COH-METRIX:
AUTOMATED COHESION AND COHERENCE SCORES TO PREDICT TEXT READABILITY AND FACILITATE COMPREHENSION

Project Funded by the
Office of Educational Research and Improvement
Reading Program

Danielle S. McNamara
Max M. Louwerse
Arthur C. Graesser

University of Memphis

September 2002 – August 2005
$1,425,000

COH-METRIX: AUTOMATED COHESION AND COHERENCE SCORES TO PREDICT TEXT READABILITY AND FACILITATE COMPREHENSION

Introduction

One of the salient missions of OERI is to improve students' ability to comprehend and learn from text. High school graduates are facing increasing expectations to comprehend complex texts in a world that is growing increasingly diverse and sophisticated. However, comprehension proficiency is not improving in the United States, and students are falling behind those in other countries (National Assessment of Educational Progress, 1998). The recent RAND report on *Reading for Understanding* (Snow, 2002) documents the pressing need to improve reading comprehension and presents a research agenda that embraces scientific research on comprehension processes, comprehension instruction, teacher training, and methods of assessing comprehension.

The RAND report's heuristic for thinking about reading comprehension includes four highly interactive components: Characteristics of the reader, the text, the comprehension activities, and the sociocultural context. In this proposal, the lens will focus primarily on the text and how text characteristics interact with the reader. Our goal is to improve reading comprehension in classrooms by providing a means to improve textbook writing and to more appropriately match textbooks to the intended students. The current practice of selecting texts for students is far from adequate. The content of selected texts is typically relevant to the particular subject matters taught in courses, but frequently there are problems with the vocabulary, syntax, cohesion, and difficulty level of the texts (Beck et al., 1989). We believe that these problems stem partially from an over-reliance on readability formulas for scaling texts on difficulty and age level, even though it has been widely acknowledged for decades that readability formulas are limited (Davison & Kantor, 1982; Hill & Erwin, 1984). The popular readability formulas, namely the Flesch Reading Ease (Flesch, 1943) and the Flesch-Kincaid Grade Level (Kincaid & McDaniel, 1974) formulas rely primarily on word length (number of letters

or syllables) and sentence length to assess difficulty level (see Appendix A for a list of commonly used readability formulas). Consequently, textbook publishers pressured to target textbooks to a certain grade can lower textbooks' grade level estimates by reducing word and sentence lengths. This approach results in short, choppy, sentences with minimal cohesion. However, texts with shorter sentences paradoxically run the risk of being more difficult to comprehend, particularly for readers with low domain knowledge and low reading proficiency. Why? Because there are fewer linguistic cues of cohesion that specify how the sentences should be conceptually related.

The concepts of *cohesion* and *coherence* are central to our proposed research, so it is important to clarify their intended meaning. Both constructs represent how words, constituents, and ideas conveyed in a text are connected on particular levels of language, discourse, and world knowledge. In the case of cohesion, the connections are grounded in explicit linguistic elements (i.e., words, features, cues, signals, constituents) and their combinations. As in the case of all symbolic and semiotic systems, such elements are interpreted within a designated sociocultural context (i.e., the intended language and cultural community). There is an objective foundation in computing cohesion, which to a large extent can be extracted by computer programs. Coherence, however, results from an interaction between text cohesion and the reader. A particular level of cohesion may lead to a coherent mental representation for one reader, but an incoherent representation for another (McNamara et al., 1996). The connections within the reader's mental representation are constructed based on the elements available in the text combined with the reader's cognitive abilities and intentions. These connections we call coherence (Louwerse & Graesser, in press).

Two decades ago it would not have been feasible to systematically investigate cohesion and coherence because of the intractability of measuring world knowledge and lack of language proficiency measures at multiple levels. However, recent advances in psycholinguistics (Gernsbacher, 1994), discourse processing (Graesser, Gernsbacher, & Goldman, in press; Kintsch,

1998), corpus linguistics (Biber, 1988; Biber et al., 1998), and computational linguistics (Allen, 1995; Jurafsky & Martin, 2000) have led to the development of sophisticated computer tools that analyze text on various dimensions. There now exist computer programs with large lexicons, syntactic parsers, semantic analyzers, essay graders, and summary critiquers. For example, latent semantic analysis (LSA) has provided a statistical representation of world knowledge that surprisingly captures judgments and intuitions of humans (Landauer & Dumais, 1997; Landauer et al., 1998; Graesser, P. Wiemer-Hastings et al., 2000; Millis et al., 2001; Shapiro & McNamara, 2000). Numerous lexicons have been developed that provide a wealth of linguistic information: e.g., WordNET (Fellbaum, 1998), FrameNet (Fillmore & Baker, 2001), CELEX (Burnage, 1990), Comlex (MacCleod et al., 1998), and MRC (Coltheart, 1981). Syntactic parsers have been implemented in numerous computational linguistic projects, ranging from speech recognition (Rosé & Lavie, 1998) to evolutionary linguistics (Cangelosi & Parisi, 2002). Parsers extract and detect a wide range of linguistic information, such as parts of speech, ambiguity, syntactic complexity, keywords, and speech acts (Abney, 1997; Brill, 1995; Early, 1970; Sekine & Grishman, 1996).

