2019

Readability Formulas and Writing Technique

Eric Andrew Bryant
Bethel University

READABILITY FORMULAS
AND WRITING TECHNIQUE

A MASTER'S THESIS PROJECT

SUBMITTED TO THE FACULTY

OF BETHEL UNIVERSITY

BY

ERIC A. BRYANT

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS

FOR THE DEGREE OF

MASTER OF ARTS

DECEMBER 2019

Abstract

Readability formulas are widely used to analyze educational texts; however, the use of these formulas has been widely criticized. Prominent among these criticisms are that classic readability formulas calculate readability using only sentence length and word length, and that writers should not directly consider readability formula factors such as word and sentence length when creating texts. This literature review attempts to identify specific writing and adaptation techniques that educators and educational writers can use to improve the comprehensibility of texts aimed at secondary-level readers. Advantages and limitations of readability formulas are discussed. Word length, sentence length, and reader characteristics are analyzed individually as factors affecting comprehensibility. Studies on text elaboration and simplification for L2 readers are discussed to gain further insight into word length and sentence length as variables affecting reader comprehension. An alternate tool for assessing text, Coh-Metrix, is discussed as a potential replacement for classic readability formulas, and specific Coh-Metrix measurements are evaluated.

Keywords: readability formulas, quantitative analysis (text), reader and task, text elaboration and simplification, Coh-Metrix

Table of Contents

Chapter I: Introduction and Purpose 4
Guiding Question 6
Definitions 7
Chapter II: Literature Review 10
Readability Formulas 10
Conventional and Modern Readability Analysis 11
Debate Concerning Readability Formulas 12
Why Might Readability Formulas Misjudge Text? 22
Can Readability Formulas Actively Mislead? 37
Readability Analysis Through Coh-Metrix 48
Coh-Metrix Analysis of Lexis 50
Familiarity and word frequency 50
Concreteness, imageability, and hypernymy 51
Meaningfulness 53
Age of acquisition 53
Polysemy 54
Type:token ratio 54
Coh-Metrix analysis of syntax 54
Coh-Metrix analysis of cohesion 57
Utilization of Coh-Metrix 63
Coh-Metrix and L2 readers 63
Coh-Metrix grading of simplified texts by difficulty 64
Coh-Metrix analysis of simplified texts' linguistic features 66
Coh-Metrix analysis of subject-area texts 69
Chapter III: Conclusion 73
Summary 73
Professional Application 76
Limitations of the Research 82
Implications for Future Research 85
Conclusion 86
References 88

Chapter I: Introduction and Purpose

Because the acquisition of information in educational contexts is highly dependent on reading, evaluating the readability of texts is of vital concern (Blau, 1982). The goal of having students engage with appropriate texts throughout their school lives requires valid and reliable measures of text complexity (McNamara, Graesser, McCarthy, & Cai, 2014; Nelson, Perfetti, Liben, & Liben, 2011). The question of how to evaluate texts has taken on particular importance in the context of ongoing changes in the curricular and assessment systems of United States K-12 education including the requirements of No Child Left Behind and the continuing implementation of the Common Core State Standards (Hiebert, 2011; McNamara, Graesser, Cai, & Kulikowich, 2011; Nelson et al., 2011). However, this concern is not new; Hiebert (2011) stated that readability formulas have been used in the American education system for nearly a century. Text selection concerns affect essentially everyone involved in the K-12 educational field; students, teachers, principals, superintendents, publishers, and assessment developers are all stakeholders in this ongoing process (Graesser, McNamara, & Kulikowich, 2011; Nelson et al., 2011).

A large proportion of students in United States schools fail to achieve reading proficiency during their school years. According to the 2017 National Assessment of Educational Progress (NAEP) Reading Report Card, 24% of 8th grade students scored below basic proficiency in reading while 36% achieved proficiency and 4% achieved advanced proficiency (The Nation's Report Card, 2017). The 2015 NAEP Reading Report Card showed similar results for 12th grade students; 28% scored below basic proficiency while 31% achieved proficiency and 6% achieved advanced proficiency (The Nation's Report Card, 2015). These results show that secondary students' ability to comprehend text is a source of concern.

The ever-changing nature of the population served by the U.S. educational system provides further impetus to improved analyses of text readability. The Current Population Survey compiled by the United States Census Bureau (2017) reported that as of October 2016 there were 42,578,000 foreign-born persons residing in the United States, of which 5,097,000 were enrolled in a United States educational institution, including 1,207,000 at the elementary level and 1,179,000 at the high school level (p. 1). Ovando and Combs (2018) found that educational discourse and learning environments continue to reflect the discourse practices of mainstream society, resulting in negative consequences for language minority students.

Frequently, teachers must design, produce, and adapt texts for their students, however relatively few teachers have training in materials development (Lesiak-Bielawska, 2015; Tomlinson, 2012). Lenz, Schumaker, and ERIC Clearinghouse on Disabilities and Gifted Education (2003) stated that "when instructional materials present a barrier to student learning, teachers often adapt the materials to allow students greater access to the information to be taught" (p. 1-2).

Adapting and creating texts can help unlock teacher success and student achievement. Kimmons (2015) noted that many textbooks used in high schools are written by college professors and argued that such textbooks may disadvantage high-school users in diverse ways. Texts may not be written at an appropriate reading level, may not align with state standards or scaffolded learning, may not include appropriate or high-quality supplementary resources (e.g., classroom activities, practices, rubrics), and may be inappropriate for diverse classrooms. Robinson, Fischer, Wiley, and Hilton (2014) found that secondary student standardized test scores increased when open textbooks were used in place of copyright-restricted textbooks. Kimmons (2015) found that teachers evaluated adaptations of open-source textbooks as 38%

higher in overall quality than traditional copyright-restricted textbooks based on 10 diverse metrics (e.g., conciseness, readability, standards alignment, supplements, timeliness, links to external resources).

Street and Stang (2008) argued that when teachers become more experienced writers, they are better able to teach students how to write effectively. Street and Stang furthermore noted that writing is a complex task and that there is no simple formula that can be followed. Klare (1976) contended that published guides to effective writing do not tend to accord with one another, and in fact commonly contradict one another. The challenge of making sense of contradictory writing advice is likely only increasing in an era where educators receive more and more information and advice through social media, blogs, and ebooks. Thus, there is a need for more research on writing techniques and linguistic features that have been experimentally demonstrated to increase or decrease readability and comprehensibility. Writers and adaptors of text deserve effective and evidence-based guidance, just as students deserve writers and adaptors who can write comprehensible, informative, and engaging texts.

Guiding Question

In light of this purpose, the guiding question of this review is: *Can research on readability formulas offer insight into effective writing and adaptation techniques for secondary educators who write or adapt text for the purpose of improved student comprehension?*

The textual analysis and text modification techniques reviewed in this paper may be useful for a broad audience of secondary-level educators. Some teachers may wish to adapt text for students who are in a lower grade level than the intended audience of a particular text, for example a 6th grade biology teacher who wants to adapt a passage from an online article intended for high school students or adults. Other teachers may adapt texts from their own

classrooms, for example to make the textbook more readable for students with limited proficiency or whose first language is not English, or simply to scaffold complex texts generally. Other teachers may create their own texts (e.g., articles, PowerPoint presentations, exam questions). A closer understanding of readability formulas and what they can or cannot measure may also be valuable for educational professionals who are responsible for designing curriculum or selecting texts.

Definitions

For the purposes of this paper, the following definitions will be used.

***Authentic Text*:** Crossley and McNamara defined authentic texts as "unmodified texts that were originally created to fulfill a social purpose in a first language community" (Crossley & McNamara, 2016, p. 2).

Coh-Metrix: According to Crossley, Louwerse, McCarthy, and McNamara (2007b), Coh-Metrix is a "computational tool that measures cohesion and text difficulty at various levels of language, discourse, and conceptual analysis" (p. 19) and "measures over 250 language and cohesion features" (p. 16). Coh-Metrix uses computational linguistics, corpus linguistics, information extraction, information retrieval, and discourse processing to analyze texts (Graesser, McNamara, Louwerse, & Cai, 2004, p. 193).

***Cohesion*:** Cohesion "refers to relations of meaning within a text, and that define it as a text" (Halliday & Hasan, 1976, p. 4). Cohesion is broader than syntax as syntax refers only to sentence-level constructions (Halliday & Hasan, 1976).

Frequency/Familiarity: Word frequency refers to measurements of "how frequently particular words occur in the English language" (Graesser et al., 2004, p. 197). Word frequency and word familiarity can be considered rough synonyms as they both refer to statistical measurements of

how often a reader is likely to have encountered a word (Coh-Metrix version 3.0 indices, n.d.; Crossley, Allen, & McNamara, 2012)

L1: L1 is the "first or native language" (Mishan, 2005, p. xii).

L2: L2 is the "second language . . . or foreign language being learned" (Mishan, 2005, p. xii).

***Lexis/Lexical*:** Lexis is the vocabulary of a language and the study of vocabulary (Matthews, 2007).

Readability: Readability is "an attribute of written text, commonly defined by factors that theoretically make text more or less difficult to read (e.g., vocabulary, sentence complexity)" (Begeny & Greene, 2013).

Readability Formula: Readability formulas generate numerical estimates of how easy or difficult it is for a reader to comprehend a text (Lenzner, 2014). Readability formulas are a quantitative method of text evaluation which can be distinguished from qualitative evaluations (e.g., human ratings of conceptual difficulty or idea density) (Oakland & Lane, 2004).

***Structural and Intuitive Approaches*:** Structural approaches to simplification rely on techniques such as word lists which are predefined by level as well as readability formulas such as Flesch-Kincaid Grade Level, whereas the intuitive approach is "more subjective and depends solely on the author's natural sense of text comprehensibility and discourse processing" (Crossley, Allen, & McNamara, 2011, p. 84).

***Syntax/Syntactic*:** Syntax is the study of grammatical relationships between words and other units within the sentence (Matthews, 2007).

***Text Modification/Simplification/Elaboration*:** Oh (2001) stated, "modifications to input can be divided into two types: simplification, in the form of less complex vocabulary and syntax, and

elaboration, in which unfamiliar linguistic items are offset with redundancy and explicitness" (p. 70).

Chapter II: Literature Review

Although many readers will be familiar with readability formulas' function, fields of use, and general method of predicting text difficulty, a short introduction of these concepts will make subsequent discussions more transparent.

Readability Formulas

Readability formulas are a quantitative method for assessing textual difficulty and predicting the ease or difficulty readers are likely to experience in comprehending a text (Lenzner, 2014; Oakland & Lane, 2004). Most readability formulas rely on factors representing two broad features of text difficulty, namely lexical sophistication and syntactic complexity, with lexical sophistication typically measured by word length and syntactic complexity typically measured by sentence length (Armbruster, Osborn, & Davison, 1985; Crossley, Skalicky, Dascalu, McNamara, & Kyle, 2017b; Graesser et al., 2004; Lenzner, 2014; McNamara et al., 2014). Readability formulas rely on mathematical methods of predicting comprehensibility and can thus be considered a structural approach rather than an intuitive approach, which is "more subjective and depends solely on the author's natural sense of text comprehensibility and discourse processing" (Crossley, Allen & McNamara, 2011, p. 84-85). The most common formulas are Flesch Reading Ease and Flesch-Kincaid Grade Level, both of which use two independent variables, namely average sentence length and average syllables per word, to calculate a readability score (Graesser et al. 2004).

Readability formulas can be used to analyze any type of text. Crossley et al. (2017b) reported that the relative simplicity and mechanical nature of readability formulas has led to their widespread adoption by teachers, testing agencies, and the print media. Readability formulas exert a powerful influence on the textbook industry and the selection of texts (Armbruster et al.,

1985; Graesser et al., 2004). To provide evidence for the widespread acceptance and use of these formulas, numerous researchers have pointed to existing legal statutes, including laws governing the field of education. Both state and federal laws exist stating that diverse texts including educational textbooks, insurance policies, warranties, legal instruments, tax forms, contracts, and jury instructions must meet certain criteria in terms of readability formulas (Bruce, Rubin, & Starr, 1981; Charrow & Charrow, 1979; Davison et al., 1980; Davison & Kantor, 1982). Tinkler and Woods (2013), for example, noted that the U.S. Department of Defense requires that all documents have a Flesch-Kincaid Grade Level score at or below the 10th grade.

According to Nelson et al. (2011), text difficulty tools such as readability formulas have a wide range of educational applications including uses consistent with the Common Core Standards (e.g., text evaluation, assessment). The potential educational utility of readability formulas in evaluating texts for prospective readers, as well as the ubiquity of these formulas, invites discussion of these formulas to determine (1) where and when the use of these formulas might be appropriate, (2) what the shortcomings of these formulas can tell us about reader comprehension and better writing technique, and (3) whether more modern forms of computer-assisted text analysis can improve our ability to achieve readability formulas' ultimate goal of predicting reader comprehension of diverse texts.

Conventional and Modern Readability Analysis

Several readability formulas which are still in common use today, namely Flesch Reading Ease, Gunning Fog, and Dale-Chall, were developed during the 1940s and 1950s (Lenzner, 2014). However, development of readability formulas is ongoing. More than 200 readability formulas have been produced since the 1970s with the goal of providing tools for measuring text difficulty more accurately and efficiently (Crossley et al., 2017b).

The advent of computer technology and recent progress in the fields of cognitive science, psycholinguistics, computational linguistics, corpus linguistics, and information retrieval has led to attempts to analyze and predict readability using linguistic measures that go far beyond word and sentence length (Crossley et al., 2011; Graesser et al., 2004; McNamara et al., 2011). Coh-Metrix, for example, is a publicly available web-based software tool which analyzes texts on over 200 measures of cohesion, language, and readability (Graesser et al., 2004). McNamara et al. (2011) stated that "Coh-Metrix is motivated by theories of discourse and text comprehension" (p. 5) and that it "was designed to move beyond standard readability formulas, such as Flesch-Kincaid Grade Level" (p. 2).

The availability of newer linguistic-processing software tools has led to changes in how older tools such as Flesch Reading Ease are described in the literature. Crossley, Dufty, McCarthy, & McNamara (2007a) used the descriptors "traditional," "classic," and "conventional" to describe readability formulas which rely solely on word and sentence length (p. 197). McNamara et al. (2011) referred to these formulas as "uni-dimensional" (p. 5). To avoid confusion, the term "readability formula(s)" as used in this review will refer solely to conventional readability formulas which are still in widespread use, though in some cases the term "conventional readability formula(s)" will be used to provide greater contrast to other measurements and tools in context. Newer tools such as Coh-Metrix will be described by name.

Debate Concerning Readability Formulas

Readability formulas have been used in education and other fields since at least 1923 (Hiebert, 2011). However, there is still lively debate concerning their pros and cons, the contexts within which they should or should not be used, and even whether they are of any real use whatsoever.

Readability formulas would likely never have become as ubiquitous as they are without significant evidence behind them and a wide range of potential use cases. Armbruster et al. (1985) reported that readability formulas are "objective, quantitative, and relatively easy to use" (p. 18). These advantages, particularly objectivity and ease-of-use, constitute part of the reason for the widespread adoption of such formulas (Crossley et al., 2017b; Davison et al., 1980). There is also strong statistical evidence for the validity of readability formulas in predicting whether target readers will be able comprehend a text. A number of classic validation studies have found readability formulas' predictive validity to be consistently high, with formula scores correlating to observed difficulty in the range of .7 to .8 (Crossley et al., 2017b). Davison et al. (1980) reported that readability formulas have shown high statistical correlations in the .8 and .9 range to other measurements of readability such as student success in answering comprehension questions, cloze testing, and publishers' assigned grade level, though the authors noted that assigned grade level was probably influenced by readability formulas in the first place and thus vulnerable to circular logic. Davison et al. also emphasized other strong attractions of using readability formulas including ease-of-use by persons without specialized training and the ability to apply the formulas to diverse text types (e.g., narratives, expository prose, technical prose).

Klare (1976) performed a review of 36 studies which had looked at the predictive validity of readability formulas. Klare defined validity in terms of whether the modification of a text into a more or less readable version of the text according to the readability formula score had led to a measurable difference in reader performance or behavior, for example improved scores on a comprehension test. Nineteen studies showed statistically significant changes in reader performance as predicted by the readability formulas, six studies showed mixed results where

significance could only be established in highly specific circumstances, and 11 studies showed no statistical significance.

McNamara et al. (2011) reported that "sentence length and word length . . . robustly predict reading time" (p. 2). Graesser et al. (2004) claimed that texts are generally more difficult to read when they contain long words and lengthy sentences. Longer words tend to occur with less frequency and words with lower global frequency take more time for the reader to access and interpret. Longer sentences, meanwhile, place higher demands on working memory and thus tend to result in a more challenging cognitive load. From this perspective, formulas that rely on word and sentence length exhibit some degree of statistical validity.

Criticisms of readability formulas are widespread in the academic literature. Crossley and McNamara (2016) stated that conventional readability formulas "are widely criticized as weak indicators of comprehensibility" (p. 3). Graesser et al. (2004) reported that readability formulas "rely exclusively on word length and sentence length, two very simple and shallow metrics" (p. 194) and argued that "two-parameter multiple regression equations will not go the distance in explaining text difficulty" (p.194). Armbruster et al. (1985) stated that readability formulas "contribute to the production of poorly written text" (p. 20).

