COH-METRIX MEASURES TEXT CHARACTERISTICS AT MULTIPLE LEVELS OF LANGUAGE AND DISCOURSE

ABSTRACT

Coh-Metrix analyzes texts on multiple measures of language and discourse that are aligned with multilevel theoretical frameworks of comprehension. Dozens of measures funnel into five major factors that systematically vary as a function of types of texts (e.g., narrative vs. informational) and grade level: narrativity, syntactic simplicity, word concreteness, referential cohesion, and deep (causal) cohesion. Texts are automatically scaled on these five factors with Coh-Metrix-TEA (Text Easability Assessor). This article reviews how these five factors account for text variations and reports analyses that augment Coh-Metrix in two ways. First, there is a composite measure called *formality*, which increases with low narrativity, syntactic complexity, word abstractness, and high cohesion. Second, the words are analyzed with Linguistic Inquiry and Word Count, an automated system that measures words in texts on dozens of psychological attributes. One next step in automated text analyses is a topics analysis that scales the difficulty of conceptual topics.

Arthur C. Graesser
UNIVERSITY OF MEMPHIS

Danielle S. McNamara
ARIZONA STATE UNIVERSITY

Zhiqang Cai
Mark Conley
Haiying Li
UNIVERSITY OF MEMPHIS

James Pennebaker
UNIVERSITY OF TEXAS, AUSTIN

THE assignment of texts to students is a central concern of teachers, principals, superintendents, and other experts in educational policy. Text difficulty is one important criterion to guide such decisions in addition to considerations of curriculum, standards, and suitability of the subject matter for the age group. Students sometimes need to be challenged by texts on difficulty

THE ELEMENTARY SCHOOL JOURNAL VOLUME 115, NUMBER 2

levels that push the envelope on what they can handle. Students at other times need a self-confidence boost by receiving easy texts they can readily comprehend. Those who advocate Vygotsky's zone of proximal development would assign texts that are not too difficult or too easy, but at an intermediate zone of difficulty. The argument can also be made that there should be a balanced diet of texts on the difficulty dimension, with adequate scaffolding for difficult texts. Whatever principles of text selection are adopted, stakeholders would benefit from an automated analysis of texts on difficulty as well as other characteristics (Hiebert & Mesmer, 2013; Pearson & Hiebert, 2010).

There is a practical, logistical perspective that needs to be seriously considered. As appealing as it might be to imagine that teachers will have the time to review each and every text carefully, the task of individual quantitative and qualitative review is simply too daunting for individual teachers or even entire school staffs. This is where automation can assist in such decisions. Automated technologies can improve text assignments at various points in the process and reduce the load on teachers and other stakeholders.

Text difficulty has been seriously addressed in the Common Core Standards for English Language Arts (Common Core State Standards Initiative, 2010). The Council of Chief State School Officers acknowledged the need for a systematic comparative study of automated text-analysis tools. A systematic comparison study of seven text-analysis tools was conducted (Nelson, Perfetti, Liben, & Liben, 2011) on five samples of texts (to be discussed in a later section). Four of the tools provide a single metric of text difficulty (i.e., sometimes called "complexity," with the opposite being "ease"): (1) Lexile Framework (Lexile), (2) Advantage/TASA Open Standard (ATOS), (3) Degrees of Reading Power (DRP), and (4) Reader-Specific Practice (REAP). These four tools primarily capture word length, sentence length, and word-frequency measures, which are typical for unidimensional readability metrics. Two additional tools have several dimensions that tap levels of language and discourse in addition to providing a single text-difficulty score: (5) SourceRater and (6) Pearson Reading Maturity Metric (PRMM). The seventh tool, Coh-Metrix-TEA (Text Easability Assessor), also has several dimensions but originally did not provide a single overall measure of text difficulty. All seven tools were evaluated by computing correlations (Spearman's rho) between the difficulty scales and the grade levels or the achievement scores of narrative and informational texts.

Four of the conclusions from the Nelson et al. (2011) report are particularly relevant to the present article. First, the six tools with a single overall text-difficulty score (i.e., 1–6 above) had respectably high correlations with grade level (.59 to .79). Therefore, the word and sentence length variables are quite diagnostic of text difficulty. Second, the metrics based on word and sentence length variables were successful in predicting difficulty for (*a*) informational texts rather than narrative texts and (*b*) the grade bands 2–3, 4–5, to 6–8—flattening out between 6–8, 9–10, and 11–12. Such flattening out at higher grade bands has been a typical pattern of readability scaling since its inception (Dale & Chall, 1948). Third, the metrics based on multiple dimensions (tools 5–6 above) did a better job handling the narrative texts and discriminating grade bands between 6–8 and 11–12 than did the metrics based on word and sentence length components (tools 1–4). Therefore, there is value in pursuing metrics that tap multiple levels of language and discourse. And, fourth, there is no solid gold standard for defining grade level.

Most of the scales used to scale the grade level of texts were based on a panel of human experts in education or literacy research. Their judgments are influenced by an assortment of theoretical perspectives, practical experiences in education, and data. Those judgments surely include traditional readability formulas, so there is circularity in the assessment methodology and unfortunately minimal grounding in solid empirical data. Nevertheless, the comparison study does provide some very encouraging results and a foundation for scaling texts on difficulty.

The Coh-Metrix group never offered a simple dimension of text difficulty, as will be discussed in the first section of this article. Instead, their assumption was that text difficulty is inherently multidimensional and that the dimensions follow a multilevel theoretical framework for language and discourse comprehension (Graesser & McNamara, 2011; Graesser, McNamara, & Kulikowich, 2011; McNamara, Graesser, McCarthy, & Cai, 2014). This framework is summarized in the next section of this article. Nelson et al. (2011) ended up reporting how the five factors of Coh-Metrix-TEA correlated with five samples of texts, as will be summarized in the second section of this article.

The present article explores a new composite metric from the Coh-Metrix-TEA components that might be considered as a single dimension of text difficulty. The metric is labeled *formality*. Stylistic variation of language and discourse has traditionally been a core interest in virtually all explorations of language use (Clark, 1996; Hymes, 1974; Labov, 1972; Olson, 1977), and formality is one important construct in these explorations. Formal speech has been defined as "the type of speech used in situations when the speaker is very careful about pronunciation and choice of word and sentence structure" (Richards, Platt, & Platt, 1997, p. 144). Social context constrains the choice of language with respect to formality (e.g., the difference between the language on a rental contract and the gossip exchanged at a party). Formal language has also been defined as "a linguistic system based on logic and/or mathematics that is distinguished by its clarity, explicitness, and simple verifiability" (Bussmann, 1996, p. 169). This definition emphasizes the explicitness and unambiguity of formal language. These definitions indicate that formal expressions are related to linguistic and discourse systems, but they do not specify specific features related to formality. An automated analysis that is inspired by a multilevel theoretical framework would ideally provide additional clarity on the formality construct.

Some researchers have identified linguistic or discourse features that are diagnostic of formality at word, phrase, syntax, or text levels (Biber, 1988; Heylighen & Dewaele, 2002; Li, Graesser, & Cai, 2013). The present study follows this tradition of automating language and text analysis in order to provide an objective foundation for grounding theoretical claims. We define formality from the standpoint of the Coh-Metrix measures and report empirical assessments of its plausibility. An alternative foundation for computing formality is briefly reported, based on the Linguistic Inquiry and Word Count (LIWC; Pennebaker, Booth, & Francis, 2007). LIWC classifies words into dozens of linguistic and psychological categories based on ratings of human experts. A formality metric is created and tested on the basis of the LIWC analysis.