The educational enterprise clearly needs a better system for grading text difficulty and the time is ripe for such a system to be created. Moreover, the investigators on this proposal are particularly suited to the task of developing a tool that analyzes texts more deeply and on more levels. Graesser and Louwerse have built a tool called QUAID (Question Understanding Aid), which critiques sentences and questions on multiple levels: unfamiliar technical terms, vague or imprecise relative terms, vague or ambiguous noun-phrases, syntactic complexity, and working memory overload (Graesser, K. Wiemer-Hastings et al., 2000; Graesser, Karnavat et al., 2001). Graesser and Louwerse have also developed and tested an intelligent tutoring system called AutoTutor, where students learn about computer literacy and conceptual physics by holding a dialog with the computer in natural language (Graesser, VanLehn et al., 2001; Graesser, Person et al., in press). AutoTutor analyzes the

quality of student contributions in the dialog (words, sentences, or several sentences) and the extent to which they are coherently related to a curriculum script (the right answer) and the discourse topic. McNamara and her colleagues (Millis et al., 2001; Levinstein et al., 2002) are currently developing the Automated Self-Explanation Reading Strategy Trainer, an interactive program to train learners to use comprehension strategies, such as building explanations of the reading material and connecting incoming clauses in the text to previous text information. The automated strategy trainer is able to judge the quality of a reader's explanation of text and provide appropriate feedback.

In addition to expertise regarding automated text analysis, the principle investigators are also leading researchers of comprehension processes. Graesser, widely published in the reading comprehension literature, is also editor of *Discourse Processes* and the *Handbook of Discourse Processes*. Louwerse and McNamara, leaders in the Society for Text and Discourse, specialize in effects of cohesion and coherence (Louwerse, 2002a, 2002b; Louwerse & Graesser, in press; McNamara, 2001; McNamara et al., 1996; McNamara & Kintsch, 1996;). The project team possesses multidisciplinary expertise in psychology, linguistics, education, literary theory, cognitive science, mathematics, and artificial intelligence. This interdisciplinary effort results in a unique combination of skills, scholarship, and accomplishments that will lead to the successful completion of the project goals.

Project Goals

The purpose of this project is to create a computer tool that measures text difficulty at various levels of language, discourse, and conceptual analysis. Text difficulties include problems with vocabulary, syntactic composition, meaning, and cohesion. We already have developed computational modules that tap multiple lexicons and syntactic parsers (in QUAID and AutoTutor); these modules are immediately available for analyzing a wide range of elements and constituents within sentences. Our primary challenge in the proposed grant is to analyze cohesion and coherence.

Our first goal is to formulate and test metrics that measure text coherence based on properties of cohesion and reader characteristics (e.g., world knowledge, domain knowledge, and reading ability). We call the tool that will calculate the coherence metrics *Coh-Metrix*. This tool will be used to (a) assess the overall cohesion of texts, thereby enhancing current readability measures, (b) provide scores regarding various cohesion characteristics, and (c) determine the coherence or appropriateness of a text for a reader. We will also develop a tool to identify the specific locations and types of cohesion gaps in text, which we call the Cohesion Gap Identification Tool, or *Coh-GIT*.

Our second goal is to further investigate effects of text cohesion. This project is motivated by research establishing that text cohesion is a driving factor of comprehension, and moreover, that text characteristics critically interact with readers' abilities (McNamara, 2001; McNamara et al., 1996; McNamara & Kintsch, 1996). However, there remain gaps in our understanding of these effects. Therefore, to develop and calibrate our tools, we need to conduct further empirical research. Due to time and budget limitations, we have chosen to concentrate on two populations, young readers and college students. Our first choice is driven by a lack of research concerning the effects of text cohesion on young readers' comprehension. A critical period occurs between the third and fifth grades, with comprehension difficulties often emerging within the fourth grade (Chall et al., 1990; Meichenbaum & Biemiller, 1998; Snow, 2002). Hence, we will focus our attention on those grade levels. Specifically, this research will address the following questions: (a) What are the effects of text cohesion on comprehension for young readers, and how do those effects depend on text genre and word-level reading skills (i.e., word recognition and letter-sound decoding)?; (b) What are the relations between reading comprehension skills demonstrated by young children when reading high and low coherence texts; and what are inter-individual differences in cognitive aptitudes and word-level reading skills? (c) What are the intra-individual differences in reading skills demonstrated by young readers who read low-cohesion texts with high and low levels of comprehension?

