comprehension model, some researchers have recently used prototype theory (Rosch *et al*. 1976) as the basis for a study incorporating semantic networks. Finally, a new software program called DeLite has been developed by German researchers and serves as a kind of bridge to the statistical language modeling methods that are discussed later in this paper. This section concludes with some guidance for future research in cognitively based text difficulty analysis.

Revising texts for better comprehension using inference analysis Using principles developed based on Kintsch's model of text comprehension (Kintsch and van Dijk 1978), Britton and Gulgoz (1991) developed principles for revising texts at a local level to improve comprehension. They took a text that was used to train Air Force recruits and used Kintsch's computer program (Miller and Kintsch 1980) to find places in the text where inferences were lacking. Having devised modification principles based on Kintsch's theory, they modified the text by linking each sentence to the previous sentence via overlapping propositions and arguments using only one term for each concept that appeared in the text, arranging sentences so that old information precedes new information, and making important implicit inferences explicit for the reader. The authors found that participants performed better on free recall tasks and multiple-choice inference questions when given the revised version rather than the original version of the text even though traditional readability statistics (e.g., Flesch–Kincaid, Coleman–Liau, and Automated Readability Index) between the passages were the same. Additionally, when the novices' ratings of relationships among terms were compared to experts' ratings (experts had read only the original text), ratings of novices who had read the revised text correlated much higher with the experts than ratings of novices who had read the original text.

Britton *et al*. (1993) later found similar results when they conducted a review of studies in which textbooks had been revised according to similar principles, providing a promising contrast to studies in which revisions made according to readability formulas had little effect (e.g., Coleman 1962; Klare 1963). These studies by Britton and colleagues demonstrate that even if readability formulas are not able to discern differences between texts, analyses of explicit inferences within a text can show that one text is more comprehensible—at least for novices—than another.

Text cohesion and reader knowledge The important distinction between high-knowledge and low-knowledge (or novice) readers is highlighted in two studies (McNamara and Kintsch 1996; McNamara *et al*. 1996) in which low-knowledge and high-knowledge participants were given either a low-cohesion or high-cohesion text to read about a given informational topic. The difference between the two types of texts lies largely in the number and type of inferences a reader has to make in order to form a coherent mental representation of the content. A low-cohesion text requires a reader to make more inferences, while a high-cohesion text tends to provide more information explicitly for the reader. Both studies found that in general, low-knowledge readers benefited from high-cohesion texts while high-knowledge readers benefited from low-cohesion texts.

The explanation for this result is based on Kintsch's construction–integration model of reading comprehension (Kintsch 1988); high-knowledge readers need some obstacles placed in their path to promote deeper-level processing. Low cohesion requires these knowledgeable readers to make inferences which activate and strengthen semantic

networks. However, low-knowledge readers need the obstacles removed because they do not have the prior knowledge necessary to construct appropriate models of text content, i.e., if inferences are not made for them, these readers will often fail to make the inferences necessary to fully comprehend a text. This high/low-cohesion distinction is similar to recent findings in cognitive load theory, demonstrating that certain efforts to decrease cognitive load for learners can be beneficial for novices but detrimental for experts (e.g., Leahy *et al*. 2003; Kalyuga *et al*. 2003, 2000, 2001a, b).

Use of Coh-Metrix in text difficulty analysis As the studies above have contributed to the understanding of what factors might make texts differentially difficult for readers of a similar general reading skill, the development of various software programs have allowed researchers to develop new methods for determining the difficulty of texts based on traditional readability features as well as features of cohesion that indicate a text's level of coherence. Note that while coherence refers to how propositions are connected in a reader's mental representation (a psychological construct) and cannot be measured using computational surface indicators, cohesion refers to surface indicators of how sentences are related to one another in a text (a text construct). This is accomplished through examining propositional or argument overlap and can be measured using software like Coh-Metrix (Graesser *et al*. 2004), which reports over 50 indices of language, cohesion, and text difficulty.

Coh-Metrix also conducts LSA. As the above-described studies (McNamara and Kintsch 1996; McNamara *et al*. 1996) point to the importance of gauging a reader's level of background knowledge when selecting texts for learning, Wolfe *et al*. (1998) attempted to use LSA to match students to texts of appropriate difficulty. They had undergraduates and medical students write brief essays about the human heart and then randomly assigned each of the participants to one of four texts about the human heart: one text designed for children, one general text for adults, one from an undergraduate textbook, and one for medical students. Participants' knowledge was tested after studying the assigned texts. After analyzing the results, the researchers predicted that had LSA been used to match readers to texts based on a comparison of the pre-essays to the texts, learning would have improved 53%. Unfortunately, though, they did not go on to test this prediction in an additional experiment.

Foltz *et al*. (1998) used LSA to determine the coherence of original and revised texts from two studies (Britton and Gülgöz 1991; McNamara *et al*. 1996). They found that LSA accurately predicted distinctions between texts and accurately predicted comprehension outcomes. Furthermore, LSA performed significantly better than simple measures of word overlap. LSA has been used for many other purposes (see volume 25 of *Discourse Processes*, issue 2/3—dedicated to research regarding LSA) which fall outside the scope of this article, but it shows promise as a method for assessing the readability of texts especially in relationship with reader background knowledge.

Other linguistic features utilized by Coh-Metrix have also been used to develop measures of text difficulty for predicting comprehension for both L1 and L2 (discussed below) English speakers. Three Coh-Metrix variables—number of words per sentence, argument overlap, and CELEX word frequency scores—were combined to develop a readability index (Crossley *et al*. 2007). Using Bormuth's (1971) 32 academic texts as their text corpus and Bormuth's mean cloze scores as the dependent variable, the three variables predicted cloze scores with an adjusted R^2 of 0.90. However, employing Bormuth's (1969) formula which utilizes number of letters per word, number of Dale–Chall words per total words (authors used the updated list), and number of words per sentence, the authors achieved an adjusted R^2 of 0.92. While the more sophisticated Coh-Metrix variables did not perform any better than the traditional variables, it is critical to note that Bormuth's corpus

(1971) is not necessarily representative of all texts. Because Coh-Metrix measures variables that are more in line with theories about how humans process text, further studies with a greater variety of texts will likely favor the newer methods. Crossley *et al*. (2007) point out that if a determination about superior methods is to be made, then studies using larger corpora are necessary.