The final section identifies future directions in computing text complexity that extend beyond the explicit text, specifically, the need for *topic analysis*. Topic difficulty (e.g., Newtonian physics is more difficult than cooking vegetables) is needed to make sense of counterintuitive findings and trade-offs between language and discourse levels in previous Coh-Metrix analyses of text difficulty. Another angle to explore is the role of text complexity in helping to understand motivation and emotions during comprehension.

Coh-Metrix Measures at Multiple Levels of Language and Discourse

Models of reading and discourse comprehension uniformly assume that multiple levels of language, meaning, and discourse must be satisfactorily encoded or constructed in order for comprehension to succeed. Lower-level *basic reading components* include phonology, morphology, word decoding, and possibly vocabulary (Perfetti, 2007; Rayner, Foorman, Perfetti, Pesetsky, & Seidenberg, 2001), although vocabulary is typically positioned at a deeper level to the extent that words are tied to world knowledge. Without mastery of basic reading, deeper comprehension skills will not develop (Cain, 2010; Kendeou, van den Broek, White, & Lynch, 2009; Pearson & Hiebert, 2010). The higher-level *deeper comprehension components* move from words into sentence interpretation, construction of inferences, use of background knowledge, reasoning, and knowledge of discourse structures (Graesser & McNamara, 2011; Graesser, Singer, & Trabasso, 1994; Kintsch, 1998; McNamara, 2007; Perfetti, 1999; Snow, 2002). Deeper reading components are more time consuming, strategic, and taxing on cognitive resources of readers.

Multilevel Framework

Graesser and McNamara (2011) articulated a multilevel theoretical framework that integrates the large body of research on reading comprehension in various fields. The framework concentrated on deeper comprehension rather than basic reading. The framework is compatible with several other models in reading, discourse processing, and education that specify multiple levels of representation and processing components (Graesser, Millis, & Zwaan, 1997; Just & Carpenter, 1987; Kintsch, 1998; Perfetti, 1999).

The Graesser-McNamara framework identified six theoretical levels: *words, syntax*, the explicit *textbase*, the referential *situation model* (sometimes called the mental model), the discourse *genre and rhetorical structure* (the type of discourse and its composition), and the *pragmatic communication* level (between speaker and listener, or writer and reader). Whereas words and sentence syntax are straightforward, the other four levels call for some clarification.

Textbase. The textbase consists of the explicit ideas in the text that preserves the meaning but not the precise wording and syntax (Kintsch, 1998; van Dijk & Kintsch, 1983). There are basic idea units (sometimes called *propositions*) that contain a predicate (main verb, adjective, connective) and one or more arguments (nouns, noun-phrases, embedded propositions). For example, in the sentence "The Congress impeached the President," the predicate is "impeached" and the arguments are "Congress" and "President." *Co-reference* is an important linguistic method of connecting propositions, clauses, and sentences in the textbase (Halliday & Hasan, 1976; van Dijk & Kintsch, 1983). Referential cohesion occurs when a noun, pronoun, or noun-phrase refers to another constituent in the text. For example, in the sentence "If Congress impeaches the President, the impeachment will stimulate the news industry," the word "impeachment" in the second clause refers to the state associated with the predicate "impeaches" in the first clause. A referential cohesion gap occurs when the words in a sentence or clause do not connect to other sentences in the text. Cohesion gaps at the textbase level increase reading time (Haberlandt & Graesser, 1985; Just & Carpenter, 1987; Kintsch, 1998) and sometimes disrupt comprehension (McNamara & Kintsch, 1996; McNamara, Louwerse, McCarthy, & Graesser, 2010).

Situation model. The situation model is the subject matter content that the text is describing. In narrative text, this includes the characters, objects, spatial settings, actions, events, processes, plans, thoughts and emotions of characters, and other details about the story. In informational text, the situation model corresponds to the substantive subject matter (i.e., domain knowledge, topics) that the text describes. For example, the brief example on the impeachment of the president ("If Congress impeaches the President, the impeachment will stimulate the news industry") would potentially activate the following background knowledge: (*a*) causal networks of the events, processes, and enabling states that explain presidential impeachment, (*b*) properties of politicians in the political system, (*c*) the mechanisms of getting the attention of the news industry, and (*d*) goal-oriented actions of politicians. At least some world knowledge about U.S. politics and the news industry is needed to comprehend the example sentence. The situation model includes inferences that are activated by the explicit text and encoded in the meaning representation (Goldman, Braasch, Wiley, Graesser, & Brodowinska, 2012; Graesser et al., 1994; Kintsch, 1998; McNamara & Kintsch, 1996; van den Broek, White, Kendeou, & Carlson, 2009; Wiley et al., 2009). Zwaan and Radvansky (1998) proposed five dimensions of the situational model that apply to the thread of deep comprehension: causation, intentionality (goals), time, space, and people. A break in text cohesion occurs when there is a discontinuity on one or more of these situation-model dimensions. Such cohesion breaks result in an increase in reading time and generation of inferences (Rapp, van den Broek, McMaster, Kendeou, & Espin, 2007; Zwaan & Radvansky, 1998). Whenever such discontinuities occur, it is important to have connectives (e.g., because, so that, however), adverbs (finally, previously), transitional phrases (in the next section, later on that evening), or other signaling devices (headers) that convey to the reader that there is a discontinuity. Connecting words and expressions play an important role in Coh-Metrix, as will be discussed later.

Genre and rhetorical structure. Genre refers to the category of text, such as whether the text is narration, exposition, persuasion, or description (Biber, 1988; Grimshaw, 2003). These major genre categories can be broken down into subcategories within a taxonomy at varying levels of detail. A text has a rhetorical composition that provides a more differentiated functional organization of the discourse. In addition to paragraph organization, there are different rhetorical frames, such as compare-contrast, cause-effect, claim-evidence, problem-solution, and so on. Readers will struggle with texts without sufficient training in the structure, pragmatic ground rules, and epistemology of the genres and rhetorical structures of texts (Deane, Sheehan, Sabatini, Futagi, & Kostin, 2006; Eason, Goldberg, Young, Geist, & Cutting, 2012; Williams, Stafford, Lauer, Hall, & Pollini, 2009). One important contrast is the distinction between narrative and informational text, as will become apparent in this article.

Pragmatic communication. Just as a speaker in a conversation has a purpose in conveying a message to the listener (Clark, 1996), the writer tries to convey a message to the reader (Rouet, 2006). A good reader asks why the article was written and why it is being read. What is the point, theme, moral, message, or utility of the text? The pragmatic communication level is exceedingly important but is beyond the scope of the present article, which investigates difficulty of the text per se, as opposed to contextual variables that situate the text in the sociocultural context.

Coh-Metrix Scaling of Texts on Multiple Levels

Coh-Metrix is a computer facility that analyzes texts on most of the levels of the multilevel theoretical framework (Graesser, McNamara, Louwerse, & Cai, 2004; McNamara et al., 2014). Coh-Metrix is available in a public version for free on the web (http://www.cohmetrix.com). The original version of Coh-Metrix had nearly a thousand measures, but approximately 100 measures are on the public website for colleagues to use. Proponents of the CCSS encouraged the developers of Coh-Metrix to simplify the analysis and converge on a smaller number of factors. Therefore, a principal components analysis (PCA) was performed on 37,520 texts to identify central constructs of text complexity (Graesser et al., 2011). These texts included almost all of the 37,651 texts in the Touchstone Applied Science Associates (TASA) corpus; outliers of 10 standard deviations eliminated 131 unusual texts. The texts had a mean length of 288.6 words ($SD = 25.4$). One important reason for selecting this corpus is that it was representative of the texts that a typical senior in high school would have encountered from kindergarten through twelfth grade. Drama, poetry, and texts with headers, graphics, or special annotations were not included in the TASA corpus. The PCA resulted in eight dimensions that accounted for 67% of the variance among texts. The top five of these dimensions were incorporated in Coh-Metrix-TEA (http://tea.cohmetrix.com). The five dimensions of Coh-Metrix-TEA were analyzed by Nelson et al. (2011) in the comparative assessment of text complexity metrics. Highlights of these results are reported later in this article.