The choice of our second population of college students is motivated partially by availability, and also by a number of questions relevant to general populations that must be answered to develop our tools. Our goal is to further our understanding of text and reader aptitude interactions, thus providing a means to better calibrate *Coh-Metrix* and *Coh-GIT*. Specifically, we will address the following questions: (a) How do the effects of text cohesion interact with prior knowledge, reading ability, reading strategy knowledge, and reading motivation; and (b) How do the effects of text cohesion vary across text genres (i.e., narrative, historical, scientific) and reader aptitude levels?

A third goal is to fine-tune and validate our coherence metrics. We have three approaches to achieve this goal. First, we will use existing text corpora and data collected during this project to calibrate and test *Coh-Metrix*. Second, we will examine and compare cohesion metrics and readability scores across a large set of K-14 basal readers and instructional textbooks. Our third approach is to use eye-tracking technology to empirically verify *Coh-GIT*'s ability to identify cohesion gaps.

Theoretical Framework

We directly adopt the RAND panel's heuristic that comprehension can be viewed as having four highly interactive components, namely the text, reader, comprehension activities, and sociocultural context. Indeed, empirical studies of reading comprehension have uncovered some intriguing, and sometimes counterintuitive, interactions among text, reader, and task variables (Coté et al., 1998; Graesser, Kassler et al., 1998; McNamara et al., 1996). For example, McNamara et al. (1996) manipulated text cohesion (high versus low), measured readers' knowledge about passage topics (high versus low), and administered several tests that tapped different levels of mental representation. High cohesion texts benefited readers with low knowledge regardless of the type of comprehension test (to no one's surprise). However, when tests of deep comprehension were analyzed, high-knowledge readers showed substantial benefits from having read texts with low cohesion (to the surprise of many). The low cohesion texts apparently encouraged the

knowledgeable readers to work harder and actively build more elaborate mental representations. There emerges the challenge of explaining a three-way interaction among the text, the reader, and the tasks at test. The destiny of comprehension researchers lies in discovering and explaining such higher order interactions, not merely in confirming simple obvious generalizations.

Our theoretical framework emphasizes the importance of dissecting the multiple levels of cognitive representation that get activated, constructed, and encoded during comprehension. Most discourse psychologists adopt Kintsch's distinctions among the *surface code*, the propositional *textbase*, and the *situation model* (Kintsch, 1988, 1998). The *surface code* preserves the exact wording and syntax of clauses. Readers retain the surface code only briefly unless there are features of the surface code that have important repercussions on meaning. The *textbase* contains explicit propositions that preserve the meaning, but not the exact wording and syntax; the textbase also includes a small number of inferences needed to establish local text cohesion. The textbase is normally retained in memory for a few hours and fades out almost completely after a few days. The *situation model* (or mental model) is the content or microworld that the text is about. The situation model for an expository text refers to the reader's prior knowledge about the text's subject matter. The situation model for a story refers to the people, setting, actions, events, emotions, and various mental states of the people in the constructed microworld. This microworld is constructed inferentially through interactions between the explicit text, background world knowledge, and the comprehension goals of the reader (Graesser, Singer, & Trabasso, 1994; Johnson-Laird, 1983; Kintsch, 1994, 1998). The situation model has the slowest decay rate from memory, lasting days, weeks, or even years (Graesser, 1981; Kintsch et al., 1990). Aside from these three levels, discourse psychologists also acknowledge the importance of the text genre (i.e., the category of text, such as a technical scientific text versus a mystery novel) and the pragmatic communication between the writer and reader (Clark, 1996; Graesser, Millis, & Zwaan, 1997; Van Dijk & Kintsch, 1983).

The process of comprehension consists of activating, constructing, and encoding representations at these various levels. This is not necessarily a bottom-up process, but rather a process of constraint satisfaction (Kintsch, 1998; McClelland & Rumelhart, 1986). Comprehension is easy when the composition is well formed within each level and when there is harmony in the links/mappings between levels. However, there may be cohesion gaps or clashes either within and/or between levels, which put the attentive, knowledgeable reader in cognitive disequilibrium (Otero & Graesser, 2001). Attention, effort, and knowledge are needed to restore equilibrium. Metacognitive strategies monitor the detection and management of cognitive disequilibrium (Hacker et al., 1998). A diligent, strategic, high-knowledge reader will be successful and learn from the experience, whereas others often give up or settle for a more shallow understanding.