McNamara et al. (2011) noted that readability measures are limited in that they consider only the features of text that tend to predict surface understanding of words and separate sentences. The authors also pointed out that even if a readability formula can predict student comprehension, it cannot identify the particular characteristics of the text that may be challenging or helpful to the student. Dowell, Graesser, and Cai (2016) similarly claimed that conventional readability formulas are not useful for identifying specific deficits in text and that

this, in turn, makes it difficult to provide support to students (e.g., provide scaffolding for specific text features that may cause comprehension problems).

The adoption of readability formulas has been driven in part by studies that show high validity for such formulas (Crossley et al., 2017b; Davison et al., 1980). However, the methods used to test the formulas' validity have been questioned (Crossley et al., 2007a; Klare, 1976). Others have expressed concern about confusing correlation for causation (Lenzner, 2014). Bruce et al. (1981) argued that there is little empirical or statistical evidence that readability formulas are accurate when predicting comprehensibility. Brucet et al. noted that readability formulas often provide their results in a grade-level format, indicating that a child of that level of reading ability should be able to comprehend the text. However, validation of such claims on actual students of that grade level was rarely if ever performed. Graesser et al. (2014) similarly noted that "there is no solid gold standard for defining grade level" (p. 211).

Klare (1976) reviewed 36 studies on readability formula validity and noted a strong publication bias which implicitly favored positive evaluations of readability formula validity. Six of the 19 studies which had found statistical significance between readability formula score and reader comprehension had been published in journals, whereas none of the 11 studies which failed to find such statistical significance had been published in journals. Klare (1976) attributed this finding to the statistical fact that you cannot prove the null hypothesis: in this case, the hypothesis that readability formulas do not predict changes in reader performance or behavior.

Some empirical research has suggested a negative correlation between readability formula scores and comprehensibility. Lockman (1957) found that naval cadets' rating of the understandability of a text correlated negatively with that text's Flesch Reading Ease score and that this result was statistically significant. Lockman (1957) concluded that Flesch Reading Ease

scores and understandability ratings as provided by readers did not measure the same thing. Charrow & Charrow (1979) found a negative correlation between Flesch readability scores and tested comprehension of jury instructions, though the effect was not found to be statistically significant. Lenzner (2014) found readability formulas tend to inaccurately distinguish between relatively confusing and relatively comprehensible survey questions. These three studies will be discussed in depth later in this review.

Concerning the reasons why readability formulas may fail to predict reader comprehension, Armbruster et al. (1985) advanced two major criticisms. First, these formulas do not consider characteristics of text that are known to affect comprehension (e.g., content difficulty, content familiarity, author style, organization). Second, these formulas do not consider the reader of the text; they neglect important factors such as motivation, interest, and purpose. Armbruster's two major lines of criticism can be found in the work of numerous researchers. Graesser et al. (2004) admitted that texts will tend to be more difficult to read when they contain longer words and lengthier sentences, but stated that readability formulas are shallow in that they "ignore dozens of language and discourse components that are theoretically expected to influence comprehension difficulty" (p. 194). Young (1999) stated that readability formulas do not consider the structure of the text and ignore students' background knowledge, language proficiency, and reading strategies. Kantor and Davison (1981) argued that text comprehensibility is located beyond sentence and word length and is primarily determined by global factors such as idea presentation, reader background knowledge, and local discourse organization such as transitions between ideas.

Addressing the question of text characteristics, Bruce et al. (1981) stated that readability formulas which rely entirely on sentence length and word difficulty ignore other vital factors

determining text comprehensibility, including "degree of discourse cohesion, number of inferences required, number of items to remember, complexity of ideas, rhetorical structure, [and] dialect" (p. 4). Bruce et al. (1981) also addressed the question of target reader, stating that the formulas are based on the isolated text and take no account of the context in which the text will be read. Reader-specific factors such as background knowledge, cultural background, motivation, interest, values, and purpose are ignored. To illustrate this point, Bruce et al. used the example of a text which contains relatively simple sentences and words but tells a story which is complex and subtle. The authors contended that in such situations, readability formulas will tend to provide wildly inaccurate numbers which greatly overestimate young readers' ability to achieve meaningful comprehension. As an illustration, Snow (2015) calculated that *The Old Man and the Sea* by Ernest Hemingway had a Flesch-Kincaid Grade Level value of four.

Bruce et al. (1981) stated readability formulas could be useful if strict criteria for the use case are met: (1) the material may be freely read, (2) the text is written to satisfy communicative goals rather than the formula itself, (3) higher-level text structures such as text organization are unimportant, (4) reader purpose is unimportant, (5) statistical averages are closely correlated to individual readers, and (6) the target reader's characteristics are similar to the characteristics of the readers upon whom the specific readability formula was validated. The authors argued that such cases were rare and nearly all important potential use cases of readability formulas violate these criteria, including adaptations of texts, selection of texts for readers of different cultural backgrounds, designing special texts for children, selection of text passages, and the design of remedial readers. The authors placed special attention on two educational use cases where readability formulas would be extremely valuable if their measurements were reliable, namely selecting an appropriate text for a child in school and as a guideline for the simplification of

existing texts. However, in these vital situations, the authors argued that readability formulas were specifically inappropriate and generally no better than intuitive methods of predicting comprehensibility, which may in fact be underutilized by educators who place too much trust in readability formulas. Davison et al. (1980) placed similar emphasis on the importance of intuitive methods of evaluating texts and on factors affecting comprehensibility "for which there is to date no objective measurement" (p. 5).

Armbruster et al. (1985) pointed out that different passages from the same book or passage might show widely different readability scores even within the same formula. The authors randomly selected four 100-word passages from a 5th grade social studies textbook and found Fry Graph readability scored the four passages as appropriate for the 4th, 7th, 8th, and 11th grades. Despite this variability, the authors reported that textbook companies do not usually provide information on sampling procedures to consumers.

Armbruster et al. also noted that different readability formulas often provide widely different scores; a randomly selected passage from a 6th grade science textbook was given grade level scores of 3.1 by the Spache formula, 4.2 by the Dale-Chall formula, 4 by the Gunning formula, and 7 according to the Fry Graph formula. Thus, the reported readability of a textbook will depend on which formula is chosen. Lenzner (2014) echoed this criticism, stating that simply changing from one readability formula to another will often result in different scores and thus different text selection. In the end, Armbruster et al. (1985) emphasized the role of intuition, stating that decisions about matching texts with readers "are probably best made by trained and experienced judges—the teachers and librarians who have worked with children and who have witnessed the interactions of a lot of children with a lot of books" (p. 20).

Conventional readability formulas such as Flesch Reading Ease and Flesch-Kincaid Grade Level rely mathematically on word and sentence length. As a result, authors or adaptors can change the predicted formula scores of a text by systematically choosing shorter words and shorter sentences. Such a process results in formula scores indicating increased predicted comprehension even if the actual comprehensibility of the text were unaffected or had in fact suffered as a result of the modification (Armbruster et al., 1985; Charrow & Charrow, 1979; Crossley et al., 2017a; Davison & Kantor, 1982; Klare, 1976).

Bruce et al. (1981) stated that many educational authors are under pressure to attain specific scores according to readability formulas, and that readability formulas have been used as "guidelines for the simplification of existing texts and documents" (p. 7). Armbruster et al. (1985) pointed out that publishers are often under pressure to produce texts with a given readability level but stated that this often results in texts that are less readable. Davison et al. (1980) contended that

> there is inescapable temptation to use these formulas as a guide to writing, especially if the writer is under an obligation to produce materials at a specific readability level. Publishers of textbooks, for children as well as college students, have recently been eager to guarantee the reading levels of their product, and the issue of guaranteeing readability levels is coming to have wider application with the institution of the 'Plain English' requirement for many legal and other documents and with new interest in accurate c aptioning for hearing impaired people. (p. 2)

Graesser et al. (2004) stated that readability formulas are commonly misused, as "textbook writers are known to shorten sentences . . . for the purpose of downsizing the grade levels of their texts" (p. 194). According to Graesser et al., this process often results in texts with

lower cohesion and coherence. Cohesion and the use of cohesive devices is a linguistic factor in text which plays an important role in reader comprehension, for example by assisting readers in generating inferences and bridging conceptual gaps (Crossley, Rose, Danekes, Rose, & McNamara, 2017a; Duran, McCarthy, Graesser, & McNamara, 2007). Plakans and Bilki (2016) noted that cohesion is critical for readers to make both local and global connections across ideas, clauses, and words in a text. Rote use of readability formulas when evaluating text inherently disregards the vital issue of cohesion.

Bruce et al. (1981) argued that although formulas may assign reasonable numerical values to existing text, they do not justify modifications of text. In fact, where writers write to the formula, "such prescriptive use magnifies the inaccuracies inherent in the formulas" (Bruce et al., 1981, p. 7). The authors further argued that despite the limitations of readability formulas, writers engaged in simplification work often produced text which considered the formula above all else, with the reader and his or her text comprehension playing a secondary role. Armbruster et al. (1985) echoed this criticism, stating that "evidence is fast accumulating that these formulas may not be very useful in selecting textbooks and that, in fact, they may adversely affect the quality of textbook writing" (p. 18). Perhaps reflecting some of these concerns, California in 1987 and Texas in 1990 changed their language arts textbook adoption guidelines to stipulate that texts "should not be manipulated to comply with readability formulas" (Hiebert & Pearson, 2010, p. 1).

Davison et al. (1980) pointed out that readability formulas cannot be truly objective due to their reliance on subjective factors, namely "the skill or common sense of the writer who is presumed to have created a coherent, well-formed text to which objective measurement may be applied" (p. 4). Since the formula depends on the good faith of the writer in attempting to create

highly comprehensible text, if the writer explicitly attempts to satisfy the formula while writing, a circular feedback loop is created, often with negative results. Davison et al. (1980) concluded that readability formulas cannot instruct a writer on how to produce a text; the phenomenon of writers and publishers creating texts to formula essentially contradicts the foundations upon which the objectivity of readability formulas is supposed to rely.

According to Lockman (1957), Flesch himself pointed out that readability formulas will not indicate whether the ideas expressed are nonsense. To this, Lockman added that the formulas are similarly unable to indicate whether a text is ungrammatical and stated that we should perhaps make a distinction between *readability* and *understandability* (Lockman, 1957, p. 195).

Davison et al. (1980) performed a sentence-by-sentence analysis of four modified texts from SRA Reading Laboratory 3b which had been adapted from adult-level texts. These modified texts were designed for students in 8th, 9th, and 10th grade who are reading at 5th and 6th grade levels. The authors found clear evidence of adaptors engaging in conscientious and careful rewriting. For example, in some cases adaptors deliberately increased sentence length, apparently because they felt this would aid student comprehension. However, they also found evidence of passages where writers had clearly placed primary importance on vocabulary lists and restrictions on sentence length and passage length rather than other factors which affect readability. In one part of this study, the authors instructed amateur editors to modify the same original texts that had been rewritten by the professionals and published in the actual textbook. The amateur editors were told to use whatever means they wished in their modification process. These editors often shortened the sentence length of strikingly long or complicated sentences, for example using splitting or paraphrasing. However, it was found that in the professional adaptations, *all* sentences above a certain length had been shortened. The authors concluded that

the readability formula itself had likely exerted strong influence upon the professional editors' modification decisions.

Davison and Kantor (1982) stated that "adaptations were found to be most successful when the adaptor functioned as a conscientious writer rather than someone trying to make a text fit a level of readability defined by a formula" and urged for more experimental research to define the "real factors constituting readability" (p. 187). This review is an attempt to identify such real factors within empirical research on readability formulas such that educators and authors will have greater guidance on grammatical, lexical, cohesive, and reader-specific factors to consider when producing text for the purpose of improved secondary student comprehension.

Why Might Readability Formulas Misjudge Text?

Crossley, Greenfield & McNamara (2008) stated that conventional readability formulas have been widely criticized by both L1 and L2 researchers due to their inability to account for deeper levels of text processing. By outlining several characteristics of text and text readers which affect deeper processing but are not measured by readability formulas, insight can be gained into language features that a writer may wish to use or avoid regardless of the effect on readability formula output.

Elfenbein (2011) stated that a major challenge in the study of text is the difficulty of controlling for multiple variables. Syntax, vocabulary, cohesion, and other aspects of text known to affect comprehension tend to be inextricably linked to one another, such that modifying text to achieve one target may result in negative tradeoffs concerning the achievement of other targets. One such linguistic pattern is the tradeoff between using shorter sentences and using conjunctions which function to combine two sentences into a single multi-clause sentence.

Lesser and Wagler (2016) argued that adding the conjunction "because" to texts explaining statistics concepts tends to improve cohesive aspects of the text. However, adding such a conjunction will also tend to increase sentence length and thus syntactic complexity as measured by readability formulas due to the required use of multiple clauses and thus longer sentences. The authors contrasted these two texts:

> LOWER COHESION: The mean is greater than the median. There are a few observations much larger than the others.
>
> HIGHER COHESION: The mean is greater than the median because of a few large observations (Lesser & Wagler, 2016, p. 155).

Applying linguistic analysis to these two sentences, the authors noted that the "lower cohesion" text had a Flesch-Kincaid score of 4.4, whereas the "higher cohesion" text had a Flesch-Kincaid score of 6.7, predicting reduced comprehensibility. However, the authors stated that "most readers would find the higher cohesive text more comprehensible due to the logical structure of the text even though [Flesch-Kincaid] alone suggests it is less comprehensible" (Lesser & Wagler, 2016, p. 156). Lesser and Wagler suggested that if one target, such as sentence length, is measured by a readability formula, whereas another target, such as cohesion, is not measured by the same formula, readability formulas may assign higher readability scores to less comprehensible text, or vice versa. The authors concluded that "using simple measures of readability, such as [Flesch-Kincaid], is not sufficient" (Lesser & Wagler, 2016, p. 156).

Armbruster et al. (1985) similarly provided specific examples where shorter sentences, especially those denuded of conjunctions between ideas in order to create shorter sentences, caused text to become less comprehensible. Simply creating new sentences while deleting conjunctions such as "and," "but," "then," "because," and "since" creates ambiguous and

confusing texts which young readers have trouble parsing. The authors provided the following somewhat humorous example: “A cell is made of living stuff. A cell can grow. It takes in food. It changes the food into more living stuff” (Armbruster et al., 1985, p. 20).

Tweissi (1998) found that the presence of appropriate conjunctions enhances comprehension “whether or not the information is also recoverable from context” (p. 206). Crossley et al. (2007b) also discussed the method of omitting conjunctions and connectives between sentences and argued that this method tends to negatively affect text cohesion and interfere with reader processing. Even if the individual sentences become easier to understand, the meaning and purpose of those sentences in relation to other sentences and to the text as a whole may become more difficult for readers to decipher.

Davison et al. (1980) found that one common method adaptors use to simplify or modify text was to reduce sentence length. This can be done by splitting a sentence into two or more independent sentences. However, sentence splitting is not always useful. The authors noted that even simple conjunctions (e.g., “and”) can signal complex relationships such as causation, sequence in time, and contrast. For example, “I toasted the muffin and (then) poured the Hollandaise over it” shows sequence in time, and “I heard a scream and (therefore) turned around” shows causation (Davison et al., 1980, p. 24). The authors stated that splitting these clauses rather than using the discourse marker “and” could reduce text continuity and make fewer clues available to the reader to infer the relationships between the two clauses. From these and other examples, the authors found that excessive clause-splitting to create single-clause sentences from longer sentences could result in “a series of unconnected clauses, thereby adding to the task of the reader” (Davison et al., 1980, p. 21).

Davison et al. (1980) contrasted the practice of clause splitting with the practice of merging clauses from separate sentences into a single sentence. In one example, an adaptor had used the merger of sentences to place the resultative clause after the causative:

> [Original text] We had water to drink after that. We set out basins and caught the raindrops.
>
> [Adapted text] We set out basins to catch the raindrops so that we could have water to drink (Davison et al., 1980, p. 25).

The authors concluded that this merger had allowed an implicit relationship to be explicitly stated, which would likely make the text easier to parse. Having provided this contrast between splitting and merging, the authors stated that "the adaptor must weigh the advantages and disadvantages of these complementary processes, because producing a readable text at a given level is not the same as producing a text which scores at that given level of readability" (Davison et al., 1980, p. 31).

Davison et al.'s (1980) overall conclusion relating to sentence length was that when adapting texts, there are inherent tensions and tradeoffs which make it impossible to rely on readability formulas as literal guides to writing. First, some methods of simplifying vocabulary, such as paraphrase, may conflict with the injunction to shorten sentences. Second, the need to simplify grammatical structure and lexicon may conflict with the need to add markers of cohesion such as conjunctions. According to the authors, clause splitting and clause merging have their pros and cons and each may have value in different contexts, however readability formulas are not sensitive to the inherent tradeoffs in deciding whether or where such methods might be valuable. No matter whether clause splitting or clause merging might be preferable in a given context, conventional readability formulas will mathematically infer increased

comprehensibility when clauses are split and decreased comprehensibility when clauses are merged.

Readability formulas predict increased comprehensibility for texts that contain shorter words. However, this conclusion is not always borne out in research. Lenzner (2014) stated that longer words are not more difficult to understand than shorter words if the longer words exhibit high frequency in language and that a large amount of research supports the idea that word frequency plays a more fundamental role in word recognition than word length. For example, Lenzner pointed out that high frequency words tend to be processed as a single unit, whereas low frequency words are processed syllable by syllable.