The five major dimensions of Coh-Metrix-TEA are succinctly defined as follows: (1) *Narrativity*: Narrative text tells a story, with characters, events, places, and things that are familiar to the reader. Narrative is closely affiliated with everyday oral conversation. (2) *Syntactic simplicity*: Sentences with few words and simple, familiar syntactic structures are easier to process and understand. Complex sentences have structurally embedded syntax. (3) *Word concreteness*: Concrete words evoke mental images and are more meaningful to the reader than abstract words. (4) *Referential cohesion*: High-cohesion texts contain words and ideas that overlap across sentences and the entire text, forming threads that connect the explicit textbase. (5) *Deep cohesion*: Causal, intentional, and other types of connectives help the reader form a more coherent and deeper understanding of the text at the level of the causal situation model.

Therefore, the five dimensions cover five of the six levels in the multilevel theoretical framework: genre (dimension 1), situation model (dimension 5), textbase (dimension 4), syntax (dimension 2), and words (dimension 3). Each of the five dimensions is expressed in terms of ease of comprehension. Text difficulty is defined as the opposite of ease, so principal component scores are reversed in measures of text difficulty.

It is beyond the scope of this article to describe how Coh-Metrix computes the measures and the five dimensions. This technical information is provided in previous journal publications (Graesser et al., 2004, 2011; McNamara et al., 2010), a book

(McNamara et al., 2014), and the help systems on the websites. Instead, this article reports how the five dimensions correlate with grade level, genre, and some other measures of text difficulty. The remainder of this section focuses on the TASA corpus, which was used to extract and norm the five dimensions, as reported in Graesser et al. (2011).

Our analysis of the TASA corpus had utility beyond it purportedly being representative of what seniors in high school would have read. TASA researchers provide measures that directly or indirectly reflect text ease/difficulty. Each text has an associated Degrees of Reading Power (DRP) score of text difficulty (Koslin, Zeno, & Koslin, 1987), with an approximate grade level associated with these values as specified in McNamara, Graesser, and Louwerse (2013). Each text is assigned to a text category by the TASA researchers. Most of the text genres were classified in language arts (n = 15,991), science (n = 5,349), and social studies/history (n = 10,438), but other categories included business, health, home economics, and industrial arts. Science and other informational texts cover topics less familiar to readers than the texts in language arts, which are predominantly narrative. The TASA measures of DRP (approximate grade level) and genre provided an objective foundation for validating the Coh-Metrix scores.

Also available for the TASA corpus was Flesch-Kincaid (FK) grade level (Klare, 1974–1975) and Lexile scores (Stenner, 2006). The three grade-level scales on text difficulty correlated highly (r = .89 to .94) when comparisons were made among FK grade level, DRP, and Lexile metrics. These unidimensional metrics of readability are all sensitive to sentence length, word length, and word frequency, so it is not surprising that they are all highly correlated. Word length and word frequency are robustly correlated in the negative direction. The high correlations among FK grade level, DRP, and Lexiles imply that they can be used interchangeably in the correlational analyses reported in this article.

Principal component scores of the five Coh-Metrix dimensions were correlated with two of the unidimensional metrics of text complexity, namely, FK grade level and Lexiles. The grade levels robustly decreased as a function of narrativity (r = −.536 for FK and −.487 for Lexile scores) and syntactic simplicity (r = −.665 for FK and −.731 for Lexile) and moderately decreased with word concreteness (r = −.208 for FK and −.075 for Lexile). Word frequency heavily loads on the narrativity dimension and sentence length on the syntax dimension, so there was no surprise about the robust negative correlations with grade level of FK and Lexiles. Referential cohesion had a small increase with grade level (r = .054 for FK and .047 for Lexile), whereas deep cohesion had a moderate increase with grade level (r = .138 for FK and .146 for Lexile). Apparently, cohesion is not on the radar of the standard readability metrics, even though discourse-processing researchers have established that cohesion is an important predictor of reading time and comprehension, as discussed in the first section of this article.

The analyses of genre had some obvious results as well as some unexpected but illuminating patterns. As would be expected, the narrativity scores were substantially higher for the language arts texts than for the two informational genres (science and social studies). The informational texts are on topics that are less familiar to readers, so they tend to be more difficult by virtue of the subject matter. Indeed, reading times are much longer for informational texts than narrative texts, whereas memory and comprehension scores tend to be lower for informational texts than narrative

(Graesser, Hauft-Smith, Cohen, & Pyles, 1980; Haberlandt & Graesser, 1985). The other dimensions of language and discourse apparently compensate for the inherent difficulty of informational texts. Compared to the language arts (narrative) genre, the science texts had substantially higher referential cohesion and simpler syntax. Narrative texts tend to occur with greater frequency at earlier grade levels than do informational texts, allegedly because of the easier vocabulary and subject matter (Hiebert & Fisher, 2007; Pearson & Hiebert, 2010). The multilevel framework of Coh-Metrix is therefore important to sort out some complex interactions among text constraints, test performance at different grade levels, and data reported in laboratory experiments.

Tests of Coh-Metrix Factors on Text Samples Analyzed by the CCSS

Nelson et al. (2011) reported analyses on four text samples in their comparisons among the seven tools for scaling texts on difficulty. The five dimensions of Coh-Metrix-TEA were included in these comparative assessments. Coh-Metrix-TEA can scale new texts on the five dimensions, based on the normative data provided on the TASA corpus. This section summarizes the results reported by Nelson et al. We were particularly interested in whether the results reported above for the TASA corpus ended up replicating for the four text samples considered by the CCSS.

The four text samples varied considerably in sample size, genre, and methods of scaling texts on grade level (for details on the samples, see Nelson et al., 2011). The text samples included (1) exemplar texts from Appendix B of the CCSS, (2) a set of standardized state test passages, (3) passages from the Stanford Achievement Test (SAT-9), and (4) comprehension passages from the Gates-MacGinitie Reading Test. The numbers of texts selected per sample were 168, 683, 97, and 98 for samples, 1, 2, 3, and 4, respectively. There was also a scale of student performance (Rasch scores) for the SAT-9 and Gates-MacGinitie texts. Criterion measures therefore included both grade level and student performance. The nonparametric Spearman's rho statistic was computed between these criterion measures and the metrics of text difficulty, including Coh-Metrix-TEA (hereafter referred to as Coh-Metrix). Follow-up analyses segregated the informational and narrative genres.

Most of the results showed correlations that replicated the TASA corpus analyses. Regarding grade level, the correlations with narrativity were negative for all four text samples, were strongly negative with syntactic simplicity for all four text samples, were leaning to negative with word concreteness for three out of four samples (near zero for the other sample), and were leaning to positive with deep cohesion for three out of four samples (near zero for the other sample). The only substantial difference between the TASA analyses and these four corpora was on the dimension of referential cohesion. Referential cohesion showed a small ($r = .05$) correlation with grade level in the TASA corpus, whereas there was a modest negative Spearman's rho with grade level in the four text samples analyzed by Nelson et al. (varying between $-.18$ and $-.41$). In summary, grade-level results replicated the TASA text analysis with the exception of referential cohesion.