National Significance

Texts provide the core information backbone of our educational system so it is imperative that adequate attention is paid to the quality of texts and their appropriateness to particular populations of readers. It is widely acknowledged that the quality of readability formulas for texts is severely limited. Moreover, such formulas may paradoxically aggravate the problem by encouraging textbook writers to modify texts in mechanical ways that have unwanted side effects on comprehension. The proposed research is designed to solve this problem by developing a better computer tool for grading texts on several dimensions of language and discourse and for matching texts to particular classes of readers. Research has clearly indicated the importance of text cohesion and coherence, but there are no readily available methods for measuring either one of them. Textbook writers and editors cannot follow formulas to improve coherence because (a) there exist no automated valid formulas, and (b) increased coherence often decreases standard readability scores. Textbook writers will be given access to our scientific understanding of language, cohesion, and discourse coherence through an improved automated system of text evaluation.

Readability Measures

Readability measures are primarily based on factors such as the number of words in the sentences and the number of letters or syllables per word (i.e., as a reflection of word frequency). Two of the most commonly used measures are the Flesch Reading Ease formula and the Flesch-Kincaid Grade Level. The output of the Flesch Reading Ease formula is a number from 0 to 100, with a higher score indicating easier reading. The more common Flesch Grade Level formula converts the Reading Ease Score to a grade level. In addition, more than 40 readability formulas have been developed over the years (Klare, 1974-1975; Appendix A).

Readability measures guide the construction of textbooks such that the writing conforms to the intended grade level. However, there are at least three major problems with readability formulas that prevent valid predictions of text comprehension. First, readability scores are based on the surface characteristics of the text. Comprehension and learning, however, depend to a greater extent on processing at the textbase and situation levels (Kintsch et al., 1990; McNamara et al., 1996). Measuring text elements that are primarily needed for surface processing does not adequately capture comprehension and learning, which is the concern of educators. Recent advances in discourse processing and computational linguistics afford more advanced measures of readability due to more precise predictions of which text characteristics improve comprehension and learning. Second, predicting reading, understanding, and learning requires consideration of the reader's knowledge, language skills, and other cognitive aptitudes. Although text characteristics can certainly predict aspects of readability, readability should be viewed as an interaction between a text and a reader's cognitive aptitudes (Kintsch, 1994; McNamara et al. 1996; Miller & Kintsch, 1980).

Third, readability formulas cannot capture text cohesion or coherence. Research has clearly shown that readers have less difficulty reading cohesive texts (Beck et al., 1991; Britton & Gulgoz, 1991; Gernsbacher, 1997a; Graesser, Gernsbacher, & Goldman, in press; McNamara & Kintsch,

1996; McNamara et al., 1996). We would therefore expect greater readability scores for high-cohesion texts than low-cohesion texts; however, this is not generally the case. For example, *One part of the cloud develops a downdraft. Rain begins to fall.* has lower causal cohesion than *One part of the cloud develops a downdraft, which causes rain to fall.*, but a lower Flesch-Kincaid grade level (3.4 and 4.9 respectively). Similar patterns are found for passages with empirically documented comprehension effects. For instance, a high-cohesion text about cell mitosis in McNamara (2001) resulted in better comprehension but had a Flesch-Kincaid grade level of 11.2 compared to 9.3 for the low-cohesion version. Many more such examples are available, but the bottom line is that increasing cohesion often requires adding words. Longer sentences result in increased grade level predictions.

In addition, readability formulas fail to capture many other types of cohesive devices. For example, fewer pronouns can increase referential coherence. However, pronouns are shorter words; thus, their presence decreases readability grades level estimates. Referential coherence also increases with greater conceptual repetition, which readability formulas cannot capture. Indeed, as described in the following section, they cannot capture most, if any, forms of cohesive devices.

Cohesion and Coherence

As explained within the introduction, we use the term cohesion to refer to properties of the text with reference to the intended readers' sociocultural knowledge, and coherence to refer to the interaction between the text and reader characteristics. The literature distinguishes various kinds of cohesion and coherence. One common distinction is between ***local*** and ***global*** levels (Givón, 1993; Louwerse, 2002b; McNamara et al., 1996; Van Dijk & Kintsch, 1983). Both cohesion and coherence are locally and globally structured. The comprehender finds local cohesion relations between adjacent clauses in the text and global cohesion links between groups of clauses and groups of paragraphs. This distinction is important, because both local and global cohesion provide links cues to comprehenders on how to organize the comprehension process. Titles, headers, and topic

sentences in paragraphs often mark global cohesion. Anaphora, conceptual (or argument) overlap, and interclausal relationships mark local cohesion. Texts that are locally cohesive but lack global cohesion tend to inhibit comprehension and recall (Bransford & Johnson, 1972; McNamara et al., 1996). Similarly, texts that are globally cohesive but lack local cohesion are sometimes difficult to read and comprehend (Van Dijk, 1977; McNamara et al., 1996).

Another distinction can be made between ***grammar-driven*** and ***vocabulary-driven*** cohesion (Givón, 1995; Kintsch, 1995; Townsend & Bever, 2001). Grammar-driven cohesion primarily refers to information in the text that cues grammar-based inferences. Vocabulary-driven cohesion primarily refers to words that cue knowledge-based inferences. Consider the following sentences:

(a) John left *his office* and stepped *into the lounge where* Mary was waiting.