Lenzner provided specific examples where longer words may be easier to comprehend than shorter words. First, many polysyllabic words are derivatives and compounds. Derivatives result from affixing prefixes (e.g., *pre-*, *co-*, *mis-*, *anti-*) or suffixes (e.g., *-er*, *-ion*, *-ing*, *-ism*) which speakers of the language tend to know the function of (Lenzner, 2014, p. 682). Derivatives provide strong clues as to the meaning of the word; longer words which are derivatives, such as *unemployment*, are relatively comprehensible compared to their length because the prefix *un-* and the suffix *-ment* are generally understood while the root word *employ* is relatively common in language. Words like *unemployment* can be contrasted to monosyllabic words such as *apt*, *dearth*, or *feint*, which are not only relatively infrequent but also provide no clues as to their meaning (Lenzner, 2014, p. 682). Readability formulas evaluate texts containing such short words as highly readable even if such words cause significant problems with reader comprehension.

In terms of compounds, Lenzer used examples such as *safeguard*, *overweight*, and *playground* to point out that some longer words are relatively easy to comprehend because their

individual word components tend to be understood quite easily (Lenzner, 2014, p. 682). Readability formulas tend to rate texts containing these multisyllabic words as less readable than texts with single-syllable words despite the fact that many single-syllable words are more difficult to parse than multiple-syllable compounds derived from frequent single-syllable words like “safe” and “ground.”

Concerns about the ability of readability formulas to evaluate derivatives and compounds are not a niche issue. Nagy and Anderson (1984) estimated that there are approximately 240,000 distinct words printed in English texts used in schools, including textbooks, workbooks, novels, poetry, and encyclopedias. Of these 240,000 words, the authors estimated that more than 170,000 of them are derived through suffixation, prefixation, and compounding. Lenzner (2014) concluded that “long words are not necessarily, or even usually, difficult to understand” (p. 683).

Although the broad range of reader characteristics which can affect text comprehension are beyond the primary scope of this review, the prominence of the argument within readability formula criticism that these formulas ignore the reader necessitates a discussion of some findings that may be directly relevant to the question of readability. For example, Klauk (1984) contended that sixth-graders better recall propositions which are placed higher in the text as opposed to propositions further down. Goedecke (2015) measured undergraduate student engagement over time when texts are more or less difficult, with difficulty measured through Coh-Metrix linguistic measurements such as syntactic complexity and word abstractness. Goedecke found that readers are more deeply engaged for the first 200-400 words, after which engagement decreases and reading times increase. These two studies support the idea that students may be better able to comprehend a text if the text is short or if the most important information is closer to the beginning, however neither text characteristic is measured by readability formulas.

Klare (1976) discussed several specific circumstances which may affect the validity of readability formulas in predicting comprehensibility. First, Klare found that readability formulas were more likely to predict reader performance when reader motivation was low. When reader motivation was high, readers showed the tendency of being able to achieve similar levels of comprehension from texts considered more difficult by readability formulas. Readers were also more likely to comprehend texts regardless of readability formula value when the content being read had a higher “interest-value" to the reader (Klare, 1976, p. 141). Klare also found that readers with more robust content knowledge as well as readers with more developed reading skills and higher intellectual ability were less likely to show increased comprehension when given texts that were easier to read according to readability formulas. Klare concluded that the usefulness and validity of readability formulas is thus strongly influenced by at least five factors within the target reader, namely the reader’s motivation, interest level in the topic, existing content knowledge, gross intellectual ability, and general reading skill. When some or all of these are factors were relatively high, readability formulas tended to be less potent predictors of reader comprehension, whereas when some or all of these factors were relatively low, readability formulas became more useful predictors of comprehension.

Bamford (1984) contended that student interest positively affects motivation levels and allows students to better comprehend text which is more difficult than the student is accustomed to reading. Beglar (2012) also noted motivation as an important factor in student comprehension, finding that pleasure reading was more effective than intensive reading for increasing reading speed over time without decreases in comprehension.

Elfenbein (2011) contended that even advanced linguistic analysis tools such as Coh-Metrix, with their impressive data output, are not necessarily sensitive to reader characteristics.

According to Elfenbein, even if two texts produced virtually identical numbers in Coh-Metrix, they might produce highly variable effects on participants.

Several authors have noted that the widespread use of readability formulas to grade educational texts by difficulty may cause educators and curriculum designers to act according to the assumption that students should only read texts identified by readability formulas as appropriate for their grade level. Dowell, Graesser, & Cai (2016) stated that learners can benefit from challenging material, particularly when the material is scaffolded effectively. At the other end of the spectrum, students can get a self-confidence boost and build self-efficacy through occasionally reading texts that are easy to comprehend. These authors contended that students need a balanced diet of texts with an emphasis on texts of intermediate difficulty. Armbruster et al. (1985) similarly recommended a balanced diet of texts and stated that "children *can* read and understand texts within a wide range of difficulty, and it is probably to their advantage to do so [emphasis in original]" (p. 20).

These findings form only a small part of the research on how reader characteristics can affect text comprehension. However, research in this area supports the conclusion that criticism of readability formulas for their inability to consider reader characteristics, including the cognitive psychology of readers generally, is well-founded. Although this review concentrates on linguistic and text factors such as cohesion and grammar, I will attempt where possible to place this research within the context of what we know about the cognitive psychology of readers and the attributes of specific readers. In this way, I hope that writers and adaptors can gain insight into potential relationships between explicit linguistic features and qualitative factors such as reader characteristics, needs, and abilities and thus be able to produce texts that are more specifically appropriate for their audience.

The question of whether simplification or elaboration is a superior strategy for adapting authentic texts for L2 learners has attracted considerable attention from researchers in L2 pedagogy; despite the fact that simplification and elaboration was not the primary focus of my research, I identified dozens of studies that attempted to answer this question. Although a full discussion of these studies is beyond the scope of this review, the debate concerning simplification and elaboration implicitly contains a debate about readability formulas in that elaborated texts tend to exhibit higher syntactic and lexical complexity compared to simplified texts (Long & Ross, 1993). As such, a short discussion of selected research in this field can help to evaluate whether readability formulas' emphasis on sentence length and lexical sophistication as measures of text comprehensibility is warranted. This discussion may also help identify linguistic features beyond sentence and word length to consider in tandem or as potential replacements.

Long and Ross (1993) argued that elaborative modification was effective in that syntactic and lexical complexity are often retained within the text but that this is compensated for by clarifying content and structure. The authors contended that shorter sentences are not necessarily easier as longer sentences often allow the writer or adaptor to "maintain clear references to unfamiliar concepts, remove pronouns with unclear antecedents, delete irrelevant details in distracting phrases, and highlight important points through pausing, stress, [and] topicalization" (Long & Ross, 1993, p. 30). To test this proposition, the authors took 13 authentic texts, adapted a simplified and an elaborated version of each, and had 483 college-level EFL students read three different versions of the texts. Thirty-question multiple-choice exams measured student comprehension.

Long and Ross (1993) provided Flesch-Kincaid and sentence-length measurements of the texts. The authentic versions of the texts had an average Flesch-Kincaid grade level of 12.8 and 23.7 words per sentence. The elaborated versions of the texts had an average Flesch-Kincaid grade level of 14 and 27.6 words per sentence, both higher than the authentic versions, while the simplified versions had an average Flesch-Kincaid grade level of 7.5 and 12.2 words per sentence, both significantly lower than the authentic versions.

The results of the comprehension tests indicated that text comprehension was highest for the simplified version and lowest for the authentic version. However, the authors found no statistically significant difference between comprehension scores for the simplified and elaborated versions of the texts. The authors concluded that readability formulas had inaccurately identified elaborated text as harder to understand than simplified text and argued that it is possible to create texts which are linguistically complex yet cognitively simpler.

Blau (1982) performed a study in which 18 paragraph-length texts with identical vocabulary and content were modified with respect to the combining of sentences. The purpose of this study was to identify the effects that syntax alone would have on EL comprehension and to evaluate student and teacher perceptions of sentence length modification in terms of how it affected comprehensibility.

For each paragraph, three different text versions were produced. Version 1 passages contained primarily short, simple sentences. Version 2 passages contained more complex sentences and thus contained fewer, longer sentences. Version 3 passages were similar to Version 2 passages in terms of sentence length but had fewer surface clues to underlying relationships. For example, Version 3 passages were more likely to delete optional relative

pronouns, to use implied conditional statements rather than using explicit markers such as the conjunction *if*, and to contain gerunds or derived nouns rather than infinitive verbs.

In order to test objective student comprehension of these texts, two separate sample groups from Puerto Rico were selected: 85 college students and 111 eighth-graders. Subjects from these two samples were asked to read one of the three text versions for each paragraph and answer multiple-choice comprehension questions. Student perception of text comprehensibility was also tested using a separate sample of 79 Puerto Rican college students, who were asked to read all three versions of three of the texts and to rate which text was easiest to comprehend. Twenty-one experienced teachers and 42 pre-service teachers were asked to rate text comprehensibility in the same manner.

Text readability was analyzed using the Fry readability scale. Version 1 texts ranged from first to fourth grade, Version 2 texts ranged from fifth to sixteenth grade, and Version 3 texts ranged from sixth to seventeenth grade.

In terms of objective comprehension, Blau found that the null hypothesis that there was no difference in student comprehension between Version 1 texts (simplified with shorter sentences) and Version 2 texts (elaborated with longer sentences) could not be definitively rejected. However, the difference ($p < .09$) approached statistical significance, suggesting that Version 2 texts were in fact easier to comprehend than Version 1 texts. Although the original hypothesis that Version 2 texts were superior to Version 1 texts was not conclusively established, Blau pointed out that the variance in Fry readability scores among Version 1 and Version 2 texts provided an alternative hypothesis. Namely, that Version 1 texts, which were consistently ranked at a lower grade-level by the Fry readability formula, would be easier to comprehend than Version 2 texts. This hypothesis could be rejected ($p < .045$), which indicated Fry readability

scores were not accurate predictors of text comprehensibility and that shorter sentences were not necessarily superior to longer ones.

In terms of student perception, for two of the three passages evaluated, Version 2 was ranked as significantly ($p < .005$) easier to comprehend than Version 1; students preferred texts with longer sentences. In contrast, Blau found that teacher evaluations of text comprehensibility for these three passages were firmly at odds with student evaluations; teachers overwhelmingly considered Version 1 texts to be most comprehensible and pre-service teachers were notably more likely than experienced teachers to rate Version 1 as most comprehensible. In essence, pre-service teachers' predictions of text comprehensibility were similar to the predictions of the Fry readability formula while students strongly disagreed with this assessment. Blau (1982) remarked that "this discrepancy should be taken as a warning to well-meaning teachers who may unwittingly be doing their students a disservice by selecting what they might mistakenly consider easy reading material" (p. 525).

Blau found interesting differences between the three groups of eighth graders, who had been sorted by proficiency into three homogenous groups. For the low-proficiency group, version 1 showed slightly higher comprehension scores, while the middle-proficiency group showed higher comprehension scores for version 2 and the high-proficiency group showed higher comprehension scores for version 3. High-proficiency eighth graders also showed higher comprehension scores for version 2 than for version 1. Blau (1982) concluded that "even eighth graders are cognitively mature enough to benefit from more mature sentence structure (p. 525). However, it could also be concluded that reduced sentence length might be effective for lower-proficiency readers whereas more complex syntax would likely benefit higher-proficiency readers. The optical illusion effect seen in Blau (1982) which caused teachers to identify shorter

sentences within text as more comprehensible was also found by Lotherington-Woloszyn in terms of student perception.

Lotherington-Woloszyn (1993) performed a case study in which two single-page authentic texts were simplified by two different editors of EFL textbooks, after which 36 intermediate ELL students at a Canadian university read all three versions of both texts. The purpose of the study was to report on how language was simplified by these editors, to evaluate subjects' comprehension of the different texts, and to evaluate subjects' perception of the comprehensibility of the different texts. The descriptions of the simplification process outlined by both editors provided useful insights. First of all, three of the methods described by Editor A could be described as elaborative rather than simplifying. Editor A stated that he or she would "clarify necessary background information," "simplify lexis by glossing," and "simplify lexis by providing a richer context for unfamiliar words" (Lotherington-Woloszyn, 1993, p. 143). These three methods have been described elsewhere in the literature as appropriate guidelines for text elaboration (Long & Ross 1993; Oh, 2001). Editor B's methods emphasized simplification rather than elaboration.

Each subject was interviewed twice. During each sitting, all three versions of a single text were read by the subject: the authentic text, the elaborated text adapted by Editor A, and the simplified text adapted by Editor B. The reading order was counterbalanced such that an equal number of students read the texts in each possible reading order. Student comprehension was assessed directly after the reading of the first text through oral recall. Evaluation of the subjects' perception of the text was accomplished in two ways: (1) by having students state which text was easiest to comprehend, and (2) by having students use a highlighter to identify the parts of each text which were perceived as difficult to comprehend.

Lotherington-Woloszyn (1993) found that there was a gap between how much students actually comprehended and student perception of which texts were easiest to comprehend. In terms of comprehension, “none of the text versions was significantly better comprehended by the subjects” (p. 144). However, in evaluating text difficulty, the subjects tended to rate the authentic texts as harder to comprehend than the simplified texts. The researcher explained that student perception of comprehensibility appeared unrelated to actual comprehension; subjects had underrated their comprehension of the authentic texts and overrated their comprehension of the simplified texts. She concluded that students had been “fooled by the apparently simplified surface features of the simplified texts, such as vocabulary difficulty and text length” (p. 148).

Lotherington-Woloszyn (1993) also found that students had identified with their highlighters a large number of comprehension problems per idea unit in the simplified texts of Editor B, yet they still tended to rate Editor B’s texts as easier to comprehend than the authentic texts. This finding strongly suggested that Editor B had created texts which appeared easy to comprehend, although they were not more comprehensible. The author attributed this finding to the fact that Editor B had placed a high priority on cutting out inessential information, which tended to reduce text length. As to whether this method was effective, the researcher found that students commonly highlighted idea units which had been reduced in size through the deletion of redundant information and concluded that redundancy was in fact useful for EL readers.

Overall, Lotherington-Woloszyn (1993) found that the use of authentic texts for intermediate English learners was justifiable and that reductive simplification was not superior to elaboration in aiding comprehension. Nonetheless, the author noted that the optical illusion qualities of simplified texts exerted a positive effect on learner confidence and may be useful for attracting readers to texts and for introducing the content of authentic texts.

The studies by Long and Ross (1993), Blau (1982), and Lotherington-Woloszyn (1993) clearly suggest that sentence length is not a highly significant factor that affects the comprehension of EL learners. If one accepts that more proficient readers are less likely to struggle with sentence length, as Blau (1982) found for eighth grade ELs, it is but a short step to hypothesize that secondary L1 readers are even less likely to struggle with longer sentences than L2 readers, so long as content difficulty and confounding linguistic factors are controlled for.

These three studies also provide a measure of warning to writers, adaptors, and teachers. Simply, the perception of the educator concerning text difficulty may not always be aligned with objective student comprehension or with student perceptions. Readability formulas inherently associate shorter sentences with increased reader comfort and comprehension and at least some teachers have shown similar tendencies; in cases where student perceptions and objective outcomes do not match these evaluations, results may be suboptimal and caution is warranted.

Conventional readability formulas attempt to measure lexical complexity through word length and syntactic complexity through sentence length; however, word and sentence length are *indicators* of lexical and syntactic complexity based on statistical correlations (Crossley et al., 2017b). Indicators are not the same as the thing to be indicated, and any ultimate evaluation of comprehensibility must consider diverse factors beyond word and sentence length which readability formulas cannot provide. Not least amongst these considerations is whether any chosen textual construction has improved the comprehension of actual readers of the text. Lockman (1957) has expressed this same idea, stating “wherever assessed understandability is low, regardless of measured readability level, revision to improve comprehension of the material in question is indicated” (p. 195). The reader, not the formula, must be the ultimate focus.

Can Readability Formulas Actively Mislead?

At least three studies have found negative correlations between readability formula scores and the assessed comprehension of readers. Because a negative correlation would imply that readability formula scores are not merely useless, but in fact actively misleading, an in-depth discussion of these studies is warranted. Two studies are particularly valuable; they discuss specific features within the texts that likely caused readability formulas to assess less comprehensible material as easier to read. These studies provide avenues not merely for criticizing readability formulas but for identifying specific linguistic features that writers and adaptors may wish to incorporate or avoid regardless of the effect of such modifications on readability formula output.

Charrow and Charrow (1979) conducted a study which sought to identify specific syntactic and lexical features within jury instructions which tended to make these texts more or less comprehensible to potential jurors. In Charrow and Charrow's first experiment, 14 texts taken from California civil jury instructions were given to 35 potential jurors. When the subjects attempted to paraphrase the original instructions, they successfully paraphrased only 38.6% of the semantic units of legal instruction identified by the authors. The remaining semantic units were either omitted or paraphrased incorrectly. The authors concluded that jury instructions are not adequately understood by the average juror.

Charrow and Charrow (1979) found that sentence length accounted for only 1.7% of the variation in subjects' comprehension of the information contained in that individual sentence. Similarly, the mean sentence length in each individual text accounted for less than 3% of the variation in subjects' comprehension of the information contained within that individual text. Texts with lengthier sentences were not significantly less comprehensible.