More recent studies using Coh-Metrix variables focus on using the tool for making writing quality distinctions (McNamara *et al*. 2010) and distinguishing between the relative cohesion of texts (McNamara *et al*. 2010). While these functions are important and seem to be areas in which Coh-Metrix stands out as a useful tool, these functions are not necessarily precise measures of readability, and so in-depth discussion of these studies is beyond the scope of this review.

Semantic networks While the studies above examine text difficulty largely in light propositional networks, Lin *et al*. (2009) believed that text difficulty would be a function of semantic networks—lexical relationships such as those described in WordNet software (Fellbaum 1998)—in accordance with prototype theory (Rosch *et al*. 1976). In light of the role that word length often plays as a measure of lexical complexity for determining text difficulty, Lin and colleagues tested the hypothesis that *relative* word length is what really matters (i.e., the basic level of a noun is typically the shortest version of that noun, e.g., *red*). Hypernymns (words that subsume the basic level noun in a semantic network, e.g., *color*) tend to be longer, while hyponyms (words that are subsumed by the basic level noun, e.g., *crimson*) are virtually always longer than their basic-level noun. Basic-level words also tend to be less morphologically complex. Thus, the authors propose that absolute word length is not really what matters when determining lexical complexity in a text; relative word length (the length of a noun as compared with the length of its basic noun form) is a more precise determiner. However, Lin *et al*. (2009) also found that sometimes hyponyms can also function cognitively as basic-level words (e.g., *card* is a hyponym of *paper*, but is often used to form compounds and tends to behave as a basic level noun).

Lin *et al*. (2009) determined to find basic-level words in a text via two filter conditions: compound ratios and word length differences between the target word and all hyponyms. Using the ratio of basic-level nouns in texts as their index, the authors compared their text difficulty levels to other readability measures—using online graded readings as the texts—and found that their measure ordinally matched the grade-level progression more accurately than the traditional readability formulas. The readability formulas used included the New Dale–Chall formula, Spache, Powers–Sumner–Kearl, Flesch–Kincaid, FOG, SMOG, and FORECAST. These formulas tended to indicate a dip in difficulty for the grade 7–8 text, meaning that the grade 4–6 text was described by the formulas as more difficult than the grade 7–8 text. The method of Lin *et al*. (2009) did not result in such an error.

A weakness in this study lies in its use of online graded texts rather than texts matched with students' comprehension scores; the authors never describe the methods used to grade these texts, so it is impossible to know whether they were "written to a formula" or authentic. Also, even though this method can be fully automated and, therefore, simple to use, it needs to be tested on a larger corpus of texts and text types for a more complete validation, something which the authors are, no doubt, considering. Because this method relies on psychological theory-based variables, it is likely that studies using larger corpora will further validate this technique.

DeLite software Bridging the gap between these cognitive theory-inspired methods and the statistical language modeling methods of computer scientists lies the DeLite/

EnLite program (vor der Brück *et al*. 2008). This program was originally designed for German, but a prototype English version has been created as well. Published testing and evaluation, however, was conducted with the German version. The appeal of this program lies in its combination of what the authors call *surface* and *deep* indicators of both syntactic and semantic complexity—and it is the inclusion of these deep indicators that the authors believe makes this system more psychologically valid. However, like the statistical language modeling methods that follow, this program conducts its final layer of text analysis by comparing the text in question with what it has "learned" about leveled texts based on prior training data. That is, program designers had over 3,000 ratings (seven-point Likert-type scale) of 500 texts by 300 participants from an online readability study. Some of these texts and readability ratings were used to train the program to recognize certain combinations of syntactic and semantic features at particular levels of readability. The rest of the texts were used for testing and cross-validation. Some of the syntactic indicators that the program examined were depth of embedded clauses (e.g., *He left the house where the woman he loved lived immediately* causes difficulty for the reader because the reader has to hold the main clause in memory while reading the subordinate clauses) and number of words per noun phrase. Some of the semantic indicators included semantic network quality (i.e., are semantic connections complete and clear), number of propositions per sentence, and length of causal and concessive chains (i.e., are there too many ideas in too few words).

Interestingly, testing and cross-validation of this system showed that the indicator with the greatest weight was the basic surface-level indicator of sentence length (vor der Brück *et al*. 2008). The use of both surface and deep indicators together, however, resulted in the best performance. DeLite's readability predictions correlated more highly with participants' difficulty ratings than did Flesch–Kincaid predictions (r=0.53 vs. 0.43, respectively). Additionally, the authors used local government texts in this study rather than newspapers or general texts, and the Flesch–Kincaid formula might not perform well with texts containing a lot of specialized language, numbers, and symbols.

While the DeLite program improved upon the simple Flesch–Kincaid formula for this corpus of texts, the program's predictions only accounted for 28% of the variance among ratings—a much lower percentage than is typically expected among text difficulty measures—and the method was not tested by comparing text difficulty predictions with actual comprehension scores. Further research using more traditional texts may result in improved performance, but the software also should be tested against more recent traditional measures like the New Dale–Chall, ATOS, and Lexile.

Evaluation and directions for research

Text difficulty analysis via methods inspired by developments in cognitive science is a field still in development. Much of this work has moved beyond traditional readability methods by explaining what might make texts difficult for different readers (McNamara and Kintsch 1996; McNamara *et al*. 1996; Wolfe *et al*. 1998), how principled revisions of texts can improve readability (Britton and Gülgöz 1991; Britton *et al*. 1993), and how complex cognitive indicators can be objectively measured through automated language processing software (Graesser *et al*. 2004; Lin *et al*. 2009; vor der Brück *et al*. 2008). While these studies have helped researchers and educators better explain text difficulty, some traditional readability indicators seem to perform as well as the newer more complex methods (e.g., in Crossley *et al*. 2007, sentence length and word frequency were the most powerful indicators).

However, these methods are still relatively new, and multiple opportunities exist to demonstrate the strengths of methods based on advances in cognitive science. One of the greatest weaknesses of solely relying on variables like sentence or word length and word frequency is that a passage in which the sentences are in random order can be judged to be just as difficult as a passage that actually makes sense. Additionally, a poorly written text with short sentences and common words might fool a traditional readability formula. Propositional analysis, however, could reveal the lack of cohesion in such texts. Combining traditional variables with more sophisticated ones could presumably provide a safety net for users of automated text difficulty analysis tools.