(b) John left *his office. He entered the lounge.* Mary was waiting *in the corner.*

(c) John left his office *opposite* the lounge. He walked *inside. There*, Mary was waiting.

In sentence (a) grammatical cues like *his*, *into* and *where* help to establish spatial coherence, which is vocabulary-driven by *his office* and *lounge*. In sentence (b) the grammar-driven cohesion devices have been left out and are replaced by lexical markers. In (c) the focus lies on grammar-driven cohesion devices, like *opposite*, *inside*, and *there*.

Finally, the following conceptual categories of cohesion and coherence can be distinguished: ***referential, temporal, locational, causal,*** and ***structural*** (Givón, 1995; Louwerse, 2002b). These categories answer the who, where, why, when and what of the events described by the text, thus facilitating processing (Gernsbacher, 1990; Graesser, Wiemer-Hastings, P. & Wiemer-Hastings, K., 2001; Zwaan & Radvansky, 1998). ***Referential*** coherence is established with the use of anaphora, repeated phrases, definite articles, and conceptual overlap (Gernsbacher & Robertson, 2002; Haviland & Clark, 1974). ***Temporal*** coherence refers to continuity in time, which is established by connectives (*before, and then*), prepositional phrases (*Later on that day*), verb tense and aspect, or with

order of mention (Anderson et al., 1983; Magliano & Schleich, 2000; Ohtsuka & Brewer, 1992; Zwaan, 1996). ***Locational*** cohesion is often cued by adverbs (*here, there*), adverbial phrases, prepositions (*above, near*), and verbs that reflect the narrator's point of view (*come* versus *go*) (Black et al., 1979; De Vega, 1995). ***Causal*** coherence is frequently established by marking the causal relations between two events with connectives (*because, enable, so that*) (Keenan et al., 1984; Myers et al., 1987; Singer et al., 1992; Van den Broek, 1994). Finally, ***structural*** coherence refers to the continuity in syntactic and conceptual form of clauses (Strunk & White, 1972; Gernsbacher, 1997b).

Cohesion markers exist within each permutation of these categories. For example, referential cohesion can be identified at the local and global levels, both with grammatical and semantic information. These cohesion markers can be captured computationally with a combination of syntactic parsing techniques, lexicons, and other CL modules. For example, modules such as LSA can be used to assess referential cohesion both globally and locally by computing degree of conceptual overlap between sentences, paragraphs, and the entire text. Similarly, topic sentences can be identified on the basis of the semantic relatedness with other sentences in a paragraph and with the whole document. Syntactic parsers can identify a large array of information such as the presence and adequacy of titles and headers, consistency in subject assignment between clauses, syntactic categories and their hierarchical structure, and verb aspect agreement. Lexicons can identify lexical and grammatical elements, as well as semantic relations between words. It is beyond the scope of this proposal to give an exhaustive account of these mechanisms. Nevertheless, these examples should illustrate that whereas it is impossible for non-computational readability formulas to capture syntactic and semantic aspects of text, this task is manageable when we make use of CL tools.

In sum, establishing coherence involves local and global processing of vocabulary-driven and grammar-driven cues in the text and monitoring the multiple referential, locational, temporal, causal, and structural relations (Taylor & Tversky, 1997; Zwaan et al., 1995). Current readability measures

cannot successfully assess these different characteristics of coherence because coherence is often not related to surface features like word or sentence length. Our metrics will satisfy this need with an assortment of cohesion and coherence metrics. A wealth of literature now elucidates how coherence is established during comprehension and how cohesion affects this process. With recent CL techniques allowing us to identify these cohesive cues, the fertilization between psycholinguistics, computational linguistics, education, and reading literacy can provide us with valuable information concerning text readability and comprehension.

PROPOSED RESEARCH

The following section describes the three goals of our project and our means to achieve these goals.

A. Develop *Coh-Metrix* and *Coh-GIT*.

Our first goal is to expand current readability measurements by developing an automated cohesion metric (*Coh-Metrix*) that considers syntactic and semantic aspects of texts at various levels. We will also develop an automated tool that identifies specific cohesion gaps in text (*Coh-GIT*). Both tools will be web-based to reach the largest group of potential users (e.g., educators, writers, readers). The software will rely on a variety of computer languages (JAVA, LISP, and C++) in Linux and Windows XP operating systems, whereas the hardware will be a configuration of Dell Pentium servers. We have documented the software and hardware components in our previous publications on QUAID and AutoTutor, so we will not elaborate on the technical details here.