Charrow and Charrow analyzed the jury instruction texts using a Flesch readability formula and found a small negative correlation between Flesch readability scores and comprehension performance. This correlation was not statistically significant, however the authors concluded that readability formulas which rely heavily on sentence length are unreliable measures of comprehensibility, and that laws requiring insurance contracts and other legal documents to meet a certain readability standard as determined by a readability formula were misguided.

By identifying which semantic units of legal instruction had been poorly- or well-comprehended by the subjects and matching those semantic units to specific lexical and syntactic features, the authors attempted to identify which of these linguistic features had affected comprehensibility. The authors then rewrote the 14 jury instruction texts in an effort to eliminate linguistic features which had reduced comprehension while retaining the exact same semantic content as the original texts. These modified texts were tested on 48 potential jurors in a second experiment meant to evaluate whether the modified versions were easier to comprehend, and if so, to identify which specific modifications had exerted a positive effect on comprehensibility. In this experiment, the authors had each subject read seven of the original jury instruction texts and seven of the modified texts. As before, comprehension was evaluated by measuring how many semantic units of legal instruction subjects could successfully paraphrase.

The authors found that the modified instructions allowed subjects to achieve higher comprehension scores on the constituent units of legal information contained in the texts; in fact, 90% of subjects showed improvement in comprehension score on the modified instructions. The authors then performed statistical analyses on the specific linguistic modifications which had been made in order to test which modifications had been effective.

In producing the modified texts, Charrow and Charrow (1979) rephrased all nominalizations, defined by the authors as nouns that have been constructed from a verb; (e.g., *stipulation*, *admission*, and *recollection)*. In the first experiment, only 28.6% of the semantic information conveyed using nominalizations had been successfully paraphrased. The authors hypothesized that the use of nominalizations such as *the incorporation of* rather than phrases such as *when you are incorporating* tended to increase abstraction and to delete the actual doer of the action, making decoding more difficult (p. 1321-1322). Charrow and Charrow (1979) found that "de-nominalizing" these structures in the modified texts increased paraphrase scores by 45% for the associated semantic information.

Charrow and Charrow (1979) also modified the texts by rewriting clauses containing misplaced phrases, defined by the authors as prepositional phrases which are given non-standard placement and thus break up the continuity of the clause. Examples provided by the authors from the original jury instructions include: "a proximate cause . . . is a cause which, in *natural and continuous sequence*, produces the injury [emphasis in original]" (p. 1323). They pointed out that the phrase *natural and continuous sequence* precedes the verb modified by this phrase and reported that subjects incorrectly tended to assume that the italicized phrase modified the previous noun, *cause*, even though a singular noun such as *cause* cannot be in a continuous sequence (p. 1323). The authors identified nine misplaced phrases in the original instructions and reported that subjects successfully paraphrased only 24% of the semantic information associated with these constructions. Furthermore, the authors reported that an inordinate number of subjects who failed to paraphrase the semantic information associated with these constructions did not merely omit the instruction, but in fact paraphrased the instruction incorrectly. Charrow and Charrow (1979) found that modifying these structures resulted in a 24% increase in paraphrase

score and that subjects who failed to paraphrase the associated semantic information were now significantly more likely to omit the information rather than paraphrase it incorrectly.

Additionally, Charrow and Charrow (1979) rewrote sentences that used multiple negatives. Examples provided by the authors from the original instructions include the phrases "*without* which the injury would *not* have occurred [emphasis in original]" and "innocent *mis*recollection is *not uncommon* [emphasis in original]" (p. 1325). In the original experiment, they reported a 37% paraphrase score for semantic information containing a single negative, which dropped to 26% for constructions containing multiple negatives. For the second experiment, Charrow and Charrow (1979) reported that sentences containing multiple negatives had to be completely rewritten and were often paired with other variables, making a definitive statistical analysis impossible. However, they emphasized the low comprehension results obtained for multiple negatives in the first experiment.

As well, Charrow and Charrow (1979) rewrote many clauses which used the passive voice. In the first experiment, Charrow and Charrow (1979) found that "passives, when viewed as a class, are not an outstanding source of confusion" (p. 1325). However, the placement of the passive was found to be instrumental in determining the likelihood of successful paraphrase; in main clauses, 53.5% of the semantic information associated with passive constructions was paraphrased correctly, however this number dropped to 27% when the passive was placed within a subordinate clause. By rewriting many of these passive constructions and including them in the main clause of the rewritten sentences, the authors achieved a 48.5% improvement in paraphrase score (Charrow & Charrow, 1979).

Charrow and Charrow (1979) also rewrote sentences which used syntactic structures known as *complement deletion* and *whiz deletion*. They explained that *whiz* was a shortening of

which is and defined *whiz deletion* as a grammatical structure where the combination of a relative pronoun (such as *which*, *that* or *who*) and a form of the verb *to be* (such as *is* or *were*) is omitted (Charrow & Charrow, 1979, p. 1323). Examples from the original jury instructions provided by the authors include "questions of fact submitted to you" and "any statement of counsel made during the trial" (Charrow & Charrow, 1979, p. 1323). The authors stated that such constructions are standard in English but hypothesized that omitting the *which is* phrase forces the listener to reconstruct the missing grammatical information and thus tends to slow linguistic processing. *Complement deletion* was defined as the omission of the relative pronoun (such as *which* or *that*); examples provided from the original jury instructions include the phrase "if you are convinced it is erroneous" (p. 1323). In the original experiment, only 24.5% of the semantic information associated with *whiz deletion* or *complement deletion* was successfully paraphrased. They also reported the elimination of *whiz deletions* and *complement deletions* could not be separated from the modification of other variables, making a definitive statistical analysis impossible;however, they emphasized the low 24.5% comprehension result found in the first experiment.

Simplification of vocabulary often had a strong positive effect on comprehension. For example, the replacement of the phrase *must be imputed* with the phrase *would transfer* improved the comprehension score for the associated semantic item from 25% to 71% (Charrow and Charrow, 1979, p. 1336). Charrow and Charrow (1979) used a frequency dictionary to help identify uncommon words. Word frequency has been identified elsewhere as important for reader processing and word frequency measures are currently available through Coh-Metrix, a computer-assisted linguistic analysis tool discussed elsewhere in this review (Graesser et al., 2004; Crossley et al., 2008; Crossley et al., 2011; Graesser et al., 2011).

The researchers also found that text creators should pay attention to discourse structure, that is, “how the individual sentences are organized relative to each other” (p. 1326). For several of the jury instructions, subjects actually commented on how poorly organized the ideas were. Charrow and Charrow recommended at least one potential strategy, which was previewing and numbering the major ideas that will appear in the instruction.

The vast majority of linguistic features discussed by Charrow and Charrow (1979) were judged to have reduced comprehensibility, however there was one exception: the use of modals such as *must* and *should* (p. 1324). It is unclear if modals would be expected to affect comprehensibility in all texts; it seems entirely possible that the nature of jury instructions, namely telling someone what his or her duty is, made such commands more prominent in the mind of the listener and thus improved recall.

The major strength of Charrow and Charrow’s study was the statistically-robust linguistic analysis offered by the authors. This statistical analysis provides writers and adapters of text with highly specific advice concerning specific grammatical modifications which may improve comprehensibility. The researchers also provided every single jury instruction text in both original and modified format and discussed the changes for each text in great detail, which allowed the reader to see the modification process up close.

One potential weakness of this study was the fact that many linguistic modifications could not be separated from other modifications. This inability to separate confounding variables did not allow the authors to statistically determine whether modifying multiple negatives, *whiz deletions*, or *complement deletions* tended to improve comprehension. However, this weakness may result from the inherent complex nature of language, in which changing one element (for example a grammatical construction) is likely to require changes in other related elements, thus

impacting statistical results. From this point of view, the authors could be commended for not analyzing statistical results that had confounding factors; that is, they did not overstate their conclusions.

Lockman (1957) performed another study that found a negative correlation between readability formula score and assessed understanding. The researcher attempted to determine whether Flesch Reading Ease scores would correlate with assessed comprehensibility measured by a questionnaire in which readers would assess how difficult they found it to understand a text. Lockman chose Naval Aviation Cadets as his subjects. Subjects were 18 to 25 years of age, had two years of college or its equivalent, and had been inducted into the selective program based on a rigorous physical exam and the Navy Flight Aptitude Rating battery. For the texts to be assessed, Lockman used directions on standard psychological tests given to such cadets (e.g., academic aptitude, spatial orientation, attitudes, temperament, and personality).

Lockman (1957) created a seven-point rating scale for his questionnaires. The seven ratings were comparable to the standard style descriptions used to assess Flesch Reading Ease scores. They consisted of the following classifications: very easy, easy, fairly easy, standard, fairly difficult, difficult, and very difficult. The rating form instructed the subject to check one of these descriptions based on the subject's judgment of the understandability of the material that had been read. Using this scale, Lockman believed that statistical comparisons could be made with readability formula scores and style descriptions seen in Flesch Reading Ease and other formulas.

Lockman (1957) first assessed the texts themselves and found the Flesch Reading Ease (Flesch RE) style descriptions of the texts ranged from "fairly easy" to "difficult," with the mode being "standard." The texts were then rated by the subjects according to Lockman's

questionnaire. Depending on test administration schedules, between 129 and 273 cadets (median of 171) filled out questionnaires for each text.

The subjects rated the texts differently than Flesch RE had. The mode of subject understandability ratings for all exam instruction texts except one was “very easy” (53 to 74 percent of the ratings). The exception was the Navy Spatial Apperception Test instructions. In this case, 32 percent of subjects rated this text as “standard” and 26 percent as “fairly easy.” Coding the understandability ratings 1 through 7, corresponding with “very easy” through “very difficult,” Lockman then compared subject understandability ratings with Flesch RE scores using rank-order analysis. He found a rho of -.65, significant at the .05 level, indicating that Flesch RE score was negatively correlated with the understandability ratings reported by the subjects. Lockman concluded that “[Flesch Reading Ease] scores and understandability ratings were not measuring the same thing” (p. 196).

One strength of this study was the relatively high number of subjects, which allowed statistical significance to be established for the rank-order analysis. The study was also well-designed in the way it correlated Flesch RE style classifications with the exact same classifications used in the questionnaire (fairly easy, very difficult, and so on). This would appear to allow a robust analysis, since the exact same language the subjects used to assess understandability was correlated 1:1 with the descriptions that an educator or other person using the Flesch RE to analyze text would read when consulting Flesch RE’s style classifications. In other words, users of Flesch RE would not have any more information about readability or understandability than the subjects had.

In terms of weaknesses, Lockman (1957) himself pointed out that the subjects were “highly selected” (p. 196), presumably referring to the fact that all subjects had at least two years

of college education and that selection for pilot training within the U.S. Armed Forces is highly competitive. For this precise reason, it is difficult to trust the comparisons of mode, since it seems entirely normal that a highly selected group who have completed some higher education and are likely to exhibit higher than average general intellectual ability would rate most texts as being easier than the average person, let alone the average secondary-level student. However, even if the modal ranking analysis is suspect, the rank-order analysis appears strong, especially with a robust *n* value. Even if the subjects on average rated everything easier than the Flesch RE style classifications, an accurate readability formula should still rank an individual text easier when subjects say it is easier, and more difficult when subjects say it is more difficult. The negative correlation indicates that what Flesch RE rated as easier, subjects often rated as more difficult, and vice versa.

In terms of the subject population, Lockman (1957) indicated all subjects were male. There seems to be little reason to consider an all-male subject population as a significant confounding factor, however it does seem unfortunate that Lockman's robust study design was not applied to a more diverse population. Lockman's results are provocative in the way they clearly question the validity of readability formulas, however it would be more useful if these results could be widely replicated.

Overall, Lockman's (1957) article is extremely short and dense. There are no tables and no reported individual results. Perhaps most importantly, very little differentiated information is given about results obtained from different texts. Only five texts were analyzed; thus, if only one or two texts contained factors that significantly affected test results for this individual text, this could have affected the overall results. On that note, a more robust analysis of individual texts

and the specific syntactic and lexical characteristics that affected the results would have greatly improved the study's applicability to the field of readability analysis.

In the end, the reader may agree with Lockman's conclusion that readability formulas and "understandability" as reported by readers are not the same, but there is not much indication as to why exactly this might be the case. Perhaps Lockman himself did not find anything in the results that would provide significant further insight, however the study is at least highly evocative. One direction for further research would be to attempt to replicate these results with a more diverse population of subjects and texts, and with more robust technical linguistic analysis.

Lenzner (2014) attempted to analyze the ability of readability formulas to identify survey questions which had been identified by human raters and researchers as less confusing and easier to understand. Lenzner searched the relevant literature to find instances where a problematic survey question (e.g., vague questions or questions exhibiting excessive syntactic complexity) had been identified and a specific alternative proposed. He found 71 such question pairs and then tested these question pairs using four readability formulas: Flesch Reading Ease (FRE), Flesch-Kincaid Grade Level (FKG), Gunning Fog (FOG), and Dale-Chall (DC).

Lenzner found FRE identified the better question 51% of the time, whereas FKG scored 49%, FOG scored 39%, and DC scored 38%. Random chance would be expected to yield 50% accuracy; thus, three of the formulas performed worse than random chance, whereas FRE managed to exceed random chance by 1%. Lenzner concluded that readability formulas could not be used to accurately measure the comprehensibility of survey questions and that many formulas were more likely to mislead than inform.

Lenzner also examined whether or not the formulas agreed with one another in terms of their predictions. He found that FRE and FKG showed 80% agreement while FOG and DC

showed 90% agreement. However, the FRE and FOG formulas agreed only 16% of the time, while the FRE and and DC formulas agreed only 19% of the time. Lenzner concluded that if survey designers were to use different formulas to evaluate questions, they would tend to make entirely different text selections.

Lenzner's analysis was useful in that he provided specific examples of why the readability formulas and the human raters had made entirely different decisions. In one case, the term "health organization," which readers of the question had rated as confusing, had been replaced by "government health organization" in the improved survey question (Lenzner, 2014, p. 689-690). The readability formulas predictably assigned lower readability to the question; the single word "government" had been added, making the sentence longer, and "government" has three syllables, much higher than average.

Lenzner (2014) also discussed the use of negatives, contrasting the sentence "Policies that do not safeguard the environment are bad" with the sentence "policies that safeguard the environment are good" (p. 690). Lenzner stated that the negatively worded question is more difficult to comprehend, but that the FRE and FKG formulas both favored it; although the negatively worded sentence is two words longer, adding these two short words ("do not") to the sentence reduces the average number of syllables per word. This cautionary note about the use of negatives concords well with Charrow and Charrow (1979). Lenzner concluded that readability formulas are often not capable of analyzing syntactic complexity, in this case because they judge logical operators such as "not" to be short and easily comprehended words despite the fact that they add logical complexity and thus increase reader processing load (Graesser et al., 2004; Lenzner, 2014).

Lenzner discussed the limitations of using readability formulas on short texts, noting that one question scored 112.1 on the FRE formula despite the fact that FRE is designed to rate texts on a 100-point scale. Lenzner found that this same question was rated at a 0.1 grade level by FKG. Despite these ratings of extremely high comprehensibility, it seems unlikely that the question "have you ever had a Pap smear or Pap test?" (p. 690) would be comprehensible for students in the first grade. Lenzner's analysis of this issue may be relatively unimportant in the context of analyzing textbook passages and other long-form educational text, however it clearly implies that readability formulas are unreliable when analyzing important components of educational language such as test questions, photo captions, bullet points, chapter or section headings, and other short texts.

Lenzner (2014) emphasized that the nature of linguistics implies trade-offs when writing or adapting text and that readability formulas are essentially incapable of evaluating such tradeoffs. He contended that the specific variables readability formulas measure are not the most influential variables and suggested potential replacements worthy of further research such as word frequency, word ambiguity, the complexity of syntactic structures, and text purpose. He did find that readability formulas sometimes correlate with text difficulty but emphasized that correlation is not causation. He concluded that because readability formulas rely completely on the formal properties of text, they "neglect the semantic, pragmatic, psycho- and sociolinguistic aspects of language" (Lenzner, 2014, p. 692).

Readability Analysis Through Coh-Metrix

Recent progress in the fields of linguistics and computation has inspired the development of text analysis tools that go far beyond the classic readability formula variables of word and sentence length. Graesser et al. (2004) cited computational linguistics, corpus linguistics,

information extraction, information retrieval, and discourse processing as fields which have been particularly influential in the development of new linguistic analysis tools. Coh-Metrix, a freely-available online linguistic analysis tool, aims to exploit these advances and create measures of readability which are more sensitive to broader profiles of language and cohesion characteristics than conventional readability formulas (Graesser et al., 2004; McNamara et al., 2014).

Coh-Metrix analyzes texts on over 50 types of cohesion relations and over 200 measures of language, text, and readability (Graesser et al., 2004). Graesser et al. (2014) note that the original version of Coh-Metrix had nearly 1,000 measures, however that as of 2014, "approximately 100 measures are on the public website" (p. 215). McNamara et al. (2011) explained that many of the measures have not been validated and that many of the measures are highly correlated; thus, such metrics have not been released to the public.

Coh-Metrix 3.0 allows the user to freely access the online tool, paste any text into its text field, and submit the text for processing (Coh-Metrix 3.0, n.d.; Graesser et al., 2004). At this point in the process, 106 data points are created and the generated data can be viewed online or output and saved onto the user's computer (Coh-Metrix 3.0, n.d.).