These cognitively based methods can be improved and validated through additional testing on large, diverse corpora. Empirical research should determine at what age group or reading level these methods can begin to be reliably used. A text designed for very young children, for instance, may not be complex enough for LSA or propositional analysis to work properly. Additionally, designing user-friendly tools and programs utilizing these methods is a critical step if these techniques are to be widely used. WordNet and Coh-Metrix are powerful programs, but the learning curve may be steep for many of the individuals who could benefit from their use. Likewise, their complexity could result in misuse. Some refinement of these programs in upcoming years to a more user-friendly interface could expand their use dramatically and result in fewer researchers exclusively using traditional readability formulas simply because of their simplicity. The DeLite software described above seems to reflect strides taken in this direction.

Findings in statistical language modeling

A common feature of several recent studies is their focus on improving readability analyses for Web pages (which did not exist when most traditional readability methods were developed) as well as more conventional informational texts. Traditional formulas have often performed poorly when analyzing Web documents (Collins-Thompson and Callan 2004; Schwarm and Ostendorf 2005; Si and Callan 2001), a phenomenon which may be attributed to the significant amount of "noise" found in web documents (i.e., punctuation errors, sidebar menus, photograph captions) as well as the large number of web pages containing fewer than 100 words. Advances in computer science through statistical language models (SLMs) and support vector machines (SVMs) have made new types of studies possible. Both methods of analyzing texts function as classifiers based on training data, and both methods are briefly introduced below prior to discussing the developmental and empirical research that has been conducted using these methods.

Tools used in text difficulty research

Statistical language modeling The type of SLM technique used in these studies is based on the probability that a particular word or words were generated by a language model of a particular grade level (e.g., a language model for grade 5) without regard to the surrounding context. For example, statistical analysis of a large text corpus reveals that the word *red* is more likely to appear in texts designed for the primary grades than the middle school grades, regardless of the topic or context. This method is well known among many educational and psychological researchers.

Here is a concrete example of how this works: to build the appropriate models, the program is given a corpus of texts which have been labeled by their grade level. This allows the program to build a language model for each grade level which basically analyzes the

probability of particular words appearing in texts at particular grade levels. Once the program has been trained, the researcher can give it a new text. Based on the statistical analysis performed on the training data, the program can analyze the text's words and/or combinations of words to assign the text to the model most likely to generate that text. So if the new text has words that are more likely to be generated by a third grade model than any other grade, the program classifies the text as a third grade text based on the texts it was trained on. While the simplest SLMs only look at the likelihood of individual words being generated by a model, free of any surrounding context, SLMs can also be designed to look at strings of words. These more complex models, however, require a greater amount of training data (Si and Callan 2001).

Support vector machines SVMs allow a researcher to take a variety of input and classify it based on previous training data. SVMs are helpful when attempting to determine the types of grammatical features and patterns that might be more common in third grade texts, for example, as opposed to sixth grade texts. The SVM, like the SLM, develops grade-level text models by learning the characteristics of texts and then determining the likelihood that a new text was generated by one of the grade-level models. While SVMs are trained and tested similarly to ordinary SLMs, the benefit of using SVMs is that they can take the information from SLMs and incorporate it into a more complex classifier that can also include traditional readability features as well as more complex grammatical parsing features. This results in very powerful grade-level text classifiers that can be customized according to the needs of the researcher. SVMs may or may not use information from SLMs; the features incorporated into an SVM classifier simply depends upon the desire of the researcher.

Text difficulty research using statistical language modeling

In the following studies, researchers have used SLMs and/or SVMs to classify texts into grade-level categories. While the studies are presented in a roughly historical manner, they are intentionally grouped categorically rather than chronologically. Basic text difficulty research using SLMs is followed by refinements to the initial SLM techniques. Studies follow which add grammatical features to SLMs in attempts to improve performance. Finally, studies which incorporate numerous features into SVM classifiers suggest considerable potential for these tools as a means of analyzing text readability.

Simple SLMs in text difficulty analysis: Si and Callan (2001) In the "classic" study by Si and Callan (2001), it was hypothesized that readability estimates could be valid on a wider variety of document types if the estimates were based on the actual content of the text rather than surface features like sentence length, syllable counts, etc. Si and Callan focused their attention on Web documents, which can be problematic due to their often non-traditional layout, formatting errors, and sentence fragments. The authors trained their three models on an admittedly small corpus of 30 science web pages that had been written for students at particular grades levels (K–2, 3–5, and 6–8), ten pages for each model. Thus, the K–2 model, for example, was able to learn which words were likely to be generated by K–2 texts. After training, the authors tested the model on the remaining 61 Web pages they had collected and found that the best performance occurred when they combined their models with sentence length, for a total of 75.4% accuracy in predicting the grade categories of documents from the test set. For comparison, they found that the Flesch–Kincaid formula only performed at 21.3% accuracy.

Subsequent studies have expanded Si and Callan's (2001) findings and contributed to the field of readability for Web content in several ways: increasing training and test corpora as well as the specificity of grade-level assignments (Collins-Thompson and Callan 2004, 2005), including grammatical features in text analyses (Heilman *et al*. 2007, 2008; Schwarm and Ostendorf 2005), examining the application of machine methods for L2 learners (discussed in the following section; Heilman *et al*. 2007; Peterson and Ostendorf 2009), and exploring the potential for online readability decisions based on user search engine queries (Liu *et al*. 2004).

Refining the technique: Collins-Thompson and Callan (2004, 2005) Since Si and Callan (2001) used a small corpus of Web texts in their study and categorized texts into only three grade categories, Collins-Thompson and Callan (2004, 2005) expanded the 2001 research by building a substantially larger corpus, classifying texts into 12 grade-level designations, using cross-validation to test reliability, and applying a "smoothing" process which removes words that only occur once or twice in the corpus and could skew the analysis. While authors found that traditional readability indices performed equal to or better than the new model set on controlled, expert-labeled test sets, traditional indices' performance dropped significantly when tested on the Web corpus, while the new model set remained stable (correlating 0.64–0.79 with the grade levels specified on the web sites). Especially significant is the finding that the authors' model set performed significantly better than a traditionally high-performing index (percent of unknown tokens) when analyzing short texts (fewer than 100 words) at the primary grade levels. The current commonly used short-text readability index—the Fry Short Passage Readability Formula (Fry 1990)—has only been recommended for fourth grade and above, while the method developed by Collins-Thompson and Callan was tested on grades 1 through 12.