There is at least one plausible explanation for the discrepancy between TASA and the four Common Core (CC) text samples with respect to the correlation between referential cohesion and grade level. Specifically, the genre distribution is very different across grade levels for TASA versus the four CC text samples, whereas there are robust differences in referential cohesion between genres. More specifically, there were three trends that need to be considered with respect to genre distributions. First, there is a widely acknowledged shift from narrative to informational genres as grade level increases in school systems. Second, there is a larger proportion of texts in the informational genre for TASA texts than for many of the CC samples. Third, the CC samples show a tendency to have complex literary narrative texts at the upper grade levels; such linguistic and discourse difficulty could potentially be offset by higher referential cohesion by the writers.

Graesser et al. (2011, Fig. 1) reported more in-depth analyses of referential cohesion when the contributions of grade level were segregated by genre. The two informational genres (science and social studies) showed a decrease in referential cohesion over grade levels, just as reported by Nelson et al. (2011) for the four CC text samples. For narrative text, there initially was a decrease from grades K–1 to 2–3, but then there was a very small increase from 2–3 to 11–12. Therefore, we conclude that differences in the distribution of genres can explain the discrepancy between the TASA texts and the CC texts in correlations between referential cohesion and grade level.

The Rasch performance measures of text difficulty showed similar trends as the grade-level analysis. For Gates-MacGinitie and SAT-9, the rho correlations were negative for narrativity, syntactic simplicity, and referential cohesion, but were mixed in sign and nonsignificant for word concreteness and deep cohesion.

Separate analyses were also conducted on the informational versus narrative genre on the CCSS passages. Grade levels tended to be higher on the informational texts, particularly for higher grade levels, on the six unidimensional metrics of text difficulty. Correlations between grade level and difficulty also tended to be higher for informational texts than narrative texts. Unfortunately, there were no head-on-head comparisons in the scores of texts in different genres for Coh-Metrix. As we argue throughout this article, analyses of text samples need to separate different genres and also the difficulty of the topics within each genre.

It is important to reiterate the point made by Nelson et al. (2011) that there was no gold standard for text difficulty in the analyses they performed. For example, the CCSS Exemplar texts had difficulty level defined by a committee of experts. The experts made serious attempts to justify their decisions, but such judgments hardly reflect an objective scientific foundation. One would need to conduct studies on reading time, comprehension, and other cognitive tasks in order to provide a more defensible gold standard.

Formality Measures from Coh-Metrix and LIWC

The Coh-Metrix team never created a unidimensional metric of text difficulty because of the commitment to the principle that difficulty varies across levels of language and discourse. However, the present article proposes a candidate construct that may serve as a possible singular dimension. The dimension is labeled *formality*. Formal discourse is the language of print or sometimes preplanned oratory when there is a need to be precise, coherent, articulate, and convincing to an educated audience. Definitions of formality were presented earlier and are elaborated below. At the opposite end of the continuum is discourse that has a solid foundation in oral conversation and narrative, replete with pronouns, verbs, adverbs, and reliance on common background knowledge. Formal language is expected to increase with grade level and with informational over narrative text. Therefore, we have formulated a composite scale on formality via Coh-Metrix that increases with more abstract words, complex syntax, cohesion, and informational text.

A formality measure was also constructed on the basis of LIWC word categories. The previous Coh-Metrix analyses had complex patterns of results at the word level. The results of word concreteness reported in Graesser et al. (2011) showed either a small or a curvilinear trend as a function of grade level. The abstract-concrete continuum has a long history in psychology (Mosenthal, 1996; Paivio, 1986), showing large effects on learning and memory, so it was expected that the variations with grade level would be robust. Nevertheless, aside from the abstract-concrete dimension, there are other psychological aspects of words that are worthy of attention. We therefore explicitly set out to add some additional and more nuanced tools to explore psychological characteristics of words, particularly the LIWC tools (Pennebaker et al., 2007).

This section reports follow-up analyses on the TASA text corpus and CC Exemplar texts with measures of formality based in Coh-Metrix and LIWC. These measures of formality were expected to enhance our scientific understanding of text difficulty.

Formality of Text

Formality is a universal dimension of stylistic variation, starting with Labov (1972) in the 1970s and carried on by other researchers (Biber, 1988; Heylighen & Dewaele, 2002). In the earliest studies of formality, researchers intuitively categorized the texts into formal and informal style according to the situation and context. For instance, academic papers or official legal documents were very formal, with careful choice of words and sentence structures (Richards et al., 1997). Personal letters or daily conversations with close friends, where there is shared knowledge between participants, tended to be less formal (Clark, 1996).

Linguistic features are diagnostic of the discourse that varies on the informal to formal continuum (Biber, 1988; Chafe, 1982). Formal language is the language of print, where a text can be inspected carefully and reinspected if it poses comprehension difficulty. Informal language lies in the oral tradition, where messages can be retrieved from memory only after being spoken (Olson, 1977). Face-to-face conversations contain more first-person pronouns due to an interpersonal, involved style (Chafe, 1982). A popular measure of formality is the F-score (formality score), which is sensitive to different word categories (Heylighen & Dewaele, 2002). Nouns, adjectives, articles, and prepositions are more frequently used in formal texts; pronouns, adverbs, verbs, and interjections are more frequent in informal texts. The F-score is computed as [(noun frequency + adjective freq. + preposition freq. + article freq. − pronoun freq. − verb freq. − adverb freq. − interjection freq. + 100)/2]. The F-score measure has successfully scaled texts on formality at the sentence level (Lahiri, Mitra, & Lu, 2011), but rarely at the text level. A more satisfactory measure would require other levels of language and discourse that are captured by Coh-Metrix.

Coh-Metrix Measure of Formality

One of the drawbacks of defining the formality of discourse on the basis of word categories is that the approach fails to consider syntax, discourse, and the goals of communication. This is where the dimensions of Coh-Metrix can lend a hand. Our underlying theoretical claim is that the goals of formal language are to increase precision of reference, analytical structure, and cohesion so that readers can accurately recover the message intended by the author. Therefore, consider the following: (1) Referring expressions (e.g., nouns, noun-phrases) need to be pitched at the optimal level of abstractness. (2) The syntactic and semantic composition of sentences needs to accurately express the intended claims. (3) The coherence and logical flow of the message needs to be laid out convincingly. Our proposed metric of formality increases with abstractness of words, syntactic complexity, cohesion (referential and deep), and the informational genre (as opposed to narrative). At the other end of the continuum, informal discourse tends to have concrete words, simple syntax, low cohesion (because knowledge-based inferences can fill the gaps), and high narrativity. We therefore computed a composite score of formality that integrated the five major dimensions of Coh-Metrix. The five dimensions were weighted equally. Given that there is a z-score for each of the five principal component dimensions of Coh-Metrix, a formality score for a text is computed according to formula 1 below.

$$\text{formality} = [\text{referential cohesion} + \text{deep cohesion} - \text{narrativity} - \text{syntactic simplicity} - \text{word concreteness}]/5. \quad (1)$$

It is important to acknowledge that the composite measure of formality does not merely consist of adding up the difficulty (opposite of ease) of the five dimensions. Formal texts are predicted to have low concreteness, syntactic simplicity, and narrativity; formal texts are at the difficult end of the continuum on these dimensions. In contrast, formal text has high referential and deep cohesion; these discourse characteristics helped comprehension rather than making it more difficult (McNamara et al., 2010). Therefore, the formality metric rests on a nuanced theoretical analysis of difficulty rather than a simple sum of scales of difficulty versus ease.