The large amount of data Coh-Metrix outputs may be overwhelming for inexperienced users. Elfenbein (2011) noted that Coh-Metrix "challenges researchers to determine which [textual features] count and when" and called for a "teacher-friendly version [which] would make Coh-Metrix data more immediately interpretable for practical purposes" (p. 247). Thus, it will be useful to describe some key metrics that Coh-Metrix analyzes.

According to Graesser et al. (2004), the fields of corpus linguistics and psycholinguistics have been particularly useful in providing avenues for deeper analyses of lexis beyond character and syllable count. Coh-Metrix uses the MRC Psycholinguistics Database, which as of 2004

contained 150,837 words and provided information about 26 different linguistic properties of these words (Graesser et al., 2004). Rarer words are often not classed in terms of some linguistic properties within MRC (Coh-Metrix version 3.0 indices, n.d.; Graesser et al., 2004). However, the MRC Database can provide significant insight into a number of lexical properties identified within psycholinguistics research as highly meaningful to human comprehension (Crossley, Allen, & McNamara, 2012; Graesser et al., 2004).

Coh-Metrix Analysis of Lexis. Graesser et al. (2004) emphasized six measurements which are particularly useful for analyzing lexis: familiarity, concreteness, imageability, Colorado meaningfulness, Paivio meaningfulness, and age of acquisition. Paivio meaningfulness appears to have been superseded within the current Coh-Metrix 3.0 tool by Colorado meaningfulness (Coh-Metrix version 3.0 indices, n.d.), however a discussion of the other five measurements, as well as several related metrics, will be useful in future discussions of the research that Coh-Metrix has made possible.

Familiarity and word frequency. Familiarity refers to word frequency in the corpus of texts used to construct the relevant database (Graesser et al., 2004). Word frequency measures attempt to estimate how often a reader is likely to have encountered a specific word. Graesser et al. (2004) stated that "word frequency is an important measure because frequent words are normally read more quickly and understood better than infrequent words" (p. 197). Lenzner (2014) also emphasized the importance of word frequency, stating that words that occur less frequently take longer to process and are more difficult to understand.

Generally, a potential weakness of word frequency measures is that vocabulary in common usage tends to change over time (Lenzner, 2014). Coh-Metrix utilizes the CELEX database, specifically the 17.9-million-word corpus compiled in 1991 (Graesser et al., 2004;

McNamara et al., 2014). Such a corpus might treat some words such as *download*, *internet*, and *ringtone* as low frequency even though these words have become relatively frequent over time (Lenzner, 2014, p. 683). Similarly, words which have become less frequent over time, such as *washtub* and *cobbler*, would be measured as having relatively high frequency compared to current actual usage (Lenzner, 2014, p. 683). Lenzner notes that writers and adaptors can consult linguistic thesauruses which provide guidance in replacing lower-frequency words with higher-frequency alternatives.

Concreteness, imageability, and hypernymy. Words that are concrete evoke mental images and may refer to things which can be heard, tasted, or touched (Coh-Metrix version 3.0 indices, n.d.; McNamara et al., 2011). In contrast, "abstract words represent concepts that are difficult to represent visually" (McNamara et al., 2011, p. 8). MRC values for concreteness were compiled based on human ratings and reflect the dichotomy between relatively concrete words like *box* and *ball* and relatively abstract words like *protocol* and *difference* (Coh-Metrix version 3.0 indices, n.d.; Graesser et al., 2004).

Brysbaert, Warriner, and Kuperman (2014) explained the psycholinguistic method of compiling human ratings; subjects are asked to evaluate the degree to which the concept denoted by a word refers to a perceptible entity. The authors noted that subjects tend to rate visual and haptic entities as more concrete even though the definition of concreteness includes entities that can be heard, smelled, or tasted. MRC currently provides concreteness values for 4,293 unique words (Coh-Metrix version 3.0 indices, n.d.).

Concreteness can be closely compared to imageability, which measures the ease or difficulty of constructing a mental image of the word (Coh-Metrix version 3.0 indices, n.d.). Words like *dogma* and *overtone* are classed as having low imageability, whereas words like

hammer and *bracelet* are classed as having high imageability (Coh-Metrix version 3.0 indices, n.d.). As with concreteness, imageability ratings stem from human ratings (Graesser et al., 2004).

Coh-Metrix also measures hypernymy, that is, the extent to which a word can be classed as subordinate to more abstract categories (Coh-Metrix version 3.0 indices, n.d.; Graesser et al., 2004). For example, the word *seat* has a high hypernymy value as it can be classed as subordinate to the succeedingly more abstract categories *furniture*, *artifact*, *object*, and *entity* (Graesser et al., 2004, p. 198). Higher hypernymy tends to be associated with high concreteness, as seen in the previous example where *seat* and *furniture* are more concrete than *entity* (Coh-Metrix version 3.0 indices, n.d.; Graesser et al., 2004). Hypernymy values are provided by the WordNet database (Crossley et al., 2012; Graesser et al., 2004).

Texts containing more abstract words are more challenging for readers to understand (Coh-Metrix version 3.0 indices, n.d.; McNamara et al., 2011). Graesser et al. (2014) contended that "the abstractness-concreteness dimension has a robust impact on a wide array of cognitive processes, including comprehension" (p. 225). Crossley, Kyle, & Salsbury (2016) found that as L2 learners developed proficiency, the concreteness of their language output decreased and that there was strong statistical correlation between lower concreteness in student utterances and increased TOEFL scores. Crossley et al. (2012) showed that increased use of concrete and imageable words was seen in simplified beginner-level texts and noted this would be expected to allow the reader to decode the text more quickly. However, Crossley et al. (2012) also found lower hypernymy in the same texts, indicating increased abstraction. The authors noted that this result could stem from the fact that verbs in beginner-level texts are often less specific and thus potentially more abstract.

Meaningfulness. Words with higher meaningfulness scores, such as *people*, are strongly associated with other words, whereas words with lower meaningfulness scores, such as *abbess*, have weak association with other words (Coh-Metrix version 3.0 indices, n.d.). Crossley et al. (2012) found that beginner-level texts had higher word meaningfulness than advanced-level texts. As with their findings concerning higher concreteness and higher imageability, the researchers noted that higher meaningfulness would tend to allow the reader to parse text more quickly.

Although Graesser et al. (2004) distinguished between Paivio meaningfulness and Colorado meaningfulness in their original publication describing the development of Coh-Metrix 1.0, the current version of Coh-Metrix 3.0 does not export values for Paivio meaningfulness (Coh-Metrix version 3.0 indices, n.d.). Paivio meaningfulness appears to have been superseded by Colorado meaningfulness, for which MRC provides ratings for 2,627 words (Coh-Metrix version 3.0 indices, n.d.). The omission of Paivio meaningfulness since the advent of Coh-Metrix 1.0 may stem from the fact that the original Paivio meaningfulness corpus contained 925 words (Clark & Paivio, 2004) compared to the 2,627 words contained in the corpus developed in Colorado by Toglia and Battig (Coh-Metrix version 3.0 indices, n.d.).

Age of acquisition. Age of acquisition metrics reflect the fact that some words appear in children's language earlier than others (Clark & Paivio, 2004; Coh-Metrix version 3.0 indices, n.d.; Graesser et al., 2004). The norms used by Coh-Metrix 3.0 are based on a corpus compiled by Gilhooly and Logie in 1980 for 1,903 unique words (Coh-Metrix version 3.0 indices, n.d.).

Clark and Paivio (2004) noted that age of acquisition metrics are closely correlated with familiarity, concreteness, and word length, indicating that children "first learn words that tend to be concrete, short, and familiar" (p. 372). Interestingly, Clark and Paivio stated that age of

acquisition's multidimensional nature and close correlation to the psycholinguistic word characteristics of familiarity, concreteness, and word length may make it superior as a potential item in multiple regression analyses compared to familiarity, concreteness, and word length individually. From this perspective, age of acquisition metrics may also offer particular value to writers and adaptors seeking a quicker way of obtaining insight into a text's readability compared to individual analysis of familiarity, concreteness, and word length.

Polysemy. Polysemy measures word ambiguity, that is, the extent to which a single word can have multiple meanings (Coh-Metrix version 3.0 indices, n.d. Graesser et al., 2004). Graesser et al. (2004) use the example *bank*, which can mean a place to store money or the land next to a body of water. Coh-Metrix uses the WordNet database to output a polysemy metric which can be used to measure word ambiguity (Crossley et al., 2012; Graesser et al., 2004). Many words have multiple distinct meanings;. Words with high polysemy slow down reader processing, especially for less skilled and less knowledgeable readers (Graesser et al., 2004).

Type:token ratio. Graesser et al. (2004) stated that type:token ratio is a way of measuring how often the same words are used in a text. Each unique word in a text is a type, and each individual instance of a particular word is a token. A type:token ratio of one would mean that each word used only appears once, whereas a type:token ratio of seven would mean that each word is used seven times on average. A lower type:token ratio indicates increased reading difficulty since many unique words need to be encoded and integrated as the reader processes the text. Conversely, a high type:token ratio indicates a lower processing load, as individual words reoccur throughout the text (Graesser et al., 2004).

Coh-Metrix analysis of syntax. Coh-Metrix is capable of analyzing an impressive number of grammatical features. At a simpler level, it can distinguish between parts of speech,

and provide incidence values for nouns, verbs, adjectives, adverbs, and pronouns. Pronoun incidence can be further classified into instances of the first person, second person, and third person forms (Coh-Metrix version 3.0 indices, n.d.). Even these relatively shallow metrics cannot be analyzed through conventional readability formulas, and can be highly useful, as the density of pronouns for example is an important metric in predicting comprehension; texts tend to be more difficult when pronoun density is higher (Graesser et al., 2004).

The Coh-Metrix category "Text Easability Principal Component Scores" contains a component called "syntactic simplicity" which measures both sentence length and the extent to which the text contains familiar syntactic structures which are easier to process (Coh-Metrix version 3.0 indices, n.d.). This component may be useful for a quick overview of syntactic complexity.

More advanced syntactic complexity metrics provide specific avenues for predicting the difficulty of parsing the syntactic composition of sentences, for example by analyzing structural density, syntactic ambiguity, the incidence of higher-level and embedded constituents, and non-grammaticality (Graesser et al., 2004). As Lenzner (2014) noted, research has suggested that sentence length itself is not a cause of comprehension difficulty; rather it depends on syntactic structure (i.e., the specific ways words are combined to form a sentence). Coh-Metrix provides insight into the density of complex syntactic elements including adverbial phrases, prepositional phrases, gerunds, infinitives, and the agentless passive voice (Coh-Metrix version 3.0 indices, n.d.). Syntactic complexity metrics also include measurements of word classes that may signal logical or analytical difficulty including negations and logical operators such as *or* and *if-then* (Graesser et al., 2004, p. 197).

In addition to the syntactic simplicity measurement discussed above, Coh-Metrix can measure more specific indicators of syntactic complexity, for example the mean number of modifiers per noun phrase as seen in phrases like “the lovely little girl” (Graesser et al., 2004, p. 198). Verb phrases can be similarly analyzed. The number, length, and complexity of noun and verb phrases have been identified as contributing to and correlating strongly with general syntactic complexity (Graesser et al., 2004; McNamara et al., 2011).

In a similar vein, Coh-Metrix can measure the number of words before the main verb, also known as left embeddedness (Coh-Metrix version 3.0 indices, n.d.). Lenzner (2014) stated that left-branching syntax, meaning the number of clauses and qualifiers the reader must process before encountering the predicate of the main clause, was a significant predictor of more difficult reader processing. Lenzner (2014) provided the following contrasting sentences:

> (1): How likely is it that if a law was considered by parliament that you believed to be unjust or harmful, you, acting alone or together with others, would *try to do* something against it?
>
> (2): How likely is it that you, acting alone or together with others, would *try to do* something against a law that was considered by parliament and that you believed to be unjust or harmful? [emphasis in original] (p. 685)

Example (1), with high left embeddedness, requires the reader to remember a large amount of information before encountering the main verb and the predicate, whereas example (2), with lower left embeddedness, provides the reader with the main verb and predicate much sooner, allowing the reader to parse the information more efficiently.

Coh-Metrix calculates Flesch Reading Ease and Flesch-Kincaid Grade Level in order to provide a reference for quick comparison (Coh-Metrix version 3.0 indices, n.d.). Coh-Metrix

also provides statistics for mean sentence length and mean word length in terms of both letters and syllables Coh-Metrix version 3.0 indices. (n.d.). Thus, it allows the two features measured by conventional readability formulas to be analyzed in isolation.

In addition to Flesch Reading Ease and Flesch-Kincaid Grade Level, Coh-Metrix calculates "Coh-Metrix L2 Readability," a readability formula developed by Crossley et al. (2008) to better assess the readability of texts for L2 learners. Crossley et al.'s L2 Readability formula incorporates three Coh-Metrix values which have been identified in cognitive linguistics research as important for L2 readers, namely content word overlap, sentence syntax similarity, and word frequency. Content word overlap measures the proportion of word stems that overlap between pairs of sentences (e.g., *heat*, *heating*, and *heated*) (Graesser et al., 2004, p. 199). Such overlap helps readers construct meaning and speeds reader processing. (Coh-Metrix version 3.0 indices, n.d.; Crossley et al., 2008). Sentence syntax similarity measures the uniformity and consistency of syntactic constructions compared to adjacent sentences and globally; readers can more speedily parse texts when their constituent sentences are structurally similar to one another (Coh-Metrix version 3.0 indices, n.d.; Crossley et al., 2008). Syntactic dissimilarity can also signal textual complexity generally; texts with highly complex ideas or discourse tends to require relatively diverse sentence structures in order to express this complexity. The three variables of content word overlap, sentence syntax similarity, and word frequency can also be viewed independently (Coh-Metrix version 3.0 indices, n.d.).

Coh-Metrix analysis of cohesion. Both lexical and syntactic factors can affect text cohesion, which influences the ability of readers to interpret the substantive ideas of a text, connect ideas with other ideas, and connect ideas to higher-level global units such as topics and themes (Graesser et al., 2004; McNamara et al., 2010; McNamara et al., 2014). Analyzing texts

according to their cohesive elements was in fact one of the inspirations for the development of Coh-Metrix; “Coh” is short for cohesion (Graesser et al., 2011, p. 224). This is reflected in the numerous textual variables affecting cohesion that Coh-Metrix attempts to analyze.

Measuring cohesion is an aspect of textual analysis that presents particular computational challenges (Graesser et al., 2004). To aid analysis, researchers commonly place cohesive elements into sub-categories such as causal cohesion, spatial cohesion, temporal cohesion, and referential cohesion (also known as co-reference) (Crossley et al., 2012; Graesser et al., 2004; Halliday & Hasan, 1976; McNamara et al., 2010). Graesser et al. (2014) stated that “reading times, memory, and comprehension for text are significantly influenced by referential cohesion, causal cohesion, and other types of cohesion” (p. 225).

Texts with high referential cohesion contain words and ideas that overlap across sentences, which helps readers make connections between ideas and information and to experience the text as a coherent and cohesive whole (Coh-Metrix version 3.0 indices. (n.d.)). Halliday & Hasan (1976) noted that text “is not just a string of sentences” (p. 293) and emphasized that every sentence except the first contains some form of cohesion with previous sentences, and often the one immediately preceding. McNamara et al. (2011) stated that texts with low referential cohesion are more difficult for readers. Reflecting this insight, referential cohesion, or the relatedness between persons and objects, is its own category within Coh-Metrix data output (Coh-Metrix version 3.0 indices, n.d.; McNamara et al., 2010). Texts can be analyzed according to noun overlap, stem overlap, argument overlap, and content word overlap, with each measurement providing two mean values, one for adjacent sentences (local) and one for all sentences (global) (Coh-Metrix version 3.0 indices, n.d.; Graesser et al., 2004). Overlap measurements are broadly similar in that noun overlap measures whether sentences use the same

noun (e.g., *heat* and *heat*), argument overlap measures whether two nouns have the same stem (e.g., *heat* vs. *heating*, the gerund), and stem overlap measures whether nouns have overlap with other nouns *and* all other word categories such as verbs and adjectives (e.g., *heat* the noun and *heat* the verb or *heated* the adjective) (Graesser et al., 2004, p. 199). Content word overlap is a more general measure which measures overlap between all nouns, verbs, adjectives, and adverbs (Coh-Metrix version 3.0 indices, n.d.). All of these measurements help analyze the number of local and global connections available to the reader as he or she parses the text (Crossley et al., 2017b; Duran et al., 2007).

Coh-Metrix also uses a statistical analysis of word and text meaning called Latent Semantic Analysis (LSA) which can provide measurements of semantic overlap between sentences and paragraphs (Graesser et al., 2004; McNamara et al., 2011). The calculation of this measurement involves complex statistical methods; however the Coh-Metrix 3.0 indices provide an example which may be illuminating:

> Text 1: The field was full of lush, green grass. The horses grazed peacefully. The young children played with kites . . .
>
> Text 2: The field was full of lush, green grass. An elephant is a large animal. No-one appreciates being lied to . . . (Coh-Metrix version 3.0 indices, n.d.)