Studies incorporating grammatical features and error analysis By enlarging prior models to include syntactic features (Heilman *et al*. 2007, 2008; Schwarm and Ostendorf 2005) and changing the evaluation approach from examining correlations to also examining error levels (thus looking at the magnitude of a judgment error rather than simply whether or not a text was classified correctly; for example, classifying a fourth grade text as an eighth grade text is a greater error than classifying it as a fifth grade text), some of the most recent studies have moved toward potentially developing high-quality alternatives to traditional formulas for classifying both Web (Heilman *et al*. 2007, 2008) and traditional text (Schwarm and Ostendorf 2005). For Web texts, the use of grammatical feature sets have had mixed results: Heilman *et al*. (2007) found that the grammatical features could moderately improve performance when added to SLMs, but Heilman *et al*. (2008) found that by expanding the set of grammatical features using a context-free grammar parser, the grammatical features alone could perform well as readability predictors for Web texts. However, Heilman *et al*. (2007) did not compare their method with any traditional readability formulas, and the comparison of Heilman *et al*. (2008) with the Flesch–Kincaid formula and a Lexile-like measure showed that both measures performed almost as well as their own. The benefit of the methods developed by Heilman and colleagues is their applicability for automatically analyzing documents on the Web, but based on the good performance of the Lexile-like measure, some of the newer commercial formulas like Lexile and ATOS might perform adequately with some minor adjustments.

Schwarm and Ostendorf (2005) compiled a corpus of Weekly Reader, Encyclopedia Britannica, Britannica Elementary, CNN, and CNN abridged texts to train their SLM classifiers and their grammatical parser. They then trained, developed, and tested their SVM

classifiers on a Weekly Reader corpus (for grades 2–5), dividing it into three smaller corpora, one for each stage of the process. Their SVM classifiers included indices like sentence length, syllables per word, Flesch–Kincaid score, the syntactic parse features, and the analyses of the SLMs. Once in the testing phase, the authors used the percentage of articles that were misclassified by more than one grade level as the error criteria. The authors' classifier performed with error rates ranging from 5.5% to 21%, Flesh–Kincaid error rates ranged from 59% to 78%, and Lexile error rates ranged from 24% to 33%. Clearly, the authors' classifier outperformed the others on this test set. The weakness of this research is simply that the SVM classifier should be developed and tested for use with a much larger and more diverse corpus in order to demonstrate its superiority over the more traditional formulas.

Utilizing Internet search engines: Liu et al. (2004) Machine methods may prove useful not only for analyzing the difficulty of texts but also for determining the reading-level category of the user. Liu *et al*. (2004) developed a method for analyzing search queries in online search engines, determining the reading-level category of the user based on the query, and returning texts that match the grade-level category of the user. Of course, search engines have improved dramatically since the 2004 study, and the authors' grade categories were very broad (K–6, 7–9, 10–12, undergraduate, graduate). However, the SVM classifier they developed using both SLMs as well as numerous syntactic features performed with 66–96% accuracy when given two-category combinations.

It is impressive that the SVM method was able to work so well given the paucity of input in a typical search engine query. Though the two-category combinations result in a very broad classification of texts, limited search engine results to those classified as appropriate for grades K–9 could be quite helpful for an elementary school student searching for information on the earth's crust, for example. However, until this type of approach is refined, it is probably not helpful for readers who can only comprehend texts within a very limited difficulty range. Of course, it would be highly convenient to be able to use search engine queries to determine a user's reading level and supply him or her with appropriate texts. Search engines, though, have developed significantly since the 2004 study, and it remains to be seen whether this method would be helpful with the current capabilities of search tools like Google or Bing.

Summary and evaluation

Language modeling provides a powerful means for expanding simple word frequency indices by analyzing the probability of a particular model (e.g., a model for fifth grade texts) generating a word or combination of words. Adding other features to the SLMs, though, seems to improve performance when classifying Web texts (Heilman *et al*. 2007, 2008) as well as informational documents (Schwarm and Ostendorf 2005). The major weakness common to all these studies is that their standard for determining performance accuracy is often a grade-level label that a text author gives to the text. It is impossible to fully compare these machine learning methods to methods like ATOS, Lexile, Read-X, and others without more information on how well readers actually comprehend these texts. Researchers developing these methods are largely computer scientists rather than reading researchers, and while machine-learning methods certainly show some promise, it is important that scientists from the fields of computer science, reading, and psychometrics combine their efforts to make substantial gains in developing more universal methods of determining readability.

Of course, the SLM/SVM methods require the classifiers to be trained on texts that have already been deemed appropriate for particular grade levels or groupings. Thus, they do not stand alone as tools for measuring text difficulty. Rather, once other methods have determined the difficulty of the texts in the training corpus, these tools could be used to quickly and automatically categorize numerous texts, including Web texts, using sophisticated statistical techniques. For now, these studies simply demonstrate the beginnings of potential for using SLMs and SVMs as graded text classifiers.

General Recommendations for Educational Use

In the past two decades, readability research has expanded far beyond simple regression equations. Readers no longer simply read published text on a page; rather, much of what modern readers consume is on the Internet. Different methods of assessing text readability have begun to be better tailored for these media differences as well as distinctions in domain and variation in human beings. Anyone recommending particular methods of readability, then, must appropriately judge these methods based on various criteria.

The recommendations here are based on several factors which have been summarized for simplicity (see Table 1):

- Is the method intended for/tested on a particular population?
- Is the method intended for/tested with particular texts?
- Has the method demonstrated predictive validity—especially in predicting readers' comprehension of texts?
- How widely available for immediate use is the method?
- How much special training is required to use the method?

Easily available methods that make valid predictions of readability on a wide range of texts for a diverse population will receive the highest recommendations. The reading populace can be divided into at least three easily definable groups; recommendations will be made for each group: adults, student readers, young emerging readers.