We will build computer tools with a simple human-computer interface to maximize usage of the tools. The user will input a text to be scaled and set optional parameters, such as the amount of background knowledge in the reader and the language skills of the targeted reader. *Coh-Metrix* will then analyze the textual information, identify the various cohesion relations in the text, and compute a score for each of these relations. Based on reader characteristics, these cohesion scores are next

converted into coherence scores to determine the appropriateness of the text for the targeted reader. For instance, scores that are too low or too high would yield a low appropriateness score.

More specifically, the computer tool will automatically determine how text elements and constituents are connected for specific types of cohesion. Suppose there are 20 (2 x 2 x 5) types of cohesion (local and global, vocabulary- and grammar-driven, referential, locational, temporal, causal, and structural). Suppose that there are N elements and constituents in a particular text. There would be N x (N-1) directional cohesion connections and [N x (N-1)/2) bidirectional connections with respect to any one of the 20 cohesion relations. We can capture the resulting set of connections in the form of a *matrix* (Graesser, Karnavat et al., 2000; Kintsch, 1998; Zwaan et al., 1995). A cell in the matrix is 0 if there is no particular relation between a pair of elements/constituents, a 1 if a solid relation, and intermediate values if there is a non-discrete metric. The *full matrix* has the entire set of connections with respect to a type of relation R, whereas the *contiguous sub-matrix* includes only those elements/constituents that are contiguous in the explicit text. We can define a *density* measure as the summation of such cell values. For example, the *causal cohesion density* would be computed as: $[\Sigma\ R(e_i, e_j)]/[N \times (N-1)]$, which is the average cell value for pairs of cells with respect to the causal relation. One could restrict this to *a contiguous, bi-directional, causal cohesion density,* which only considers the contiguous elements and constituents, computed as $[\Sigma\ R(e_i, e_j \mid e_i \ \&\ e_j \text{ are contiguous})]/[N+1)]$. When integrating over all types of cohesion markers, one can compute an overall cohesion density score for a particular text. More importantly, however, these density scores would specify how much an entire text has cohesion markers with respect to a particular type of relation. For instance, it might be the case that different readers (low vs. high knowledge, low vs. high reading ability) rely more on one type of cohesion relation than another.

A fine-grained recall analysis can be used to test the validity of the coherence and cohesion metrics produced by *Coh-Metrix*. Suppose that a recall protocol is collected from a sample of subjects

on a text with N elements/constituents. Each cell in the matrix would have different types of recall measures. For example, a recall proportion is the proportion of observations in which both nodes e_i & e_j are present in a particular recall protocol, with values varying from 0 to 1. A multiple regression analysis can indicate whether the [N x (N-1)]/2 recall proportions are predicted by each of the cohesion matrices as predictor variables (i.e., one for referential, one for causal, etc.). There is a large number of cells in such analyses; if N = 40, then the number of cells is 780. Similarly, we can do this for contiguous recalls, for distance in recall order, for time measures, and so on.

This method would give us overall weights for cohesion relations, which can then be put in a formula to determine readability, comprehension, learning, and appropriateness scores. However, as we mentioned before, different readers might rely on different cohesion relations. Therefore, in addition to these overall scores, we want to include differences between readers. If the user has information on the knowledge of the reader (high – low) or reading ability (high – low), this information can be entered in the computational model to tailor the formula to particular groups of readers, thus increasing the accuracy of the scores.

In addition to *Coh-Metrix*, we will develop *Coh-GIT* to analyze *where* the relevant cohesion relations are located in the text. This way, writers and educators will be able to not only predict the readability, comprehensibility, learnability, and appropriateness of a text for a particular reader group but also improve the problematic aspects of that text.

B. Empirically examine interactive effects of text cohesion and reader aptitudes

As explained earlier, our empirical research will focus on two populations, young readers and college students. All of the proposed experiments include aptitude measures as quasi-experimental variables (Shaddish et al., 2001). This begs the question of how analyses will be conducted. One method is to categorize participants as high or low for each variable in question and conduct ANOVAs. However, when considering multiple, correlated aptitude measures, it is often difficult to

fill certain cells (e.g., high on two variables, but low on a third). Also, this procedure tends to reduce power. Including aptitude measures as covariates is not generally an option in our case because the majority of our hypotheses regard interactions between aptitude and experimental variables. An alternative is to conduct ANOVAs using multiple regression (Judd & McClelland, 1989). Multiple regression analyses allow us to examine effects of categorical experimental factors and continuous measures of reader aptitude simultaneously, and thus the ability to examine potential interactions between experimental and individual difference factors. Our approach will be comprehensive. For all proposed experiments, we adopt the latter approach (multiple regression) but will conduct follow up ANOVAs (and/or MANOVAs) to further clarify the results. It is essential that the results using categorical variables mimic those using continuous variables because our tool will make use of categorical aptitude judgments. For all experiments proposed (except Experiment 3), the independent quasi-experimental variables include the reader aptitudes assessed. In terms of design, aptitude variables for which we have explicit hypotheses are implicitly treated as between-subject variables, and thus the number of subjects is adjusted accordingly. Random assignment will be used for all experimental variables.