Text 1 would be analyzed as having much higher LSA scores than Text 2 since the words in Text 1 tend to be thematically related to a pleasant day in the park; whereas, the sentences in Text 2 tend to be unrelated. McNamara et al. (2011) provided further insight into LSA measurements, having stated ”LSA considers meaning overlap between explicit words and also words that are implicitly similar or related in meaning. For example, *child* in one sentence will have a relatively high degree of semantic overlap with *infant* and *mother* in another sentence” (p. 3).

LSA can provide statistical comparisons which even a human mind might have difficulty spotting or keeping track of, for example the similarity of sentence to paragraph, sentence to text, paragraph to paragraph, and paragraph to text (Graesser et al., 2004). This type of semantic co-referentiality is a powerful indicator of text cohesion (Crossley et al., 2012). LSA analysis could be very useful in quickly measuring an authentic or proposed text in terms of whether students are likely to find it semantically cohesive and thus be able to construct a coherent representation of the ideas presented.

Additional cohesive aspects of text are measured using components placed under the Coh-Metrix 3.0 category "Situation Model" (Coh-Metrix version 3.0 indices, n.d.). Situation model, in cognitive science, refers to "deeper meaning representations that involve much more than the explicit words" (McNamara et al., 2011, p. 4). The situation model is commonly referred to as the mental model (Graesser et al., 2011). For example, in narrative text, the situation model would include the plot (McNamara et al., 2011).

Situation model variables include the incidence of causal verbs such as *kill* or *enable* and the incidence of causal particles such as *since* and *because* (Graesser et al., 2004, p. 200). Graesser et al. (2004) stated that the "causal cohesion" metric is a ratio of causal particles to causal verbs, however the current Coh-Metrix output labels it directly as "Ratio of casual [*sic*] particles to causal verbs". The idea behind this metric is that causal cohesion is reduced when a text contains many causal verbs but few causal particles to signal how these events and actions are linked. Graesser et al. (2004) stated that causal cohesion may be unimportant in some texts, for example those describing static scenes, however they may offer significant insight when a text refers to events and actions which are related causally, as in many science texts and stories with an action plot. Situation model variables also include a metric for temporal cohesion, which

measures the repetition of tense and aspect (Coh-Metrix version 3.0 indices, n.d.; McNamara et al., 2011). Verb overlap, another situation model component, measures the extent to which verbs, which link actions, events, and states, are repeated across the text (McNamara et al., 2011).

The Coh-Metrix category "Text Easability Principal Component Scores" contains some elements that have been previously discussed, including syntactic simplicity and word concreteness. However, this section primarily measures cohesive components, including referential cohesion, deep cohesion, verb cohesion, connectivity, and temporality (Coh-Metrix version 3.0 indices, n.d.). Referential cohesion measures overall overlap of words and ideas between sentences while verb cohesion measures the extent of verb overlap between sentences. Deep cohesion measures the existence of connectives which show causation and logical relationships, whereas connectivity measures all connectives. Temporality measures cues concerning tense and aspect (Coh-Metrix version 3.0 indices, n.d.). Some of these components are broken down further within other Coh-Metrix components. For example, referential cohesion is later broken down into noun, argument, stem, and content word overlap, while connectives are broken down into distinct classes such as causal, logical, temporal, and additive connectives (Coh-Metrix version 3.0 indices, n.d.). Nonetheless, these broader components may be used to give a writer or adaptor quick insight into the cohesive elements of the text considered most important based on factors such as the genre of the text and the level of the prospective reader.

The final element of "Text Easability Principal Component Scores" is called narrativity. McNamara et al. (2011) stated that narrative text "tells a story, with characters, events, places, and things" (p. 7). They contended there is significant evidence that narrative texts are easier to read than informational texts. However, they also noted that narratives may have informational content, for example sections explaining the setting or context, while informational texts could

have narrative content, for example a science text narrating the journey of a water molecule. McNamara et al. found that narrativity is highly correlated with word familiarity, world knowledge, and everyday oral language, all of which support comprehension. On the other hand, narrative texts tend to have low referential cohesion and low verb cohesion; as the story travels through time and space, new entities, situations, and actions are encountered. By themselves, low referential cohesion and low verb cohesion would imply that students may have difficulty comprehending a text. However, because narratives tend to use more frequent words and often have high causal and temporal cohesion, the overall effect of narrative tends to help readers form a coherent mental model of the text; this robust mental model can compensate for challenging sentences and low word and concept overlap. McNamara et al. (2011) contrasted narrative texts with science texts, stating that science texts use rare words but that oftentimes authors offset this disadvantage by reducing syntactic complexity and increasing word and concept overlap, thus improving referential cohesion. These examples go far in explaining how individual components within Coh-Metrix can provide a picture of both the potential problems and the potential scaffolds that a text may or may not provide.

Cohesion has different effects on different readers. Dowell et al. (2016) stated that highly cohesive text has been shown to help students with low background knowledge. These readers require explicit cohesive clues, such as connectives and content word overlap, to effectively bridge the gaps between ideas. However, Dowell et al. (2016) found that students with adequate background knowledge could actually benefit from lower cohesion texts. The authors noted that this result appears counterintuitive, however lower cohesion text forces these readers to generate inferences and thus make connections between their background knowledge and the ideas presented in the text. This active generation of inferences can result in “deeper comprehension

and enhanced understanding of the situation model" (Dowell et al., 2016, p. 80). The decision of whether to emphasize the cohesive elements described above when producing text should be informed by the characteristics of the potential reader.

Utilization of Coh-Metrix

Coh-Metrix is a relatively new tool, however researchers are beginning to use it in the field to analyze the texts that writers and adaptors are currently writing and that students are currently reading. This section provides examples of ways that Coh-Metrix has been used to analyze several text types including texts for L2 readers and subject-area texts.

Coh-Metrix and L2 readers. Coh-Metrix's detailed measurements could theoretically allow writers and adaptors to design texts for L2 readers that emphasize linguistic features this population particularly relies on or tends to understand while excising linguistic features that are known to cause particular confusion among this population. In order to make such a process feasible, it would first be necessary to learn more about the linguistic features of the texts these students are actually reading and to study these students to learn which linguistic features they find most helpful, most understandable, and most troublesome.

Texts are commonly simplified for language learners and authors and adapters of such texts rely on a variety of approaches in order to achieve the goal of increased comprehensibility (Crossley, Allen, &McNamara, 2011). When simplifying texts, there are two major approaches available: the structural approach and the intuitive approach (Allen, 2009; Crossley et al., 2011). The structural approach utilizes wordlists and lists of linguistic structures which have been graded in terms of familiarity and complexity (Allen, 2009). The intuitive approach "relies on an author's subjective judgement of what learners at a particular level are able to comprehend and read" (Crossley & McNamara, 2016, p. 3). In forming this subjective judgment of what would be

easier for the reader to comprehend, authors use their experiences as a language teacher, language learner, and/or materials developer to provide guidance (Crossley & McNamara, 2016).

Crossley and McNamara stated "of these two approaches to text simplification (intuitive and structural), intuitive approaches are more common" (Crossley & McNamara, 2016, p. 3). Simensen (1987) found that even when provided with advice from publishers on how to adapt texts using structural methods, authors relied heavily on intuition. Young (1999) similarly found that even professors of theoretical linguistics tasked with providing explanations for modifications made to an authentic text self-reported as making a number of modifications based on intuition concerning what they believed students would find more comprehensible.

Crossley, Allen, and McNamara (2012) stated that the relatively vague nature of the intuitive approach implies could cause any author's text modification procedures to differ considerably that of others. For this reason, careful attention and consideration should be paid, and for this purpose authors should be aware of the positive and negative implications of specific intuitive modifications they might make.

Coh-Metrix grading of simplified texts by difficulty. Crossley et al. (2011) performed an analysis of texts from an English teaching website that had been intuitively simplified for L2 learners. The authentic texts used by the English teaching website (www.onestopenglish.com) had been taken from the *Guardian Weekly*, a British-based publication with significant international readership. From the original 100 authentic texts, a team of adaptors created three adaptations per text at three different difficulty levels, namely beginner, intermediate, and advanced. These 300 texts, with 100 in each difficulty category, became the corpus for Crossley et al.'s (2011) study. The goal of the study was to determine whether the Coh-Metrix L2 Readability measure, developed by Crossley et al. in 2008 and discussed earlier in this review,

would be superior to Flesch Reading Ease and Flesch-Kincaid Grade Level at classifying the three text difficulty levels in the same way they had been classed through human judgment.

Crossley et al. (2011) found that Flesch RE and Flesch-Kincaid performed similarly when classifying texts. Flesch RE correctly classified 49.3% of the texts, whereas Flesch-Kincaid correctly classified 44.3%; the difference between the two formulas was found to be not statistically significant at $p = .015$. Coh-Metrix L2 Readability performed better, correctly classifying 60% of the texts into the same category as the human adaptors. Comparing Coh-Metrix L2 Readability to both Flesch RE and Flesch-Kincaid, the authors found the difference in accuracy statistically significant at $p < .01$ for Flesch-Kincaid and at statistically significant at $p < .001$ for Flesch RE. The authors concluded that although conventional readability formulas showed some ability at text classification, Coh-Metrix L2 Readability was better able to classify texts based on their levels of intuitive text simplification as judged by L2 material writers. The authors ascribed this superiority to the contention that the Coh-Metrix L2 index has "stronger conceptual overlap to variables featured in psycholinguistic and cognitive accounts of reading" (p. 96).

In terms of limitations of this study, Crossley et al. (2011) first noted that more research was needed to improve the Coh-Metrix L2 index, perhaps by the inclusion of more variables. The Coh-Metrix L2 index only uses three variables, whereas "the process of intuitive text simplification likely modifies a much larger number of linguistic features" (p. 98).

The study showed one limitation common within the Coh-Metrix literature, namely that no actual readers were directly involved. As such, validation of difficulty levels is only available from the intuitive judgments of the adaptors themselves. Future research in this area could

consider validating the authors' intuitive judgments using student comprehension tests and thus laying a stronger foundation from which conclusions can be made.

Crossley et al.'s (2011) study might also have benefitted from a more specific linguistic analysis of the language features the adaptors had used when intuitively simplifying. However, the three authors of the 2011 study published a second paper in 2012 that analyzed the exact same corpus of texts in terms of their specific language features. The results of this research are described in the next section.

Coh-Metrix analysis of simplified texts' linguistic features. As with the 2011 study reviewed just above, Crossley et al. (2012) analyzed the corpus of modified texts from www.onestopenglish.com that included beginner, intermediate, and advanced versions of 100 authentic news texts. The purpose of the study was to investigate the linguistic effects of the three levels of intuitive text simplification in terms of language features used or omitted, and "to examine the benefits or disadvantages of intuitive simplification across proficiency levels" (p. 90). The authors used Coh-Metrix variables to perform their analysis.

Crossley et al. (2012) found a number of linguistic differences between texts of the three difficulty levels which would tend to cause beginner texts to contain more text features related to comprehensible input than advanced texts. At the most basic level, beginner-level texts were shorter while advanced-level texts were longer. In lexical terms, beginner-level texts had the lowest lexical diversity, while advanced-level texts had the highest. The beginner-level texts used words rated as being more frequent and more familiar; advanced texts had lower average word frequency and familiarity, indicating the use of less common words.

In syntactic terms, beginner texts had lower syntactic complexity than intermediate and advanced texts as measured by the number of words in each sentence before the main verb.

Similarly, beginner texts had higher levels of syntax similarity across sentences, while advanced texts had the lowest syntactic similarity, indicating that advanced texts used more varied grammatical constructions.

In terms of cohesion, beginner texts showed higher scores for noun overlap, while advanced texts had less noun overlap, indicating lower referential cohesion. Beginner texts were also rated higher in causal cohesion than advanced texts as they were more likely to use words such as *because*, *since*, *so*, and *then*. Beginner texts used more negative operators such as *not*, *cannot*, *no*, and *neither*; the researchers described these negations as capable of maintaining explicit links across texts. However, this conclusion does not match some of the research discussed above, which occasionally cautioned that a large number of negations can be indicators of reduced comprehensibility and increased complexity (Charrow & Charrow, 1979; Graesser et al., 2004). Some cohesive elements, such as connectives and indicators for temporal cohesion, showed no differences among the three text levels.

In addition to modifications which would be expected to make beginner texts easier to comprehend, Crossley et al. (2012) also found three major differences between the text difficulty levels which might cause beginner texts to be less comprehensible. Two differences were lexical in nature, while the other involved spatial cohesion. In terms of spatial cohesion, the advanced texts contained more prepositions of motion, which may help the reader comprehend important aspects of the text such as relative position and movement and thus help the reader create coherent mental models. However, the authors noted that some other indicators of spatial cohesion, such as motional verbs, locational nouns, and locational prepositions, did not show significant differences between the texts. The authors concluded that the overall effect on spatial cohesion was likely negligible.

In terms of lexis, advanced texts tended to use more specific verbs while beginner texts tended to use less specific verbs. Less specific verbs are potentially more abstract and have been shown to be produced later by L2 learners; L1 studies have similarly demonstrated that these less specific verbs are more difficult to acquire (Crossley et al., 2012). For example, the verbs *be*, *go*, and *have* may seem simple at first glance, but these words in fact have a wide variety of potential meanings (Crossley et al., 2012, p. 104). It might also be noted that be, go, and have are the basic word units of some advanced grammatical constructions; for example, go is used to construct a future tense (I am going to do x), while have is used to construct perfect and continuous aspects (I have gone, I am running).

In addition to using less specific verbs, beginner texts also had the highest scores for the use of ambiguous words (polysemy), while advanced texts had the lowest polysemy scores. These two differences, namely verb specificity and polysemy generally, are in fact quite similar to one another. Crossley et al. (2012) concluded that both lexical modifications likely stem from the use of more frequent words that tend to exhibit high polysemy. Thus, these two variables may present a tradeoff to adaptors, who will in some cases be forced to choose between a combination of high frequency and high polysemy words or a combination of low frequency and low polysemy words. In terms of navigating this tradeoff, Crossley et al. contended that the advantages of using more frequent words likely outweighed the costs in terms of increased polysemy. However, it seems clear from this discussion that text adaptors should closely consider the potential issue of ambiguity when choosing verbs and lexical items (e.g., the doctor made them well vs. the doctor made them skillfully).

The primary strength of Crossley et al.'s 2012 study was that it analyzed the texts using a wide variety of Coh-Metrix variables relating to many important factors theoretically affecting

comprehensibility, including lexical, syntactic, and cohesion factors. The fact that the authors were studying intuitively simplified texts means that the study has definite relevance for many writers and adaptors, as intuitive simplification is more common than structural simplification (Crossley & McNamara, 2016; Simensen, 1987) and is likely to be the method used by classroom teachers to modify texts for their students. The primary limitation of the 2012 study, as with Crossley et al.'s 2011 study discussed earlier, is that no students or readers were involved. For example, in concluding that the decision to use higher frequency and higher polysemy words seemed justified, the authors were essentially trusting the adaptors and/or relying on their own instincts and beliefs. Further research would be required to definitively determine whether readers of various proficiency levels feel the same way about how to navigate this tradeoff.

Coh-Metrix analysis of subject-area texts. Smolkin, McTigue, and Yeh (2013) stated that most analyses of Coh-Metrix have been conducted by researchers associated with developing this project. This contention is reflected to some extent in the contents of this review, which in the context of Coh-Metrix research commonly cites authors such as Graesser, Crossley, McNamara, McCarthy, Louwerse, Cai, and others, who developed and continue to develop Coh-Metrix (Graesser et al., 2004; McNamara et al., 2014). As such, it will be useful to discuss an example of a study performed by researchers not involved in Coh-Metrix development in order to show how Coh-Metrix can be used by writers, adaptors, teachers, and other educational professionals to evaluate texts within educational contexts.

Smolkin et al. (2013) performed an analysis of the explanatory content of science textbooks and science trade books commonly used by science instructors. The authors noted that although explanation is central to science, recent observational studies examining K-12

classroom science instruction found that explanation was not central to science instruction. The authors contended that science texts tended to offer more explanatory content than classroom discourse, but stated that the explanatory content was still often inadequate, with a disproportionately large percentage of both classroom discourse and science texts consisting primarily of "facts and description" rather than explanation (Smolkin et al., 2013, p. 1370). The authors hypothesized that textbooks could be a strong source of explanatory content for students and thus attempted to identify specific science texts that could effectively provide explanatory models to students.

Smolkin et al. (2013) explained that Smolkin, McTigue, Donovan, and Coleman (2009) had previously performed an analysis of explanatory language in science texts using human raters by coding individual clauses according to their use of explanatory linguistic elements (e.g., "*if*, *as a result, because, for this reason* [emphasis in original]" (Smolkin et al., 2013, p. 1370). Through this coding process, the authors measured what they called a "high-inference causal variable," which was an aggregate measure of linguistic markers showing "*condition*, *purpose*, *cause*, and *effect* [emphasis in original]" (Smolkin et al., 2013, p. 1370). However, the authors stated that this process was time- and resource-intensive and would thus likely be infeasible in many real-world contexts. The authors hypothesized that several Coh-Metrix components could be used to perform effective analysis of the same factors in a shorter time frame using fewer resources, for example by identifying linguistic elements such as conjunctions and connectives that underlie explanatory processes. Thus, the authors sought to find statistical correlations between Coh-Metrix components and the high-inference causal variable identified by human raters in the previous study.