Recommendations for use with adults

Much of the quantitative readability work in the past two decades has focused on analyzing text difficulty for skilled readers or at least readers who have acquired fundamental reading skills. This work also varies by purpose—whether the metric or method focuses on general texts, specialized documents, Web pages, domain-specific texts, or specific populations. Though many adults still enjoy the novels and essays that are often required reading for school children, adults also have to navigate their way through news, technical writing, and informational texts intended for members of special fields as well as the general public. Not many of the methods utilized in these areas, though, are available for general use (for research in specialty texts and special populations, see Kirsch and Mosenthal 1990; Meyer *et al.* 1993; Kim *et al.* 2007; Feng 2009; Green *et al.* 2010; Heilman *et al.* 2007; Peterson and Ostendorf 2009).

Generally, developments in text difficulty assessment in the past two decades demonstrate that some of the more traditional readability variables still work well for typical books and texts. For general reading of connected text, the New Dale–Chall Readability Formula (Chall and Dale 1995), Lexile (Smith *et al.* 1989), and ATOS (School Renaissance Inst., Inc. 2000) have been widely tested and proven to be largely reliable and valid for predicting comprehension of texts among readers of wide-ranging ages and abilities. The Coh-Metrix formula (Crossley

Table 1 Summary of general-purpose text analysis methods

Name or author of method	Tested population	Texts	Evidence of validity	Widely available	Easy to use
New Dale–Chall Readability Formula	Grades 1 through adults	Continuous prose	Yes. R^2=0.85 with cloze scores R^2=0.91 with Bormuth's (1971) cloze scores	Yes	Yes Manual and automated
Lexile scale	Grades 1 through adults	Continuous prose - books	Yes. R^2=0.71 with test item difficulties R^2=0.69 with rank order of basal texts	Yes—commercial	Yes Automated
ATOS	K–12	Continuous prose	Yes. $R^2 \geq 0.80$ with comp scores	Yes—commercial for schools	Yes Automated
Read-X	None Formula is in development	Web texts	None	No	Yes
LSA for matching texts to readers (Wolfe *et al.* 1998)	College students	Academic texts	N/A	Yes	No—must collect prior knowledge samples from readers
LSA for predicting text comprehension based on coherence (Foltz *et al.* 1998)	College students	Academic texts	N/A	Yes	Yes—with access to Coh-Metrix
Coh-Metrix formula (Crossley *et al.* 2007)	None—tested against Bormuth's corpus	Bormuth's corpus (school texts)	Yes. R^2=0.90 with Bormuth's (1971) mean cloze scores	No—no formula was designed	No—requires conducting a regression analysis
WordNet relations (Lin *et al.* 2009)	None—tested against online graded texts	Online graded texts	Little. No computed statistics, only graphs.	No	No
DeLite/EnLite	Adults	Municipal texts	Yes. R^2=0.28 with participant difficulty ratings	No	No—not developed yet
SLM by Si and Callan (2001)	None—K-8 texts	Science Web pages labeled by grade level	Yes. 75% accuracy in predicting grade-level category (K–2, 3–5, 6–8)	No	No
SLM by Collins-Thompson and Callan (2004, 2005)	None—grades 1–12 texts	Web pages labeled by grade level	Yes. R^2=0.41–0.62 with designated grade levels of texts	No	No

Table 1 (continued)

Name or author of method	Tested population	Texts	Evidence of validity	Widely available	Easy to use
SLM by Heilman *et al.* (2008)	None—grades 1–12 web texts	Web pages labeled by grade level	Yes. R^2=0.58 with labeled grade levels	No	No
SLM by Schwarm and Ostendorf (2005)	None—grade 2–5 texts	Web pages labeled by grade level	Yes. Error rates ranged from 5.5% to 21%, much lower than FK and Lexile	No	No
Liu *et al.* (2004)	None—K-6, 7–9, and 10–12 texts	Web pages labeled by grade	Yes, but limited. 66–96% accuracy in predicting which grade-level 'bin' a text belonged in when given two-category combinations	No	No

et al. 2007) is promising, but still underdeveloped; however, because of its inclusion of deeper cognitive features, it presumably will be applicable to a wider variety of texts than traditional formulas. Further testing would need to be done to continue to validate this method on corpora other than Bormuth's (1971) text corpus. Also, the Coh-Metrix formula has not yet been tuned to determine the particular difficulty level or grade level of a text; it has only been used to judge the relative difficulty of texts. Coh-Metrix itself, however, could and should be used for analyzing texts for literate adult readers. A text's cohesion affects a reader's ability to comprehend a text, as illustrated by the studies in cognitively based text difficulty analysis. Thus, LSA and variables like argument and propositional overlap can actually provide means of measuring the difficulty as well as the quality of texts that may be too sophisticated for typical readability formulas (which have nearly always been developed with school children in mind). Furthermore, analysis using LSA and propositions can be used to revise poorly written texts, a task at which traditional readability formulas have performed notoriously poorly.

Recommendations for use with school children

Prior to the late 1980s, many readability scales for children were only validated for grades 4 and above. Many methods discussed in this review, however, are recommended for use with texts as simple as those designed for children at emergent reading levels. The category of "schoolchildren" here, though, is defined as those students who have reached the stage of "reading to learn" rather than simply "learning to read." Thus, recommendations in this section consider students largely in grades 4–12. The appropriateness of readability methods for emergent readers will be discussed in a following section.

If a school wants to use a single method for determining the readability of texts for readers of all ages and skill levels, the New Dale–Chall Readability Formula (Chall and Dale 1995), Lexile (Smith *et al*. 1989), and ATOS (School Renaissance Inst., Inc. 2000) methods seem to be specifically directed toward this population and have been studied and researched extensively. At times, however, difficulty in determining readability can arise when examining content area textbooks. For this, Chall *et al*. (1996) Qualitative Readability Scales can be helpful in resolving ambiguities. Teachers, reading experts, and students are likely able to detect readability issues that automated scales cannot. Since the criterion passages of Chall *et al*. were tested against user ratings for appropriate leveling, these scales can provide additional support for decision making when student texts are selected. For a more automated approach, ATOS researchers found non-fiction books to be 0.42 grade levels more difficult, on average, than fiction books. Thus, educators and parents could make that simple adjustment when choosing books for children.

Finally, though they are not yet ready for widespread use, the most promising tools for analyzing the difficulty of Web texts for school children appear to be variations on SLMs and SVMs. This research is so new that these methods are all still in development, but with continued research testing and refining these tools with large text corpora and validating them with actual reader comprehension data, these methods should allow for considerable flexibility when analyzing already existing texts.