We conducted analyses on the TASA text corpus in order to assess the plausibility of the formality scale. Table 1 presents correlations on the TASA texts. Consider first the top half of the table that concentrates on Coh-Metrix, as opposed to the analyses of LIWC in the bottom half. The overall Coh-Metrix formality score had high correlations with FK grade level (.716) and Lexile scores (.664). As would be expected with two readability metrics, FK grade level and Lexiles were highly correlated (.902). The five components of Coh-Metrix had the anticipated relations with FK grade level and Lexiles that were compatible with the previously discussed relations with DRP scores. Specifically, the grade levels and Lexile scores dramatically decreased with narrativity and syntactic ease, modestly decreased with word concreteness, and modestly increased with the referential and deep cohesion.

Table 2 presents the same analysis on 246 texts from the CCSS Exemplar corpus. There was a high ($r = .721$) correlation between the formality score and FK grade level. As with the TASA corpus, the grade levels decreased with narrativity, syntactic ease, and word concreteness, but increased with deep cohesion. As reported in Nelson et al. (2011), but not previously with TASA, grade level decreased with referential cohesion. The formality metric is calibrated by the TASA norms, so we can now

Table 1. Correlations of Principal Component Scores from Coh-Metrix and Linguistic Inquiry and Word Count (LIWC) with Readability and Formality Metrics: TASA Corpus

	Flesch-Kincaid Grade Level	Lexile	Coh-Metrix Formality	LIWC Formality
Coh-Metrix:				
Coh-Metrix formality	.716	.664		.343
Narrativity	−.536	−.487		−.590
Syntactic ease	−.665	−.731		−.330
Word concreteness	−.208	−.075		.212
Referential cohesion	.054	.047		−.076
Deep cohesion	.138	.146		.135
LIWC:				
LIWC formality	.600	.601	.343	
Narrativity	−.595	−.500	−.622	
Processes, procedures, planning	−.181	−.215	.112	
Social relations	−.145	−.205	−.227	
Negative emotion	.142	.145	−.010	
Embodiment	−.217	−.182	−.184	
Collection	.264	.327	.084	

Note.—Pearson correlations of $r > .02$ are statistically significant at $p < .01$.

examine how well five components of formality correlate with the new CC Exemplar texts. As shown in Table 2, the formality scores decrease with narrativity, syntactical simplicity, and word concreteness, but increase with referential and deep cohesion. This pattern is precisely what was specified in the formula for formality.

Figure 1 plots the Coh-Metrix formality scores on a sample of texts in the TASA corpus as a function of the three genres (language arts, social studies, and science) and the six grade bands acknowledged by the CCSS. The formality scores increase linearly as a function of the grade bands. The relative ordering of formality shows the

Table 2. Correlations of Principal Component Scores from Coh-Metrix and Linguistic Inquiry and Word Count (LIWC) with Readability and Formality Metrics: 246 Texts from the Common Core Exemplar Corpus

	Flesch-Kincaid Grade Level	Coh-Metrix Formality	LIWC Formality
Coh-Metrix:			
Coh-Metrix formality	.721		.233
Narrativity	−.387	−.196	−.611
Syntactic ease	−.811	−.663	−.432
Word concreteness	−.153	−.436	.432
Referential cohesion	−.071	.339	−.209
Deep cohesion	.316	.515	.234
LIWC:			
LIWC formality	.538	.233	
Narrativity	−.630	−.539	−.584
Processes, procedures, planning	−.105	.269	−.638
Social relations	−.184	−.238	−.164
Negative emotion	.247	.202	.124
Embodiment	−.388	−.428	−.107
Collection	.344	.264	.636

Note.—Pearson correlations of $r > .16$ are statistically significant at $p < .01$ level.

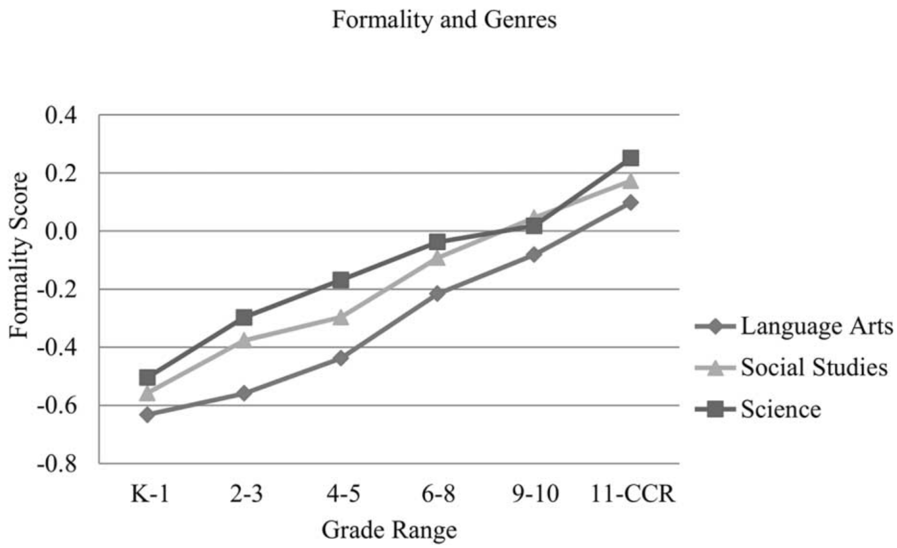

Figure 1. Mean Coh-Metrix formality scores as a function of three genres and six grade bands.

expected ordering of science > social studies > language arts. The two informational genres were quantitatively close in formality scores. It is important to note that this simple pattern of scores in the genre by grade-level plot is rather different than the more complex interaction plots when the five Coh-Metrix dimensions are analyzed separately (see plots in fig. 1 of Graesser et al., 2011).

Figure 2 presents an example TASA text with a very low formality score (−1.26) and contrasts it with a text with a very high formality score. Below each text are listed the *z*-scores on each of the five Coh-Metrix dimensions and also the FK grade level. The stilted formal language of the second text dramatically contrasts with the simplicity of the language and discourse of the first text. The cohesion scores are much higher for the second text than the first text, whereas the second text has more complex syntax and abstract words. Interestingly, the two texts are not too far apart on narrativity. The first text has a higher narrativity score than the second text even though the first was classified by TASA as social studies and the second as language arts. This illustrates how the formality score is not completely explained by text genre.

LIWC Measure of Formality

Formality scores were also computed on the basis of LIWC (http://www.LIWC.net; Pennebaker et al., 2007). The 2007 English LIWC dictionary contains 4,500 words that are classified or rated by experts on 64 word categories: 22 standard linguistic categories (e.g., pronouns, verb, tenses), 32 psychological categories (e.g., affect, cognition, biological processes), 7 personal categories (e.g., work, home, leisure), and 3 paralinguistic dimensions (assents, fillers, nonfluencies). Each word in a text is matched to a word in the dictionary and associated word characteristics are extracted. The LIWC tool computes the percentage of words in a text that fit into these linguistic or psychological categories.

LIWC categories have been shown to be valid and reliable markers of a variety of psychologically meaningful constructs (Pennebaker, 2011; Pennebaker et al., 2007).

Low Formality (From Nathan128.0101-DRP-40 in Social Study)

I can't wait to get out of bed. There is so much to do. My name is james quick. I have four brothers, but only george and daryl sleep in my room. I can't sit still so i get my brothers up. I am glad today is saturday. Mom does not have to work today. My dad is not living with us. My brothers and i are the men around the house. We have to take care of mom. Mom lets me do some of the food shopping. When i roll my cart around the store, i want to try everything. That is one of the things i like about living in the city. There are so many different kinds of foods. That is because the people living in the city come from all over the world. I bet you could try a new food every day for a year and not eat the same thing two times. After i put the food away, i take care of my kittens. I found this black-and-white cat in the street. It must have been days since she had eaten. It took a lot of talking, but mom let me take her in. Well, a few weeks later she had five kittens. Were we surprised! I named the mother cat spot. I did not name the kittens yet. When they are older i am giving the kittens to my friends. Then they can give them a name. I think i am going to feel very sad when the kittens are gone. There is an old car in a lot.