1. Experimental studies with third-grade readers

Experiment 1: This experiment examines the effect of text cohesion for young readers and how those effects depend on text genre and reading skill. Third-grade children (n=60) will read two expository and two narrative passages, including one high-cohesion version and one low-cohesion version for each text genre. The low-cohesion version of the passages will have a Flesch-Kincaid grade level score between 1.5 and 2.5, whereas the high-cohesion version will have a grade level score between 3.5 and 4.5, with a difference of at least 1.5 grade levels between low-and high-cohesion versions. For each passage, comprehension will be assessed with: (a) oral recall, (b) four orally presented bridging-inference questions that require linking separate ideas from the text to

answer correctly, and (c) 12 orally presented True/False questions concerning the passages. Children will also complete the Letter-Word Identification, Word Attack, and Picture Vocabulary tests from the Woodcock-Johnson III (WJ III; Woodcock et al., 2001; see Appendix B). The first two of these tests form a Basic Reading Skills cluster score, whereas the third is a measure of word knowledge. The independent experimental variables will include the within-subject variables of passage coherence (i.e., high/low) and text genre (i.e., narrative/expository). The dependent measures (i.e., recall and question accuracy) will be analyzed separately and as an aggregate score.

We expect children to better comprehend the high-cohesion than the low-cohesion versions of the passages (despite the increase in Flesch-Kincaid grade-level difficulty), and that the cohesion modifications will have the greatest benefits for expository passages and for less-skilled readers. We expect little effect of passage cohesion for skilled readers' comprehension of the narrative passages. We also predict that these results will be most pronounced for assessments of deeper comprehension (i.e., bridging inference questions). This pattern of results will confirm our hypothesis that traditional readability statistics are most misleading in the most critical circumstances, that is, when struggling young readers attempt to learn new information from text.

Experiment 2: To examine the myriad of inter-individual differences in reading aptitudes identified by Snow (2001) and others (Evans et al., in press; Morris et al., 1998; Scarborough, 1998; Vellutino et al., 2000; Windfuhr & Snowling, 2001), a battery of cognitive ability and reading tests will be completed by students in grades 3, 4, and 5 (n=50 in each grade level) to predict reading comprehension of high- and low-cohesion expository passages. Participants will read four expository passages that vary linearly in terms of coherence. Comprehension of the passages (i.e., the DV) and reading skills will be assessed as in Experiment 1. The children will also complete the following six WJ III tests: Verbal Comprehension, Visual-Auditory Learning, Numbers Reversed,

Sound Blending, Visual Matching, and Rapid Picture Naming. The aptitude tests will require approximately 50 minutes to administer.

One purpose of this study is to provide information concerning which aptitudes are most important for calibrating *Coh-Metrix*. We will conduct exploratory analyses to examine which of the measures are most predictive of comprehension and which of them interact with text cohesion. That is, it is important for us to determine what reader characteristics must be considered to calculate text coherence. In addition to examining the relative contributions of reading skills and cognitive abilities, we will also examine the relative effects of various types of cohesion (as measured by *Coh-Metrix*) and how each interacts with reader characteristics.

Experiment 3: To compare intra-individual differences in reading skills that may influence comprehension of low-cohesion text, we will analyze reading skill profiles of (a) children who are considered skilled comprehenders of low-cohesion text and (b) children who are considered poor comprehenders of low-cohesion text. Fourth-grade children (n=80) will read two low-cohesion passages. Based on their comprehension of these texts, they will be categorized as skilled or less-skilled comprehenders of low-cohesion text. [Note that here, comprehension of the passages constitutes an IV, not a DV.] To address a variety of reading skills measured using varying test paradigms, children will complete the Letter-Word Identification, Word Attack, Passage Comprehension, Reading Vocabulary, and Reading Fluency tests from the WJ III (Woodcock et al., 2001) and the Reading Comprehension subtest from the Wechsler Individual Achievement Tests, Second Edition (Wechsler, 2001; see Appendix B). These nationally standardized and normed instruments measure word identification, letter-sound decoding, reading vocabulary skills, reading fluency, and reading comprehension (using both modified cloze and oral open-ended response formats). [Note that in contrast to Exp. 1 & 2, performance on these measures comprises the DVs.]

Profile analyses will determine shape, level, and dispersion of reading skills for the two groups of children. Analyses will indicate whether the groups have parallel profiles (i.e., shape), whether overall differences exist between groups (i.e., level), and whether groups perform similarly on all reading skills (i.e., dispersion). In order to identify sources of variability, contrasts following the profile analysis will be conducted to investigate differences in scores between the two groups on the reading skills measures.