Recommendations for use with emergent readers

Assessing text difficulty for emergent and developing readers provides a particular challenge for text difficulty researchers. While many of the currently available formulas and automated systems are advertised as being appropriate for use with children as young as first grade, some

researchers would argue that there are relevant features in texts for young children that simply cannot be measured via an automated or generalized system, and multiple criteria are necessary for determining appropriate texts for developing readers (e.g., Hiebert 1999).

Qualitative methods of assessing readability have seemed especially appropriate and have been popular when used for emerging readers (Hiebert and Pearson 2010). While readers are learning to read, factors like phonological regularity, number of words on a page, quality and specificity of images, size of print, number of syllables, and other factors can play a role in whether or not the child will have a successful reading experience while developing his or her reading skills (Hiebert 1999; Rog and Burton 2002). For this reason, the quantitative methods discussed in this review may not be the most appropriate for very young beginning readers. Teachers in the primary grades might consider using qualitative leveling methods (e.g., Fountas and Pinnell 1999, 2001; Peterson 1991; Rog and Burton 2002) when selecting books for their students. If, in a particular school district, teachers are not the primary persons responsible for choosing and leveling particular books in their classrooms, then teachers and administrators might request that their publishers begin to provide text levels based on some of these qualitative leveling methods.

Popular newer readability methods like Coh-Metrix variables and Lexiles have not been previously tested as tools for discriminating between multiple levels of texts within the early grades. However, a recent study compared the effectiveness of these two tools in discriminating texts for beginning readers (Hiebert and Pearson 2010) in grades K–2. Hiebert and Pearson (2010) tested Lexiles and Coh-Metrix variables against other common readability formulas for detecting differences among leveled texts. Lexiles performed better than other indices in distinguishing between the seven levels of texts used, but neither the Lexiles nor the Coh-Metrix variables were consistently sensitive enough. Thus, with care, a tool like the Lexile system can be used as an initial indicator of appropriate texts for young readers, but teacher judgment and perhaps more qualitative methods are necessary for making final determinations.

Conclusion

While controversy has surrounded the development and use of readability formulas for decades, researchers continue to develop methods to overcome past weaknesses. It seems quite possible that advances in natural language processing and other computerized language systems will reach a point when not only can text difficulty be accurately assessed for an individual but automated adaptation can also make informational texts more accessible to a particular reader. No doubt, when the time comes, controversy will surround such developments as well. In the meantime, educators and researchers alike can begin to see how the readability research field applies to the many facets of public and private life in a literate society. Traditional readability formulas have served their purpose in leveling typical books for school children, but more advanced and psychologically valid methods have been developed which are modeled after cognitive processing theories. These cognitively motivated approaches seem particularly appropriate for analyzing more complex informational texts for adolescents, college students, and adults in general. Finally, the most recent advances in machine-learning approaches point to a promising future in which non-traditional texts like those found on many web sites can be categorized for greater accessibility. With all these developments taking place in the past couple decades, the coming decade will hopefully present us with automated user-friendly readability tools for use in both research and practice.

References

Anderson, J. R. (1983). A spreading activation theory of memory. *Journal of Verbal Learning and Verbal Behavior, 22*, 261–295.

Anderson, R., & Pichert, J. (1978). Recall of previously unrecallable information following a shift in perspective. *Journal of Verbal Learning & Verbal Behavior, 17*(1), 1–12. doi:10.1016/S0022-5371(78)90485-1.

Bailin, A., & Grafstein, A. (2001). The linguistic assumptions underlying readability formulae: A critique. *Language & Communication, 21*(3), 285–301.

Blackburn, B. (2000). Best practices for using Lexiles. *Popular Measurement, 3*(1), 22–24.

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

Bormuth, J.R. (1969). *Development of readability analyses*. Final Report, Project No. 7-0052, Contract No. 1, OEC-3-7-070052-0326. Washington, DC: U.S. Office of Education.

Bormuth, J. R. (1971). *Development of standards of readability: Toward a rational criterion of passage performance.* Final report, U.S. Office of Education, Project No. 9-0237. Chicago: University of Chicago.

Britton, B., & Gülgöz, S. (1991). Using Kintsch's computational model to improve instructional text: Effects of repairing inference calls on recall and cognitive structures. *Journal of Educational Psychology, 83*(3), 329–345. doi:10.1037/0022-0663.83.3.329.

Britton, B., Gülgöz, S., Glynn, S. (1993). Impact of good and poor writing on learners: Research and theory. *Learning from textbooks: Theory and practice* (pp. 1–46). Hillsdale, NJ: Lawrence Erlbaum.

Carroll, J. B., Davies, P., & Richman, B. (Eds.). (1971). *Word frequency book*. New York: Houghton Mifflin.

Caylor, J.S., Sticht, T.G., Fox, L.C., Ford, J.P. 1973. *Methodologies for determining reading requirements of military occupational specialties*: *Technical report No. 73-5*. Alexandria, VA: Human Resources Research Organization.

Chall, J. S., & Dale, E. (1995). *Readability revisited: The new Dale–Chall readability formula*. Cambridge: Brookline Books.

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

Coleman, E. (1962). Improving comprehensibility by shortening sentences. *Journal of Applied Psychology, 46*(2), 131–134. doi:10.1037/h0039740.

Collins-Thompson, K., & Callan, J. (2004). A language modeling approach to predicting reading difficulty. In S. Dumais, D. Marcu, & S. Roukos (Eds.), *HLT-NAACL 2004: Main proceedings* (pp. 193–200). Morristown: Association for Computational Linguistics.

Collins-Thompson, K., & Callan, J. (2005). Predicting reading difficulty with statistical language models. *Journal of the American Society for Information Science & Technology, 56*(13), 1448–1462. doi:10.1002/asi.20243.

Crossley, S. A., Dufty, D. F., McCarthy, P. M., & McNamara, D. S. (2007). Toward a new readability: A mixed model approach. In D. S. McNamara & G. Trafton (Eds.), *Proceedings of the 29th Annual Conference of the Cognitive Science Society* (pp. 197–202). Austin: Cognitive Science Society.

Dale, E., & Chall, J. S. (1948). A formula for predicting readability. *Educational Research Bulletin, 27*(1), 11–20–28.