Formality = -1.26; narrativity = .63; syntactic simplicity = 3.78; word concreteness = 2.00; referential cohesion = .38; deep cohesion = -.26; Flesch-Kincaid grade level = .16

High Formality (From Xanthe01.07.01-DRP-77 in Language Arts)

It was to my uncle toby's eternal honour, though i tell it only for the sake of those, who, when cooped in betwixt a natural and a positive law, know not for their souls, which way in the world to turn themselves, that notwithstanding my uncle toby was warmly engaged at that time in carrying on the siege of dendermond, parallel with the allies, who pressed theirs on so vigorously, that they scarce allowed him time to get his dinner, that nevertheless he gave up dendermond, though he had already made a lodgment upon the counterscarp; and bent his whole thoughts towards the private distresses at the inn; and except that he ordered the garden gate to be bolted up, by which he might be said to have turned the siege of dendermond into a blockade, he left dendermond to itself, to be relieved or not by the french king, as the french king thought good; and only considered how he himself should relieve the poor lieutenant and his son. That kind being, who is a friend to the friendless, shall recompence thee for this. Thou hast left this matter short, said my uncle toby to the corporal, as he was putting him to bed, and i will tell thee in what, trim. In the first place, when thou madest an offer of my services to le fever, as sickness and travelling are both expensive, and thou knowest he was but a poor lieutenant, with a son to subsist as well as himself out of his pay, that thou didst not make an offer to him of my purse; because, had he stood in need, thou knowest, trim, he had been as welcome to it as myself.

Formality = 2.21; narrativity = 1.61; syntactic simplicity = -5.59; word concreteness = -.63; referential cohesion = 1.81; deep cohesion = 4.61; Flesch-Kincaid grade level = 28.14

Figure 2. Example excerpts with low and high Coh-Metrix formality.

The different categories of content words would be expected to predict psychological dimensions. For example, negative emotion words would be diagnostic of gloomy texts. Interestingly, LIWC researchers have documented that the function words (particularly pronouns; Pennebaker, 2011) are diagnostic of social status, personality, and various psychological states. There are gender, age, and social class differences in function word use. For example, first-person singular pronouns (e.g., *I, me, my*) have higher usage among women, young people, and people of lower social classes.

We conducted analyses with LIWC to explore whether word features alone can predict text difficulty and formality. Scores for 64 LIWC categories were computed on the 37,520 TASA texts. A principal components analysis (PCA) with varimax rotation was conducted to reduce the 64 measures to fewer dimensions. The same procedures were followed in our analysis of Coh-Metrix (Graesser et al., 2011). When a PCA was conducted with LIWC indices, six principal components accounted for 40% of the variance between texts. This is a sizable effect, although less than the 67%

variance reported for the eight principal components (PCs) of Coh-Metrix and the 51% variance for the top five PCs included in Coh-Metrix-TEA (see Table 1). Six LIWC dimensions represent a suitable number because there was a leveling off in the eigenvalues on a scree plot after six factors.

The bottom half of Table 1 presents the six LIWC dimensions as well as correlations between the six PC scores and other measures of text difficulty. The strongest dimension was narrativity, which accounted for 14% of the text variance. The incremental percentages of text variance for the other five dimensions were processes, procedures, and planning (9%); social relations (6%); negative emotion (5%); embodiment (3%); and collection (3%). It is noteworthy that LIWC narrativity was the only dimension that had an analogue with the five Coh-Metrix dimensions, also labeled as narrativity. Narrativity is a robust factor that emerges in many text-analysis tools (see also Biber, 1988). Moreover, there were high negative correlations between LIWC narrativity and FK grade level (−.595) and Lexile scores (−.500); the corresponding correlations were also high between Coh-Metrix-TEA narrativity and FK grade level (−.536) and Lexile scores (−.487). The other five PCs of LIWC had modest correlations with the two text-complexity metrics (correlations between −.217 and .327).

Labels for the LIWC dimensions were constructed after observing the LIWC word categories that loaded highly on the PCs and by examining texts with very high versus low PC scores. The *processes, procedures, and planning* dimension included texts that (*a*) describe actions and events in procedures or processes that are conveyed in the present tense or (*b*) forecasted events, goals, plans, or recommendations for the future. The present and future tenses in these passages contrast with the past-tense verbs in the first narrativity dimension. The *social relations* dimension had many words in such LIWC categories as social, family, humans, friend, and positive emotions. The *negative emotion* dimension had many words in the categories negative emotions, anger, affect, sad, anxiety, and death. The *embodiment* dimension had many words in the categories biology, body, ingest, health, and feeling. The *collection* dimension had words in the categories conjunction, inclusion, we, and they.

The bottom half of Table 2 shows a comparable analysis for the CCSS Exemplar texts. Lexile scores were not available in this analysis, but there are FK grade level scores, which correlate highly with Lexiles (.90 in TASA). The correlations of LIWC components with FK grade level scores were identical in sign and very similar in magnitude. Therefore, the results of the TASA corpus and the CC Exemplar text corpus were very compatible.

A LIWC formality composite score was computed from three of the dimensions. Higher LIWC formality scores were predicted to occur for texts with low narrativity, low processes-procedures-planning, and high collection PC scores. As shown in Table 1, this LIWC formality metric had high correlations with FK grade level (.600) and Lexile scores (.601), but they were less robust than Coh-Metrix formality scores (.716 and .664). The correlation between LIWC formality scores and Coh-Metrix formality scores was modest (.343), so the two tools were picking up some different aspects of formality. The bottom half of Table 2 replicates the above analysis for the CC Exemplar texts. The LIWC formality scores correlated .538 with FK grade level, which is lower than the .721 correlation between FK and Coh-Metrix formality for CC texts. The Coh-Metrix and LIWC formality scores again showed a modest .233 correction.

These results for LIWC formality show that a deep analysis of the linguistic and psychological characteristics of words can go a long way toward explaining text difficulty and uncovering the robust dimension of narrativity. However, Coh-Metrix formality goes a giant step further by being sensitive to sentence syntax and discourse cohesion.

Next Steps in Automated Analyses of Text Complexity

The text analyses in this article have conveyed the value of both unidimensional metrics of text difficulty and also multilevel component analyses. The unidimensional scales are provided by the six tools analyzed by Nelson et al. (2011) as well as FK grade level and now the formality scales presented in this article based on Coh-Metrix and LIWC. Most of the unidimensional text characteristics rely on word length, word frequency, and sentence length in the metric. These metrics are highly correlated and provide scales on grade level for informational texts and texts at grades K through 8. However, a multilevel analysis brings added value in providing a sensitive scale at higher grade levels and of narrative texts. The SourceRater and Word Maturity scales provided such multilevel components and helped remedy the limitations of unidimensional scales based on word length, word frequency, and sentence length. The Coh-Metrix-TEA measures of narrativity, sentence simplicity, word concreteness, referential cohesion, and deep cohesion also bring similar measures with added value. A practical advantage of the multilevel approach is that it provides more specific guidance on characteristics of texts that potentially give students problems.