Smolkin et al. (2013) found that the Coh-Metrix “causal cohesion” and “positive causal connectives” components showed moderate correlation with the high-inference causal variable previously identified, both significant at $p < 0.05$. The authors reported even stronger correlation between the high-inference causal variable and two Coh-Metrix components, namely argument overlap and stem overlap, indicating that texts which had been previously measured as high in causal explanation also showed strong referential cohesion (i.e., co-reference between sentences).

The authors then separated the original corpus into two sets of 10 texts. Texts selected for the first set had previously shown the highest values for the high-inference causal variable, whereas texts selected for the second set had previously shown the lowest values. The authors stated that the means between these two sets were not statistically different for “causal cohesion” or for “positive causal connectives” but that the set identified as containing higher levels of explanatory content also had higher Coh-Metrix values for argument overlap and stem overlap. This provided further support for referential cohesion as an important indicator of explanatory content. Smolkin et al. (2013) hypothesized that the use of repeated arguments such as shared pronouns, nouns, verb phrases, and stems between clauses and sentences was likely predictive of high-quality explanatory content showing causation. However, the authors concluded that the existence of overlap was likely necessary but not sufficient to establish the existence of high-quality causal content.

Smolkin et al. (2013) noted the limitations of Coh-Metrix in analyzing causation. For example, Coh-Metrix uses WordNet to identify causal particles and causal verbs, but the authors stated that human readers were also capable of identifying nouns and associated verbs that are typically associated with causal outcomes, for example in sentences such as “the tornado touched

down" (p. 1377). In this instance, WordNet is not capable of identifying "touch down" as a causal verb nor tornado as a noun implying a likely result even though a reader will likely infer that a tornado touching down will cause destruction. The authors concluded that a total lack of causal connectives and causal content according to Coh-Metrix does not imply a lack of high-quality explanatory content linking cause and effect.

Overall, the authors acknowledged the limitations of Coh-Metrix. They also recognized the limitations of their own study and that of many Coh-Metrix studies, in that textual analysis should eventually be directly validated in terms of student comprehension in order to demonstrate that certain Coh-Metrix components empirically correlate with more readable text. However, the authors stated that Coh-Metrix showed clear superiority to conventional readability formulas as a screening tool for text and suggested that Coh-Metrix's causal cohesion components could be used as rough guidelines.

The authors also noted the potential value of Coh-Metrix as a training tool for teachers. By having teachers enter text into Coh-Metrix and compare specific measurements with the linguistic characteristics of the text at hand, this might increase the teachers' attention to important linguistic components of science text and stimulate strong discussion and contemplation of the factors that might make science text a high-quality resource for students.

In some sense, it is to be hoped that this review functions in a similar sense. The research reviewed thus far has shown that linguistic analysis can be highly complex and that it is often difficult to give unqualified or universal advice. However, by understanding the limitations of both readability formulas and Coh-Metrix and by analyzing potential tradeoffs between linguistic features, writers and adaptors can become more cognizant of specific language features that may affect the quality of the texts their students will read.

Chapter III: Conclusion

This review sought to analyze research on readability formulas in order to identify specific writing and adaptation techniques that would make these texts more comprehensible for secondary-level students. Analysis has shown that this research can indeed offer a great deal of useful advice for the purpose of text selection, creation, and modification. On the other hand, such advice is far from systematic and often contains significant caveats. A great deal of work remains to be done in using scientific methods to evaluate specific linguistic factors that contribute to or detract from readability.

Summary

Graesser et al. (2004) stated that readability formulas "rely exclusively on word length and sentence length, two very simple and shallow metrics" (p. 194). Readability formulas have significant strengths, particularly objectivity, quantitativeness, ease-of-use, and statistical validity (Armbruster et al., 1985; Crossley et al., 2017b; Davison et al., 1980; Graesser et al., 2004). However, analyzing the weaknesses of readability formulas in evaluating text comprehensibility provided a useful line of questioning for this review; by identifying linguistic features which affect comprehension but are not measured by readability formulas, one can find specific areas where human analysis or more advanced computational linguistic analysis can intervene with potentially effective modifications.

Some weaknesses of readability formulas were found through human analysis or intuition. Davison et al. (1980) found that sentence merging and sentence splitting both had value in different situations. Lesser & Wagler (2016) found that combining sentences and adding a conjunction or connective could improve cohesion and readability despite increased sentence length. Lenzner (2014) found that longer words are often more comprehensible than shorter

words due to the widespread occurrence of compound and derivative words in English. Coh-Metrix researchers such as Graesser et al. (2004) and McNamara et al. (2014) defended a wide range of linguistic factors and measurements affecting comprehensibility which readability formulas cannot measure including lexical factors such as word frequency and word concreteness, syntactic factors such as left embeddedness and noun and verb phrase density, and cohesive factors such as connectives and word overlap.

Other weaknesses of readability formulas are evidenced through statistical comparisons of readability formula output and objective student comprehension. Charrow and Charrow (1979) found that Flesch readability formula values of jury instructions were negatively correlated with subject comprehension. Lockman (1957) found that readers disagreed with Flesch Reading Ease's prediction of the comprehensibility of test instructions and suggested that *readability* and *understandability* may not be the same thing (p. 195). Lenzner (2014) found that readability formulas were actively misleading when analyzing single sentences in that they were less effective than random chance at identifying superior and inferior sentences as identified by test readers and survey designers. Long and Ross (1993) found no difference in L2 reader comprehension between elaborated texts with an average Flesch-Kincaid grade level of 12.8 and simplified texts with an average Flesch-Kincaid grade level of 7.5. Blau (1982) found that L2 readers both preferred and found more comprehensible a group of texts with Fry readability scores ranging from 5th to 16th grade to a group of texts ranging from 1st to 4th grade.

Blau (1982) found that language teachers and particularly pre-service language teachers tended to believe that texts with shorter sentences and words were more comprehensible whereas students preferred and had better comprehension results with texts with higher syntactic and lexical complexity. Textbook authors and publishers were found in some cases to have a similar

bias; they closely considered readability formula output when creating texts despite the fact that writing specifically to formula may have a number of harmful effects on text quality and comprehensibility (Armbruster et al., 1985; Bruce et al., 1981; Davison et al., 1980; Graesser et al., 2004; Hiebert & Pearson, 2010). As Davison et al. (1980) pointed out, readability formulas rely on "the skill or common sense of the writer who is presumed to have created a coherent, well-formed text to which objective measurement may be applied" (p. 4). Another potential bias, specifically publication bias, was identified by Klare (1976), who found that six of 19 studies supporting readability formula validity were published in journals, whereas zero of 11 studies that did not support readability formula validity had been published.

Readability formulas cannot analyze reader characteristics, however numerous authors have argued that reader characteristics have a strong impact on text comprehension. Klare (1976) found that when reader motivation, interest level, existing content knowledge, gross intellectual ability, and/or general reading skill were relatively high, students could comprehend text which had been judged as too difficult for them. McNamara et al. (2011) and Dowell et al. (2016) offered further insight into consideration of readers, noting that even if a readability formula can predict student comprehension, it cannot identify the exact characteristics of the text that may be challenging or helpful to the student. For this purpose, Coh-Metrix may be of significant use, however the writer and adaptor must inevitably call upon his or her own knowledge about readers and language including the factors beyond readability formula analysis identified in this review. Intuition plays a huge role in any text creation process and intuitive modification modulated by the author or adaptor's knowledge and intentions has been defended as appropriate and useful by several authors discussed in this review (Allen, 2009; Armbruster et al., 1985; Crossley et al., 2011; Davison et al., 1980).

Taken together, these sections of the review suggested that readability formulas are not only of limited use in assessing textual features and text comprehensibility, but may be actively misleading. The relationship between writer and reader is simply too complex for any mathematical formula to offer more than relatively shallow insight.

From these points of view, the use of readability formulas is perhaps most defensible for the purpose of gaining a quick quantitative snapshot of a text's readability or grade level, after which further qualitative analysis should be applied by human raters. Statistically, the features that readability formulas measure, namely word length and sentence length, are correlated with readability and comprehensibility; this is reflected in the findings of validation studies. However, even if a statistical measure is roughly correct the majority of the time, it is still wrong a minority of the time. Thus, readability formulas can hardly be accepted as the final word. The true problem, perhaps, is that the ease-of-use and simple quantitative nature of readability formulas has caused them to become so ubiquitous and widely-used that they are over-applied and over-trusted.

This review also attempted to describe and analyze new forms of linguistic analysis through Coh-Metrix. This analytical tool builds on readability formulas' essential goal of using computational techniques to assess text readability, but uses more advanced analytical and statistical techniques to offer deeper insight. Coh-Metrix and similar tools have their shortcomings and require further development and research, however they undoubtedly offer a glimpse into the future of readability analysis.

Professional Application

This review focused on what research on readability formulas can tell us about techniques for producing more comprehensible text for students. As such, this section will

attempt to provide actionable recommendations to writers and adaptors based on the literature reviewed.

Users of readability formulas should be aware of their weaknesses, not place excessive trust in formula output, not write to formula, and look beyond the formula number when appropriate. Looking beyond the formulas may involve closer consideration of linguistic features such as lexis, syntax, and cohesion as well as closer consideration of the target reader's characteristics, needs, and abilities.

Readability formulas are fast, easy to use, and provide a quick method of rough measurement. For these reasons, they may be highly useful in specific situations, for example to provide a guideline when comparing texts or text versions. However, their limitations are legion, and they may be of little help or in fact actively misleading when analyzing certain types of texts or in the context of certain target readers. Several authors have also noted that different readability formulas often give widely varying results from one another, meaning that writers and educators do not have clear guidance on which formula, if any, is more trustworthy (Armbruster et al., 1985; Lenzner, 2014).

Numerous authors have commented on cases where increased syntactic and lexical complexity can benefit readers. Science texts may benefit from longer sentence length due to the ability to use more conjunctions and causative connectives which provide the reader clues to the relationships between objects, processes, causes, and effects (Armbruster et al., 1985; Smolkin et al., 2013). Texts for L2 readers or readers with lower background knowledge can include elaborative elements, for example clarifications, glossing, restatements, and even redundancy (Lenzner, 2014; Long and Ross, 1993; Lotherington-Woloszyn, 1993). It is also worth noting that readers can sometimes benefit from reading texts that are more difficult than what they are

used to; such texts can be used to introduce new language features and discourse types and to force the reader to generate more inferences and thus make stronger connections to their existing background knowledge (Armbruster et al., 1985; Dowell et al., 2016).

At least two authors directly tested student comprehension and found sentence length itself was not a statistically significant factor in reader comprehension (Blau, 1982; Charrow & Charrow, 1979). Other authors provided numerous examples where increasing sentence length tended to improve comprehension, for example by providing clarifying information or through increased use of cohesive elements such as connectives and conjunctions, which help elucidate relationships between objects and concepts (Armbruster et al., 1985; Lenzner, 2014; Lesser & Wagler, 2016; Smolkin et al., 2013; Tweissi, 1998). The literature on text elaboration and simplification also supports the idea that sentence length is relatively unimportant compared to other factors such as clarity, cohesion, and supporting information (Long & Ross, 1993; Lotherington-Woloszyn, 1993). Excessively long sentences are not to explicitly recommended as they increase cognitive load (e.g., working memory requirements) for the reader and can often be effectively split into different sentences without losing meaning (Davison et al., 1980; Graesser et al., 2004). However, the literature strongly suggests that shortening sentences should be of relatively low priority compared to the addition of clarifying information or cohesive elements where such additions seem appropriate.

Klare's (1976) analysis of reader characteristics describes situations where consideration of the target reader can give insight into whether readability formula output is more or less useful in a specific situation. Klare identified five reader characteristics which affect the value of readability formulas in predicting comprehensibility, namely the reader's motivation, interest level in the topic, existing content knowledge, gross intellectual ability, and general reading skill.

Where any of these factors are high, readability formulas may significantly underestimate students' ability to comprehend a text. This, in turn, may cause the educator to create or select texts which are too simple and of less value both in transmitting new knowledge and in providing opportunities to strengthen reading and thinking skills. Long and Ross (1993) noted that when text is excessively simplified, readers may be robbed of opportunities to be introduced to and gain insight into new lexical items and more advanced syntactic constructions.

A number of authors have warned against writing to formula, that is, always assuming shorter sentences and shorter words aid comprehension (Armbruster et al., 1985; Blau, 1982; Bruce et al., 1981; Davison et al., 1980; Graesser et al., 2004; Hiebert & Pearson, 2010). One of the most important insights an educator can gain from a close study of readability formulas is that the student should always be considered above the formula.

Vocabulary selection is a crucial factor affecting text comprehensibility. Word frequency (also called familiarity) has been emphasized by many authors as correlating strongly with whether readers are likely to recognize and understand a word. Higher frequency words can also be parsed more quickly, resulting in higher reading speed (Charrow and Charrow, 1979; Crossley et al., 2008; Crossley et al., 2012; Graesser et al., 2004; Lenzner, 2014). Lenzner (2014) found longer words can in fact be more comprehensible than shorter words due to the widespread occurrence of compound and derivative words in English. Compound words (i.e., words containing prefixes like *pre-* or *anti-* and/or suffixes like *-ion* or *-ism*) and derivative words (e.g., *playground*) often contain words or word parts that readers tend to understand relatively easily; writers and adaptors can take this into account when comparing the comprehensibility of words like *unemployment* with words like *apt* or *dearth*. (Lenzner, 2014, p. 682). Lenzner emphasized that word frequency was generally far more important than word

length in determining comprehensibility. Using computerized word frequency measures or simply considering the relative frequency of words in language allows human raters to more closely consider how the lexical properties of a text will affect student comprehension and make specific adjustments.

Authors and adaptors can also consider the concreteness and imageability of words, as higher concreteness and imageability make text easier to parse. Analysis of concreteness and imageability can be intuitive and/or machine-assisted, for example through Coh-Metrix. The use of words showing high polysemy, or the potential to have multiple meanings in different contexts, can make texts more difficult to comprehend and should also be closely considered. Age of acquisition is another highly interesting measurement; Clark and Paivio (2004) noted that age of acquisition is closely correlated with word frequency, concreteness, and word length; as such, this measurement may be useful both as an item in multiple regression analysis and as a quick yet effective measure of lexical difficulty.

Cohesive relationships are vital for helping readers understand texts. Cohesion helps readers make links between ideas, connect ideas to higher-level features such as topics and themes, and construct mental maps of the text as a coherent whole (Graesser et al., 2004; McNamara et al., 2010; McNamara et al., 2014). Cohesive factors such as conjunctions, connectives, logical organization of ideas, and the excision of information of questionable relevance should be closely considered by writers and adaptors.

Referential cohesion, for example by using the same or similar words over multiple sentences, helps readers make connections between sentences and may be particularly useful in texts which contain other difficult elements; repetition gives readers a common thread or foundation to which they can link multiple related ideas. Type-token ratio offers a similar

measurement in that it can identify the extent to which words are repeated throughout a text. A higher type-token ratio indicates that many words are repeated, which speeds up reader processing.

Narrativity is another feature that helps students form a mental model of the text. Authors in fields which tend to exhibit high proportions of expository text such as science and social studies may wish to consider writing or adapting texts more in the form of narrative stories rather than the expository style which is common in textbooks.

Left embeddedness is a potentially interesting measurement in that left-embedded sentences where the subject and verb are quickly introduced tend to be easier for the reader to process. Similarly, higher noun and verb phrase density tends to increase cognitive load as the reader must process a number of modifiers before reaching the modified noun or verb.

Connectivity is a single measurement of how many connecting words there are in a text. Such words can help readers better understand the relationships between disparate elements, for example elements in different clauses. Connectivity can be broken down into subordinate categories. Causal connectives may be of particular interest in science and other explanatory texts, as causal cohesion can measure the extent to which the reader is given clues about the relationships between causes and effects and between subjects acting and objects being acted upon. Temporal connectives, meanwhile, may be of particular interest in judging narrative texts.

This review has discussed a wide range of Coh-Metrix measurements, however Coh-Metrix does not have to be explicitly used to analyze a text in order for it to be useful. When writers and educators explore Coh-Metrix measurements and contemplate the rationale for their development and potential deployment, they are training themselves to make stronger intuitive decisions. For example, an author or adaptor does not necessarily need to measure left

embeddedness; he or she can simply attempt to introduce the main subject and verb of sentences and clauses relatively quickly to smooth reader processing. Similarly, measurements of connectivity and causal cohesion can act as reminders to provide readers with the explicit connective words needed for the reader to parse elements like time, sequence, space or position, cause and result, comparison and contrast, summary, and purpose. From this point of view, practicing textual analysis with Coh-Metrix may be useful as a teaching or training tool for writers, adaptors, and educators who want to understand more about language and how different elements of language and text may affect reader comprehension and text processing.

Limitations of the Research

A great deal of research on readability formulas and readability generally is limited by the sheer complexity of the subject being studied. The reading process involves all aspects of language in all their complexity and interrelatedness; this process is then moderated by the mind of the reader, an even more complex entity which is even more difficult to study. It can be difficult to design experiments or perform rigorous statistical analyses concerning any topic that has so many potential variables. This difficulty is reflected in the fact that much research on readability formulas and readability generally is theoretical rather than experimental in nature. To do experiments on readers, one must pass not only the obstacle of assembling large numbers of readers but also the obstacle of testing specific linguistic features which have many confounding variables in the context of readers and minds that have even more confounding variables. At the end of the day, it is often the case that one can theorize about text comprehensibility but cannot convincingly prove or disprove any single hypothesis. Solid facts backed by experimental science are much harder to come by than informed conjecture, no matter how logical or convincing such conjecture might be.