Dam, G., & Kaufmann, S. (2008). Computer assessment of interview data using latent semantic analysis. *Behavior Research Methods, 40*(1), 8–20. doi:10.3758/BRM.40.1.8.

Davison, A., & Kantor, R. (1982). On the failure of readability formulas to define readable texts: A case study from adaptations. *Reading Research Quarterly, 17*(2), 187–209.

DuBay, W.H. 2004. *The principles of readability.* Retrieved 30 August 2010 from http://www.impact-information.com/impactinfo/readability02.pdf.

Fellbaum, C. (Ed.). (1998). *WordNet: An electronic lexical database*. Cambridge: MIT.

Feng, L. (2009). Automatic readability assessment for people with intellectual disabilities. *ACM SIGACCESS Accessibility and Computing, 93*, 84–91. doi:10.1145/1531930.1531940.

Foltz, P., Kintsch, W., & Landauer, T. (1998). The measurement of textual coherence with latent semantic analysis. *Discourse Processes, 25*(2–3), 285–307. doi:10.1080/01638539809545029.

Fountas, I. C., & Pinnell, G. S. (1999). *Matching books to readers: Using leveled books in guided reading, K-3*. Portsmouth: Heinemann.

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

Fry, E. (1990). A readability formula for short passages. *Journal of Reading, 33*(8), 594–97.

Graesser, A. C., Gernsbacher, M. A., & Goldman, S. R. (1997). Cognition. In T. A. van Dijk & T. A. van Dijk (Eds.), *Discourse as structure and process: Discourse studies: A multidisciplinary introduction, vol. 1* (pp. 292–319). Thousand Oaks: Sage.

Graesser, A., McNamara, D., Louwerse, M., & Cai, Z. (2004). Coh-Metrix: Analysis of text on cohesion and language. *Behavior Research Methods, Instruments & Computers, 36*(2), 193–202.

Green, A., Ünaldi, A., & Weir, C. (2010). Empiricism versus connoisseurship: Establishing the appropriacy of texts in tests of academic reading. *Language Testing, 27*(2), 191–211. doi:10.1177/0265532209349471.

Heilman, M., Collins-Thompson, K., Callan, J., Eskenazi, M. (2007). Combining lexical and grammatical features to improve readability measures for first and second language texts. In *Proceedings of the NAACL Human Language Technology Conference* (pp. 460–467). Morristown, NJ: Association for Computational Linguistics.

Heilman, M., Collins-Thompson, K., Eskenazi, M. (2008). An analysis of statistical models and features for reading difficulty prediction. In *EANL '08 Proceedings of the Third Workshop on Innovative Use of NLP for Building Educational Applications*, June 19–19 (pp.71–79). Morristown, NJ: Association for Computational Linguistics.

Hiebert, E. (1999). Text matters in learning to read. *Reading Teacher, 52*(6), 552–66.

Hiebert, E. H., Pearson, P. D. (2010). *An examination of current text difficulty indices with early reading texts* (Reading Research Report No. 10-01). Santa Cruz, CA: TextProject, Inc.

Just, M., & Carpenter, P. (1980). A theory of reading: From eye fixations to comprehension. *Psychological Review, 87*(4), 329–354. doi:10.1037/0033-295X.87.4.329.

Kalyuga, S., Chandler, P., Sweller, J. (2000). Incorporating learner experience into the design of multimedia instruction. *Journal of Educational Psychology*, *92*, 126–136.2000-03003-01110.1037/0022-0663.92.1.126. doi:10.1037/0022-0663.92.1.126.

Kalyuga, S., Chandler, P., & Sweller, J. (2001). Learner experience and efficiency of instructional guidance. *Educational Psychology, 21*, 5–23. doi:10.1080/0144341012468110.1080/0144341012468120001-16707-001.10.1080/01443410124681.

Kalyuga, S., Chandler, P., Tuovinen, J., & Sweller, J. (2001). When problem solving is superior to studying worked examples. *Journal of Educational Psychology, 93*, 579–588. doi:10.1037/0022-0663.93.3.57910.1037/0022-0663.93.3.5792001-18059-013.10.1037/0022-0663.93.3.579.

Kalyuga, S., Ayres, P., Chandler, P., Sweller, J. (2003). The expertise reversal effect. Educational Psychologist, *38*, 23–31. doi:10.1207/S15326985EP3801. 10.1207/S15326985EP3801.

Kim, H., Goryachev, S., Rosemblat, G., Browne, A., Keselman, A., & Zeng-Treitler, Q. (2007). Beyond surface characteristics: A new health text-specific readability measurement. *American Medical Informatics (AMIA) Annual Symposium* (pp. 418–422). Washington, DC.

Kintsch, W. (1988). The role of knowledge in discourse comprehension: A construction–integration model. *Psychological Review, 95*(2), 163–182. doi:10.1037/0033-295X.95.2.163.

Kintsch, W., & Keenan, J. (1973). Reading rate and retention as a function of the number of propositions in the base structure of sentences. *Cognitive Psychology, 5*(3), 257–274. doi:10.1016/0010-0285(73)90036-4.

Kintsch, W., & van Dijk, T. (1978). Toward a model of text comprehension and production. *Psychological Review, 85*(5), 363–394. doi:10.1037/0033-295X.85.5.363.

Kirsch, I. S., & Mosenthal, T. B. (1990). Exploring document literacy: Variables underlying the performance of young adults. *Reading Research Quarterly, 25*, 5–30.

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

Klare, G. (1974). Assessing readability. *Reading Research Quarterly, 10*, 62–102.

Landauer, T., Foltz, P., & Laham, D. (1998). An introduction to latent semantic analysis. *Discourse Processes, 25*(2–3), 259–284. doi:10.1080/01638539809545028.

Leahy, W., Chandler, P., Sweller, J. (2003). When auditory presentations should and should not be a component of multimedia instruction. Applied Cognitive Psychology, *17*, 401–418.2003-00690-00410.1002/acp.877. doi:10.1002/acp.877.

Lin, S., Su, C., Lai, Y., Yang, L., & Hsieh, S. (2009). Assessing text readability using hierarchical lexical relations retrieved from WordNet. *Computational Linguistics and Chinese Language Processing, 14*(1), 45–84.

Liu, X., Croft, W.B., Oh, P., Hart, D. (2004). Automatic recognition of reading levels from user queries. *SIGIR '04 Proceedings of the 27th annual international ACM SIGIR conference on research and development in information retrieval* (pp. 548–549). New York, NY: ACM.