The narrativity and syntax dimensions have consistently proven to be major predictors of text difficulty. Indeed, they have the highest correlations with the simple unidimensional text-difficulty scales (i.e., Lexiles, DRP, FK). In contrast, the cohesion dimensions and word-concreteness dimensions have had small or modest correlations with the simple unidimensional text-difficulty metrics. These unimpressive correlations begin to expose the blemishes of unidimensional metrics that rely on word frequency, word length, and sentence length. It is well established that reading times, memory, and comprehension for text are significantly influenced by referential cohesion, causal cohesion, and other types of cohesion at the situation-model level (Kintsch, 1998; McNamara et al., 2010; O'Reilly & McNamara, 2007; Zwaan & Radvansky, 1998). It is also well established in the cognitive literature that the abstractness-concreteness dimension has a robust impact on a wide array of cognitive processes, including comprehension (Mosenthal, 1996; Paivio, 1986). There is a fundamental limitation in unidimensional metrics if they are insensitive to cohesion and concreteness. Fortunately, the Coh-Metrix formality metric incorporates the cohesion and concreteness dimensions. Therefore, we argue that the formality unidimensional metric is superior to the metrics that rely on word familiarity, word length, and sentence length. The validity of this claim needs to be empirically tested in future projects that collect reading times, memory, comprehension, and other objective assessments.

Additional blemishes with the unidimensional difficulty metrics were exposed when we saw that syntactic simplicity and cohesion interacted with text genre in some interesting ways. When the text topic is difficult, as in the case of science texts, then writers make it easier on the reader by using (intentionally or unintentionally) simpler syntax and higher cohesion. Stated differently, simple syntax and high text cohesion may compensate for the difficulty of the topic. Such trade-offs bolster the value of the multilevel analysis of texts. Informational texts are intrinsically more difficult than narrative, but they tend to have less difficult syntax and higher referential cohesion. The five dimensions of Coh-Metrix do not swim together in ease or difficulty; hence they can detect some of these nuances in the difficulty of the subject matter.

These trade-offs have convinced us that it will be imperative in future research on text difficulty to consider the familiarity and complexity of the topics covered in the text. The need to consider topic difficulty has been explored in research on reading (Wixson, Peters, Weber, & Roeber, 1987) and listening comprehension for second-language learners (Schmidt-Rinehart, 1994). Emerging research in mathematics and the sciences about learning progressions is adding to the conversation about topic complexity. The goal of learning-progressions research is to propose and validate developmental pathways where learners gain increasingly more complex kinds of knowledge. Much of this research eschews the use of language like topic difficulty or complexity in favor of "requisite knowledge" for understanding more complex kinds of knowledge (Battista, 2011; Johnson & Tymms, 2011). The notion of learning progressions poses the idea that texts can change in complexity, both within a text and across texts. Correspondingly, the experiences of learners' reading from less to more complex topics also changes. Current research about text complexity tells us very little about what that is like (reading from simpler to more complex topics and back again), whereas the science and mathematics education research communities are attempting to figure this out. Researchers who focus on text complexity need to join this challenge.

We are now at a point in the history of text analysis when topic ontologies can be automatically derived from large text corpora through statistical techniques in computational linguistics, cognitive science, computational semantics, and machine learning (Jurafsky & Martin, 2008; McNamara, 2011). Automatically derived topics can also be automatically scaled on novelty, rarity, familiarity, links to other topics, and similarity to other topics. It may be difficult to scale topics on inherent complexity or changes in complexity, but the possibility of this is well worth exploring in the future. When the topics are exceptionally difficult, some compensatory tactics are to write texts with a simpler syntax and to help link sentences through referential cohesion and connectives. Such trade-offs would never be captured by standard unidimensional readability formulas, whereas we have shown that such trade-offs can be detected with the multilevel theoretical framework.

Another important direction for future research is to validate alternative automated measures of text complexity on psychological data. There are a variety of cognitive measures that can be collected, such as ratings of text difficulty, text reading times, text recall, think-aloud protocols, summarization, and psychometrically validated test scores (Sabatini & Albro, 2013). The emotions and affective states of readers are expected to be influenced by different dimensions of text difficulty (Graesser & D'Mello, 2012). For example, a reader may become bored, confused, or frustrated when the text is far too difficult for the reader to handle. The reader may tune out and the mind wander when text complexity is not aligned with his or her zone of proximal development (Feng, D'Mello, & Graesser, 2013). The impact of text difficulty on the psychological experience of the reader is not confined to cognition, but also stretches into realms of emotion and motivation.

Note

This research was supported by the National Science Foundation (BCS 0904909, DRK-12-0918409. REESE 1108845), the Institute of Education Sciences (R305G020018, R305A080589, R305C120001) and the Gates Foundation (Student Achievement Partners). Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of these funding agencies. We thank Jack Stenner at MetaMetrix for supplying the Lexile scores on the TASA corpus.

References

Battista, M. (2011). Conceptualizations and issues related to learning progressions, learning trajectories and levels of sophistication. *Mathematics Ethusiast*, **8**, 507–570.

Biber, D. (1988). *Variation across speech and writing*. Cambridge: Cambridge University Press.

Bussmann, H. (1996). *Routledge dictionary of language and linguistics*. London: Routledge.

Cain, K. (2010). *Reading development and difficulties*. Oxford: Wiley-Blackwell.

Chafe, W. L. (1982). Integration and involvement in speaking, writing, and oral literature. In D. Tannen (Ed.), *Spoken and written language: Exploring orality and literacy*. Norwood, NJ: Ablex.

Clark, H. H. (1996). *Using language*. Cambridge: Cambridge University Press.

Common Core State Standards Initiative. (2010). *Common Core State Standards for English language arts & literacy in history/social studies, science, and technical subjects*. Washington, DC: CCSSO & National Governors Association. Retrieved from http://www.corestandards.org/.

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

Deane, P., Sheehan, K. M., Sabatini, J., Futagi, Y., & Kostin, I. (2006). Differences in text structure and its implications for assessment of struggling readers. *Scientific Studies of Reading*, **10**, 257–275.

Eason, S. H., Goldberg, L. F., Young, K. M., Geist, M. C., & Cutting, L. E. (2012). Reader-text interactions: How differential text and question types influence cognitive skills needed for reading comprehension. *Journal of Educational Psychology*, **104**, 515.

Feng, S., D'Mello, S. K., & Graesser, A. (2013). Mind wandering while reading easy and difficult texts. *Psychonomic Bulletin & Review*, **20**, 586–592.

Goldman, S. R., Braasch, J. L. G., Wiley, J., Graesser, A. C., & Brodowinska, K. (2012). Comprehending and learning from internet sources: Processing patterns of better and poorer learners. *Reading Research Quarterly*, **47**, 356–381.

Graesser, A. C., & D'Mello, S. K. (2012). Moment-to-moment emotions during reading. *Reading Teacher*, **66**, 238–242.

Graesser, A. C., Hauft-Smith, K., Cohen, A. D., & Pyles, L. D. (1980). Advanced outlines, familiarity, text genre, and retention of prose. *Journal of Experimental Education*, **48**, 209–220.

Graesser, A. C., & McNamara, D. S. (2011). Computational analyses of multilevel discourse comprehension. *Topics in Cognitive Science*, **3**, 371–398.

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

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

Graesser, A. C., Millis, K. K., & Zwaan, R. A. (1997). Discourse comprehension. *Annual Review of Psychology*, **48**, 163–189.

Graesser, A. C., Singer, M., & Trabasso, T. (1994). Constructing inferences during narrative text comprehension. *Psychological Review*, **101**, 371–395.

Grimshaw, A. D. (2003). Genre, register, and contexts of discourse. In A. C. Graesser, M. A. Gernsbacher, & S. R. Goldman (Eds.), *Handbook of discourse processes* (pp. 25–82). Mahwah, NJ: Erlbaum.

Haberlandt, K. F., & Graesser, A. C. (1985). Component processes in text comprehension and some of their interactions. *Journal of Experimental Psychology: General*, **114**, 357–374.