Another limitation of research on readability formulas is that although criticism of readability formulas is widespread and credible, it is often unclear what precisely should replace them; criticism without a valid alternative is of limited use. Readability formulas do have reasonable use cases and can be effectively deployed if one understands their limitations and carefully considers the context in which they are being used. Coh-Metrix and other more advanced tools are certainly promising as potential replacements, but these tools need further research, refinement, and integration into educational institutions including training programs for educators and textbook authors before they can truly come into their own.

Coh-Metrix is an interesting and useful tool with a number of limitations. First, Coh-Metrix and many of the measurements it outputs simply have not been studied enough. A relatively large number of studies of Coh-Metrix were performed by people involved with its development, whereas the number of studies done by other researchers and educational professionals is relatively small. Analyses of college-level textbooks are starting to appear, however analyses of secondary and elementary materials are quite rare and at least one study has argued that Coh-Metrix is of limited effectiveness for the implementation of Common Core Standards by virtue of the fact that it does not provide a single number that can be used to compare texts side-by-side (Nelson et al., 2011). Coh-Metrix's complexity is both a strength and a weakness.

Coh-Metrix developers and researchers have put forth claims that its measurements are strongly grounded in cognitive science and psycholinguistics. However, even if one finds the theoretical foundations convincing, many of the claims about the measurements and linguistic features expected to affect comprehension have not been studied in-depth on readers in an educational context. As such, it is difficult to determine which of the many measurements are

particularly useful and actionable; a human mind is simply not capable of looking at 106 data points and synthesizing a coherent picture of text comprehensibility. Elfenbein (2011) has called for a "more teacher-friendly version of the system" which would be vital towards the "sometimes elusive goal of translating the best research in education into results for the classroom" (p. 247).

Another limitation of Coh-Metrix is its black box nature. Even when one reads the explanatory notes, how a measure is calculated and/or its significance often remain unclear. As a simple example, the explanatory notes have a references list at the bottom, but footnotes or APA citations are few and far between (Coh-Metrix version 3.0 indices, n.d.). Thus, as far as the user is concerned, much of the stated information could have come from any of the studies, and it can be challenging to find out more about how a measure is calculated or the usefulness of a particular measurement.

The lack of footnotes and citations in the online user guide is an indication of another limitation of Coh-Metrix, namely that it has simply not received enough investment. The website could use a design upgrade; it sometimes outputs error messages when text is submitted, perhaps owing to a poor captcha system or other web design problems. It would also be useful to offer explanatory notes which show up when the user places the mouse cursor over a single measurement. As it stands, if one is not an expert, one has to constantly jump between the user guide and the data output on separate web pages to make any real sense of the data.

Increased investment might also allow the development of different versions as called for by Elfenbein (2011). For example, there could be one version for linguistics researchers which provides huge data output and a separate pared-down version that teachers and writers can use as a quick reference into the most important factors of readability as identified by the relevant research. A tool like this, carefully designed and with strong theoretical and experimental

backing, could provide both the evidence and user-friendliness needed for educators and school districts to abandon the readability formulas of the last century and embrace the linguistic analysis possibilities current technology theoretically affords us.

Implications for Future Research

The most important avenue for future readability research is bringing more readers into the process. A number of studies discussed in this review had interesting findings; however, in some cases it was difficult to establish statistical significance. Similarly, readability formulas and other methods of assessing text difficulty such as Coh-Metrix need to be validated at larger scale. Validation should consider the grade level and reading skills of the reader as well as other factors such as the genre and subject area of the text and whether the reader is L1 or L2.

As textual analysis moves beyond readability formulas into more advanced systems such as Coh-Metrix, it will become vital to understand which of the hundreds or thousands of linguistic measurements one might make are of primary importance in predicting readability. Is word frequency more important than word concreteness? Is sentence syntax similarity more important than left embeddedness? Is referential cohesion more important than narrativity? Thousands of such questions could be formulated and potentially tested on readers, and for all of these questions, we might further ask whether the answer depends on the grade level of the student, the genre or subject area of the text, and other context-specific factors. These are not easy questions, however answering questions such as these would allow us to build a cohesive system of empirically-tested claims about the effect of specific lexical, syntactic, and textual forms on reader comprehension. With a wealth of such research, educators could finally move beyond the confusing and contradictory advice about writing which was described by Klare (1976) and is also reflected in the ongoing debate about readability formulas. Such formulas

have been in use since at least 1923 (Hiebert, 2011); even so, the question of when or whether sentence or word length actually matters has not nearly been put to bed in a satisfactory manner.

Conclusion

This review attempted to review the research on readability formulas in order to determine whether such literature can offer specific advice to writers and adaptors about how to create more comprehensible text. It was possible to find such advice; however, the broad picture of this advice often appeared as a patchwork of potential techniques, sometimes with significant caveats, rather than a unified or systematic whole.

For diverse reasons, readability formulas have been widely criticized as weak and inherently flawed indicators of comprehensibility. However, newer tools such as Coh-Metrix may not yet be ready for widespread use. Progress in the field of readability is slow, perhaps because the interface between linguistics (a highly complex field) and the reader (an even more complex entity) has so many facets of potential interaction. Even where evidence exists about a linguistic feature that may increase or reduce comprehensibility, there is always the possibility that a different group of readers, a different type of text, or the addition or subtraction of a confounding linguistic feature might upset this result.

On the other hand, evidence concerning a number of specific linguistic modifications was strong and widely applicable, particularly in terms of considering whether word length and sentence length as measured by conventional readability formulas can justifiably be used as effective indicators of lexical and syntactic complexity. Word frequency appears to be a far more robust predictor of comprehensibility than word length, and newer tools allow word frequency to be measured in ways that were not possible several decades ago. Even more than word length, sentence length appears to be a relatively weak predictor of comprehensibility. This insight

potentially frees up the writer or adaptor to write longer sentences with more clarified content, stronger organizational markers, and more explicit connections between ideas.

Perhaps readability formulas cannot be trusted. Even so, by considering their weaknesses, it is possible to gain insight into the complexity they were invented to measure.

References

Allen, D. (2009). A study of the role of relative clauses in the simplification of news texts for learners of English. *System, 37*(4), 585-599.

Armbruster, B. B., Osborn, J. H., & Davison, A. L. (1985). Readability formulas may be dangerous to your textbooks. *Educational Leadership, 42*(7), 18.

Bamford, J. (1984). Extensive reading by means of graded readers. *Reading In A Foreign Language, 2*(2), 218-60.

Beglar, D., Hunt, A., & Kite, Y. (2012). The effect of pleasure reading on Japanese university EFL learners' reading rates. *Language Learning, 62*(3), 665-703.

Blau, E. K. (1982). The effect of syntax on readability for ESL students in Puerto Rico. *TESOL Quarterly, 16*(4), 517-28.

Bruce, B., Rubin, R., & Starr, K. (1981). Why readability formulas fail. Reading Education Report No. 28.

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

Charrow, R., & Charrow, V. (1979). Making legal language understandable: A psycholinguistic study of jury instructions. *Columbia Law Review, 79*, 1306-1374.

Clark, J., & Paivio, A. (2004) Extensions of the Paivio, Yuille, and Madigan (1968) norms. *Behavior Research Methods, Instruments, & Computers, 36*(3), 371-383.

Coh-Metrix 3.0 (n.d.). Retrieved from: http://tool.cohmetrix.com/

Coh-Metrix version 3.0 indices. (n.d.). Retrieved from: http://141.225.41.245/cohmetrixhome/documentation_indices.html

Crossley, S. A., Allen, D. B., & McNamara, D. S. (2011). Text readability and intuitive

simplification: A comparison of readability formulas. *Reading In A Foreign Language, 23*(1), 84-101.

Crossley, S. A., Allen, D., & McNamara, D. S. (2012). Text simplification and comprehensible input: A case for an intuitive approach. *Language Teaching Research, 16*(1), 89-108.

Crossley, S. A., Dufty, D. F., McCarthy, P. M., & McNamara, D. S. (2007a). Toward a new readability: A mixed model approach. *Proceedings of the 29th annual conference of the Cognitive Science Society,* pp. 197–202. Nashville, TN: Cognitive Science Society.

Crossley, S. A., Greenfield, J., & McNamara, D. S. (2008). Assessing text readability using cognitively based indices. *TESOL Quarterly: A Journal For Teachers Of English To Speakers Of Other Languages And Of Standard English As A Second Dialect, 42*(3), 475-493.

Crossley, S., Kyle, K., & Salsbury, T. (2016). A usage-based investigation of L2 lexical acquisition: The role of input and output. *Modern Language Journal, 100*(3), 702-715.

Crossley, S. A., Louwerse, M. M., McCarthy, P. M., & McNamara, D. S. (2007b). A linguistic analysis of simplified and authentic texts. *Modern Language Journal, 91*(1), 15-30.

Crossley, S. A., & McNamara, D. S. (2016). Text-based recall and extra-textual generations resulting from simplified and authentic texts. *Reading In A Foreign Language, 28*(1), 1-19.

Crossley, S. A., Rose, D. F., Danekes, C., Rose, C. W., & McNamara, D. S. (2017a). That noun phrase may be beneficial and this may not be: Discourse cohesion in reading and writing. *Reading And Writing: An Interdisciplinary Journal, 30*(3), 569-589.

Crossley, S. A., Skalicky, S., Dascalu, M., McNamara, D. S., & Kyle, K. (2017b). Predicting text comprehension, processing, and familiarity in adult readers: New approaches to

readability formulas. *Discourse Processes: A Multidisciplinary Journal, 54*(5-6), 340-359.

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

Davison, A., Kantor, R. N., Hannah, J., Hermon, G., Lutz, R., & Salzillo, R. (1980). Limitations of readability formulas in guiding adaptations of texts. Technical Report No. 162. Cambridge: Bolt Beranek and Newman Inc.

Dowell, N. M., Graesser, A. C., & Cai, Z. (2016). Language and discourse analysis with Coh-Metrix: Applications from educational material to learning environments at scale. *Journal Of Learning Analytics, 3*(3), 72-95.

Duran, N. D., McCarthy, P. M., Graesser, A. C., & McNamara, D. S. (2007). Using temporal cohesion to predict temporal coherence in narrative and expository texts. *Behavior Research Methods, 39*(2), 212-223.

Elfenbein, A. (2011). Research in text and the uses of Coh-Metrix. *Educational Researcher, 40*(5), 246-248.

Goedecke, P. J., Dong, D., Shi, G., Feng, S., Risko, E., Olney, A. M., & ... International Educational Data Mining, S. (2015). Breaking off engagement: Readers' disengagement as a function of reader and text characteristics, presented at International Conference on Educational Data Mining, Madrid, 2015.

Graesser, A. C., McNamara, D. S., Cai, Z., Conley, M., Li, H., & Pennebaker, J. (2014). Coh-Metrix measures text characteristics at multiple levels of language and discourse. *Grantee Submission*.

Graesser, A. C., McNamara, D. S., & Kulikowich, J. M. (2011). Coh-Metrix: Providing

multilevel analyses of text characteristics. *Educational Researcher, 40*(5), 223-234.

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

Halliday, M. K. & Hasan, R. (1976). *Cohesion in English*. London: Longman.

Hiebert, E. H. (2011). Using multiple sources of information in establishing text complexity. *Reading Research Report #11.03. Online Submission*. Online Submission.

Hiebert, E. H., & Pearson, P. D. (2010). An examination of current text difficulty indices with early reading texts. Reading Research Report #10-01.

Kantor, R., & Davison, A. (1981). Categories and strategies of adaptation in children's reading materials. In A. Rubin (Ed.), *Conceptual readability: New ways to look at text,* 4-14.

Kimmons, R. (2015). OER quality and adaptation in K-12: Comparing teacher evaluations of copyright-restricted, open, and open/adapted textbooks. *The International Review of Research in Open and Distance Learning, 16*(5), 39-57.

Klare, G. R. (1976). A second look at the validity of readability formulas. *Journal of Reading Behavior, 8*, 129-152.

Klauk, E. R. (1984). Staging and text comprehensibility: It's what's "up front" that counts.

Lenz, K., & Schumaker, J., & ERIC Clearinghouse on Disabilities and Gifted Education, A. V. E. S. P. (2003). Adapting language Arts, Social Studies, and Science Materials for the Inclusive Classroom. ERIC/OSEP Digest.

Lenzner, T. (2014). Are readability formulas valid tools for assessing survey question difficulty? *Sociological Methods & Research, 43*(4), 677–698.

Lesiak-Bielawska, E. D. (2015). Key aspects of ESP materials selection and design. *English*

for Specific Purposes World, 46.

Lesser, L., & Wagler, A. (2016). Tools for assessing readability of statistics teaching materials. *Journal Of Computers In Mathematics And Science Teaching, 35*(2), 153-171.

Lockman, R. F. (1957). A note on measuring "understandability." *Journal of Applied Psychology, 40*, 190-196.

Long, M. H., & Ross, S. (1993). Modifications that preserve language and content.

Lotherington-Woloszyn, H. (1988). On simplified and simplifying materials for ESL reading. *TESL Talk, 18*(1), 112-22.

Matthews, P. H. (2007). *The Concise Oxford Dictionary of Linguistics* (2nd ed.). Oxford: Oxford University Press.

McNamara, D. S., Graesser, A. C., Cai, Z., & Kulikowich, J. M. (2011). Coh-Metrix easability components: Aligning text difficulty with theories of text comprehension. In *annual meeting of the American Educational Research Association, New Orleans, LA*.

McNamara, D. S., Graesser, A. C., McCarthy, P.M., & Cai, Z. (2014). *Automated Evaluation of Text and Discourse with Coh-Metrix*. New York: Cambridge University Press.

McNamara, D. S., Louwerse, M. M., McCarthy, P. M., & Graesser, A. C. (2010). Coh-Metrix: Capturing linguistic features of cohesion. *Discourse Processes: A Multidisciplinary Journal, 47*(4), 292-330.

Mishan, F. (2005). *Designing authenticity into language learning materials*. Bristol: Intellect.

Nagy, W., & Anderson, R. (1984). How many words are there in printed school English? *Reading Research Quarterly, 19*(3), 304-330. doi:10.2307/747823

Nelson, J., Perfetti, C., Liben, D., & Liben, M. (2011). *Measures of text difficulty: Testing their predictive value for grade levels and student performance*. New York: Student

Achievement Partners.

Oakland, T., & Lane, H. B. (2004). Language, reading, and readability formulas: Implications for developing and adapting tests. *International Journal of Testing, 4*(3), 239–252.

Oh, S. (2001). Two types of input modification and EFL reading comprehension: Simplification versus elaboration. *TESOL Quarterly, 35*(1), 69-96.

Ovando, C. J., & Combs, M.C. (2018). *Bilingual and ESL Classrooms: Teaching in Multicultural Contexts* (6th ed.). Lanham, Maryland: Rowman & Littlefield.

Plakans, L., & Bilki, Z. (2016). Cohesion features in ESL reading: Comparing beginning, intermediate and advanced Textbooks. *Reading In A Foreign Language, 28*(1), 79-100.

Robinson, T., Fischer, L., Wiley, D., & Hilton, J. (2014). The impact of open textbooks on secondary science learning outcomes. *Educational Researcher, 43*(7), 341-351.

Smolkin, L. B., McTigue, E. M., Donovan, C. A., & Coleman, J. M. (2009). Explanation in science trade books recommended for use with elementary students. *Science Education*, *93*(4), 587–610.

Smolkin, L. B., McTigue, E. M., & Yeh, Y. Y. (2013). Searching for explanations in science trade books: What can we learn from Coh-Metrix?. *International Journal Of Science Education, 35*(8), 1367-1384.

Snow, S. (2015, January 28). *This surprising reading level analysis will change the way you write*. Retrieved from https://contently.com/2015/01/28/this-surprising-reading-level-analysis-will-change-the-way-you-write/

Street, C., & Stang, K. (2008). Improving the teaching of writing across the curriculum: A model for teaching in-service secondary teachers to write. *Action in Teacher Education, 30*(1),

37–49.

The Nation's Report Card (2015). *2015 Mathematics & Reading at Grade 12 National Achievement Level Results.* Retrieved from https://www.nationsreportcard.gov/reading_math_g12_2015/#reading/acl

The Nation's Report Card (2017). *NAEP Reading Report Card National Achievement-Level Results*. Retrieved from https://www.nationsreportcard.gov/reading_2017/nation/achievement?grade=8

Tinkler, S., & Woods, J. (2013). The readability of principles of macroeconomics textbooks. *Journal Of Economic Education, 44*(2), 178-191.

Tomlinson, B. (2012). Materials development for language learning and teaching. *Language Teaching, 45*(2), 143-179.

Tweissi, A. I. (1998). The effects of the amount and type of simplification on foreign language reading comprehension. *Reading In A Foreign Language, 11*(2), 191-204.

United States Census Bureau (2017). *School Enrollment in the United States: October 2016 – Detailed Tables* Retrieved from: https://www.census.gov/data/tables/2016/demo/school-enrollment/2016-cps.html

Young, D. J. (1999). Linguistic simplification of SL reading materials: Effective instructional practice?. *Modern Language Journal, 83*(3), 350-66.