McClelland, J., & Rumelhart, D. (1981). An interactive activation model of context effects in letter perception: I. An account of basic findings. *Psychological Review, 88*(5), 375–407. doi:10.1037/0033-295X.88.5.375.

McNamara, D., & Kintsch, W. (1996). Learning from texts: Effects of prior knowledge and text coherence. *Discourse Processes, 22*(3), 247–288. doi:10.1080/01638539609544975.

McNamara, D., Kintsch, E., Songer, N., & 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. doi:10.1207/s1532690xci1401_1.

McNamara, D., Crossley, S., & McCarthy, P. (2010). Linguistic features of writing quality. *Written Communication, 27*(1), 57–86. doi:10.1177/0741088309351547.
McNamara, D., Louwerse, M., McCarthy, P., & Graesser, A. (2010). Coh-Metrix: Capturing linguistic features of cohesion. *Discourse Processes, 47*(4), 292–330. doi:10.1080/01638530902959943.
Meyer, B., Marsiske, M., & Willis, S. (1993). Text processing variables predict the readability of everyday documents read by older adults. *Reading Research Quarterly, 28*(3), 234–249. doi:10.2307/747996.
Miller, T. (2003). Essay assessment with latent semantic analysis. *Journal of Educational Computing Research, 29*(4), 495–512. doi:10.2190/W5AR-DYPW-40KX-FL99.
Miller, G. R., & Coleman, E. B. (1967). A set of thirty-six prose passages calibrated for complexity. *Journal of Verbal Learning and Verbal Behavior, 6*(6), 851–854.
Miller, J., & Kintsch, W. (1980). Readability and recall of short prose passages: A theoretical analysis. *Journal of Experimental Psychology: Human Learning and Memory, 6*(4), 335–354. doi:10.1037/0278-7393.6.4.335.
Millis, K., Magliano, J., Wiemer-Hastings, K., Todaro, S., McNamara, D. (2007). Assessing and improving comprehension with latent semantic analysis. *Handbook of latent semantic analysis* (pp. 207–225). Mahwah, NJ: Lawrence Erlbaum.
Milone, M. (2009). *The development of ATOS: The renaissance readability formula*. Wisconsin Rapids: Renaissance Learning.
Miltsakaki, E., Troutt, A. (2007). Read-X: Automatic evaluation of reading difficulty of web text. *Proceedings of E-Learn 2007, sponsored by the Association for the Advancement of Computing in Education*. Quebec, Canada.
Miltsakaki, E., & Troutt, A. (2008). Real-time web text classification and analysis of reading difficulty. In J. Tetreault, J. Burstein, & R. De Felice (Eds.), *EANL '08 Proceedings of the Third Workshop on Innovative Use of NLP for Building Educational Applications* (pp. 89–97). Morristown: Association for Computational Linguistics.
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: Heinemann.
Peterson, S., & Ostendorf, M. (2009). A machine learning approach to reading level assessment. *Computer Speech and Language, 23*(1), 89–106.
Rog, L., & Burton, W. (2002). Matching texts and readers: Leveling early reading materials for assessment and instruction. *Reading Teacher, 55*(4), 348–56.
Rosch, E., Mervis, C. B., Gray, W., Johnson, D., & Boyes-Braem, P. (1976). Basic objects in natural categories. *Cognitive Psychology, 8*(3), 382–439.
Rumelhart, D., & McClelland, J. (1982). An interactive activation model of context effects in letter perception: II. The contextual enhancement effect and some tests and extensions of the model. *Psychological Review, 89*(1), 60–94. doi:10.1037/0033-295X.89.1.60.
School Renaissance Inst., Inc. (2000). The ATOS[TM] readability formula for books and how it compares to other formulas. Madison, WI: School Renaissance Inst., Inc. (ERIC Document Reproduction Service No. ED449468).
Schriver, K. A. (2000). Readability formulas in the new millennium: What's the use? *ACM Journal of Computer Documentation, 24*(3), 105–106.
Schwarm, S.E., Ostendorf, M. 2005. Reading level assessment using support vector machines and statistical language models. *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics*, pp. 523–530, Ann Arbor, MI.
Si, L., Callan, J. 2001. A statistical model for scientific readability. *CIKM'01: Proceedings of the Tenth International Conference on Information and Knowledge Management*, pp. 574–576.
Smith, R. (2000a). How the Lexile framework operates. *Popular Measurement, 3*(1), 18–19.
Smith, R. (2000b). The Lexile community: From science to practice. *Popular Measurement, 3*(1), 20–21.
Smith, D., Stenner, A.J., Horabin, I., Smith, M. (1989). The Lexile scale in theory and practice: Final report. Washington, DC: MetaMetrics (ERIC Document Reproduction Service No. ED307577).
Stenner, A.J. (1996). Measuring reading comprehension with the Lexile framework. Paper presented at the 4th North American Conference on Adolescent/Adult Literacy, Washington, DC.
Stenner, A.J. (1999). *Instructional uses of the Lexile framework*. Durham, NC: MetaMetrics, Inc. (ERIC Document Reproduction Service No. ED435976).
Stenner, A., Burdick, D. (1997). *The objective measurement of reading comprehension: In response to technical questions raised by the California Department of Education Technical Study Group*. Durham, NC: MetaMetrics, Inc. (ERIC Document Reproduction Service No. ED435978).
Stenner, A.J., Burdick, H., Sanford, E.E., Burdick, D.S. (2007). The Lexile framework for reading technical report. MetaMetrics, Inc.

vor der Brück, T., Hartrumpf, S., & Helbig, H. (2008). A readability checker with supervised learning using deep indicators. *Informatica, 32*(4), 429–435.

Wolfe, M., Schreiner, M., Rehder, B., Laham, D., Foltz, P., Kintsch, W., et al. (1998). Learning from text: Matching readers and texts by latent semantic analysis. *Discourse Processes, 25*(2–3), 309–336. doi:10.1080/01638539809545030.

Wright, B., Stenner, A. (1998). *Readability and reading ability.* Paper presented to the Australian Council on Education Research (ACER) (ERIC Document Reproduction Service No. ED435979).

Wright, B., & Stenner, A. (2000). Lexile perspectives. *Popular Measurement, 3*(1), 16–17.