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

Heylighen, F., & Dewaele, J. M. (2002). Variation in the contextuality of language: An empirical measure. *Foundations of Science*, **7**, 293–340.
Hiebert, E. H., & Fisher, C. W. (2007). The critical word factor in texts for beginning readers. *Journal of Educational Research*, **101**, 3–11.
Hiebert, E. H., & Mesmer, H. A. E. (2013). Upping the ante of text complexity in the Common Core State Standards: Examining its potential impact on young readers. *Educational Researcher*, **42**, 44–51.
Hymes, D. H. (1974). *Foundations in sociolinguistics: An ethnographic approach* (Vol. **65**). Philadelphia: University of Pennsylvania Press.
Johnson, P., & Tymms, P. (2011). The emergence of a learning progression in middle school chemistry. *Journal of Research in Science Teaching*, **48**(8), 849–877.
Jurafsky, D., & Martin, J. (2008). *Speech and language processing*. Upper Saddle River, NJ: Prentice-Hall.
Just, M. A., & Carpenter, P. A. (1987). *The psychology of reading and language comprehension*. Boston: Allyn & Bacon.
Kendeou, P., van den Broek, P., White, M. J., & Lynch, J. S. (2009). Predicting reading comprehension in early elementary school: The independent contributions of oral language and decoding skills. *Journal of Educational Psychology*, **101**(4), 765–778.
Kintsch, W. (1998). *Comprehension: A paradigm for cognition*. Cambridge: Cambridge University Press.
Klare, G. R. (1974–1975). Assessing readability. *Reading Research Quarterly*, **10**, 62–102.
Koslin, B. I., Zeno, S., & Koslin, S. (1987). *The DRP: An effective measure in reading*. New York: College Entrance Examination Board.
Labov, W. (1972). *Sociolinguistic patterns* (Vol. **4**). Philadelphia: University of Pennsylvania Press.
Lahiri, S., Mitra, P., & Lu, X. (2011). Informality judgment at sentence level and experiments with formality score. *CICLING*, **2**, 446–457.
Li, H., Graesser, A. C., & Cai, Z. (2013). Comparing two measures of formality. In C. Boonthum-Denecke & G. M. Youngblood (Eds.), *Proceedings of the 26th International Florida Artificial Intelligence Research Society Conference* (pp. 220–225). Palo Alto, CA: AAAI Press.
McNamara, D. S. (Ed.). (2007). *Reading comprehension strategies: Theory, interventions, and technologies*. Mahwah, NJ: Erlbaum.
McNamara, D. S. (2011). Computational methods to extract meaning from text and advance theories of human cognition. *Topics in Cognitive Science*, **2**, 1–15.
McNamara, D. S., Graesser, A. C., & Louwerse, M. M. (2013). Sources of text difficulty: Across the ages and genres. In J. P. Sabatini & E. Albro (Eds.), *Assessing reading in the 21st century: Aligning and applying advances in the reading and measurement sciences*. Lanham, MD: R&L Education.
McNamara, D. S., Graesser, A. C., McCarthy, P., & Cai, Z. (2014). *Automated evaluation of text and discourse with Coh-Metrix*. Cambridge: Cambridge University Press.
McNamara, D. S., & Kintsch, W. (1996). Learning from text: Effects of prior knowledge and text coherence. *Discourse Processes*, **22**, 247–287.
McNamara, D. S., Louwerse, M. M., McCarthy, P. M., & Graesser, A. C. (2010). Coh-Metrix: Capturing linguistic features of cohesion. *Discourse Processes*, **47**, 292–330.
Mosenthal, P. (1996). Understanding the strategies of document literacy and their conditions of use. *Journal of Educational Psychology*, **88**, 314–332.
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.
Olson, D. R. (1977). From utterance to text: The bias of language in speech and writing. *Harvard Educational Review*, **47**, 257–281.
O'Reilly, T., & McNamara, D. S. (2007). The impact of science knowledge, reading skill, and reading strategy knowledge on more traditional "high-stakes" measures of high school students' science achievement. *American Educational Research Journal*, **44**, 161–197.
Paivio, A. (1986). *Mental representations: A dual coding approach*. Oxford: Oxford University Press.
Pearson, P. D., & Hiebert, E. H. (2010). National reports in literacy: Building a scientific base for practice and policy. *Educational Researcher*, **39**, 286–294.
Pennebaker, J. W. (2011). *The secret life of pronouns: How our words reflect who we are*. New York: Bloomsbury.

Pennebaker, J. W., Booth, R. J., & Francis, M. E. (2007). *LIWC2007: Linguistic inquiry and word count*. Austin, TX: LIWC.net.

Perfetti, C. A. (1999). Comprehending written language: A blueprint of the reader. In C. M. Brown & P. Hagoort (Eds.), *The neurocognition of language* (pp. 167–208). Oxford: Oxford University Press.

Perfetti, C. A. (2007). Reading ability: Lexical quality to comprehension. *Scientific Studies of Reading*, **11**, 357–383.

Rapp, D. N., van den Broek, P., McMaster, K. L., Kendeou, P., & Espin, C. A. (2007). Higher- order comprehension processes in struggling readers: A perspective for research and intervention. *Scientific Studies of Reading*, **11**, 289–312.

Rayner, K., Foorman, B., Perfetti, C., Pesetsky, D., & Seidenberg, M. (2001). How psychological science informs the teaching of reading. *Psychological Science in the Public Interest*, **2**, 31–74.

Richards, J., Platt, J., & Platt, H. (1997). *Dictionary of language teaching and applied linguistics*. London: Longman.

Rouet, J. (2006). *The skills of document use: From text comprehension to web-based learning*. Mahwah, NJ: Erlbaum.

Sabatini, J. P., & Albro, E. (Eds.). (2013). *Assessing reading in the 21st century: Aligning and applying advances in the reading and measurement sciences*. Lanham, MD: R&L Education.

Schmidt-Rinehart, B. (1994). The effects of topic familiarity on second language listening comprehension. *Modern Language Journal*, **77**, 179–189.

Snow, C. (2002). *Reading for understanding: Toward an R&D program in reading comprehension*. Santa Monica, CA: RAND.

Stenner, A. J. (2006). *Measuring reading comprehension with the Lexile framework*. Durham, NC: Metametrics, Inc. Retrieved from http://files.eric.ed.gov/fulltext/ED435977.pdf.

van den Broek, P., White, M. J., Kendeou, P., & Carlson, S. (2009). Reading between the lines: Developmental and individual differences in cognitive processes in reading comprehension. In R. K. Wagner, C. Schatschneider, & C. Phythian-Sence (Eds.), *Beyond decoding. The behavioral and biological foundations of reading comprehension* (pp. 107–123). New York: Guilford.

van Dijk, T. A., & Kintsch, W. (1983). *Strategies of discourse comprehension*. New York: Academic Press.

Wiley, J., Goldman, S. R., Graesser, A. C., Sanchez, C. A., Ash, I. K., & Hemmerich, J. A. (2009). Source evaluation, comprehension, and learning in internet science inquiry tasks. *American Educational Research Journal*, **27**, 255–265

Williams, J. P., Stafford, K. B., Lauer, K. D., Hall, K. M., & Pollini, S. (2009). Embedding reading comprehension training in content-area instruction. *Journal of Educational Psychology*, **101**, 1–20.

Wixson, K. K., Peters, C. W., Weber, E. M., & Roeber, E. D. (1987). The state of assessment in reading. *Reading Teacher*, **40**, 749–754.

Zwaan, R. A., & Radvansky, G. A. (1998). Situation models in language comprehension and memory. *Psychological Bulletin*, **123**, 162–185.