2. Experimental studies with college students

Experiments 4 and 5. Our first set of experiments will further explore effects of text cohesion to examine possible interactions of domain (science) knowledge, general knowledge (literature, arts, history), reading ability (including reading strategy knowledge), and motivation (see Appendix C for descriptions of the aptitude measures that will be administered to the participants). Although separate studies have investigated these factors, there have been no studies to examine them simultaneously with respect to effects of text cohesion. These data will answer the theoretical question of whether these abilities can compensate for one another, and thus negate the negative impacts of coherence gaps in texts. These studies will also help us to better calibrate *Coh-Metrix* according to reader aptitudes.

In **Experiment 4**, 100 college students will read two science passages, including one high-cohesion and one low-cohesion version (≅400-word passages). In **Experiment 5**, 200 college students will read either the high- or low-cohesion science passage and readers' motivational factors (i.e., goals) will be manipulated with instructions to read the passage to either (a) prepare to answer essay questions about the passage, or (b) prepare for multiple-choice questions about the content. After reading the passages, the participants will answer open-ended essay questions of varying difficulty (i.e., the dependent measure is proportion correct). Half of the questions will assess comprehension of single sentences within the passage (i.e., text based), and half of the questions will

assess comprehension of relationships between separate sentences (i.e., bridging inference). Within-subject experimental variables include text cohesion and question type.

As found in previous research, we expect that low-cohesion texts will inhibit low-knowledge readers' comprehension but yield benefits for high-knowledge readers. However, greater reading skills should help low-knowledge readers overcome negative effects of cohesion gaps. We further expect greater reading motivation/interest to minimize cohesion effects for high-knowledge readers. Specifically, we expect high-knowledge readers to perform well on the comprehension questions regardless of text cohesion if (a) in Experiment 4, they show a greater interest in science and greater motivation, and (b) in Experiment 5, they are asked to read to prepare for essay questions rather than multiple-choice questions. These results will further our theoretical understanding of reading comprehension and provide important information for the development of our cohesion metrics.

Experiment 6. This experiment examines cohesion effects across a variety of text genres. Here, we will use natural texts rather than experimenter-manipulated passages, with cohesion verified using an overall *Coh-Metrix* score. Participants (n=300) will read 12 passages, including four 350-word passages from three text genres (narrative, historical, and scientific), which will vary linearly in terms of overall cohesion. Thus, across text genres, there will be three passages of low, medium-low, medium-high, and high coherence. The passages will vary in readability scores from 9th to 12th grade levels such that readability scores and cohesion scores are not correlated. Passages will be presented on a computer screen in a cumulative sentence-by-sentence technique such that each sentence appears one at a time, with previous sentences remaining visible. Participants will read the passages and type their recall of the text either after each one (immediate recall) or after reading all of the passages (delayed recall). Participants' knowledge, reading ability, and motivation will be assessed. Independent experimental variables include the within-subjects factors of text genre and passage coherence, and the between-subjects factor of recall delay.

We hypothesize that passage recall (e.g., number of idea units recalled) and reading time will vary as a function of cohesion (and not readability) for each of the three text genres. However, cohesion effects should be more pronounced for expository (historical, scientific) texts than for narrative texts because the former contain fewer familiar concepts. We further expect that the effects of cohesion will vary as a function of readers' knowledge and reading ability. However, cohesion effects should have the greatest interdependence with knowledge for expository texts but depend more on reading ability for narrative texts. Specifically, we expect that cohesion for expository texts will have a large effect for low-knowledge participants, compared to negligible effects for high-knowledge readers. These results should be more pronounced with delayed than immediate recall. Such a pattern of results would support our assumption that coherence measures are particularly important for instructional textbooks – for which the majority of the audience possesses a relatively low amount of knowledge within the target domain.

The recall protocols will also be used to assess the different categories of coherence at a more fine-grained level, as described earlier. A recall coocurrence matrix will be prepared for each subject and for groups of subjects with particular reader profiles. The likelihood that a pair of text elements are recalled together will serve as the dependent variable, whereas the cell values of each coherence category will serve as predictor variables. As discussed earlier, if there are 40 text elements, there are 780 pairs of elements that serve as units of analysis. Multiple regression analyses can be used to assess the unique contributions of dozens of coherence measures.

C. Fine-tune and establish validity of *Coh-Metrix* and *Coh-GIT*

Calibrating and Testing *Coh-Metrix*. To achieve this goal, we will take advantage of the extensive amount of data available from previous studies of comprehension. A large amount of data is readily available to us from our own previous research and we will also request published data samples from colleagues. We will use half of our available data sets (including those data sets

described in Research Goal B) to calibrate the cohesion metrics. We will use the other half to examine the metrics' ability to predict comprehension and to determine which specific metrics account for greater variance as a function of the reading situation. Our goal is to determine the most parsimonious measures and optimal weights for *Coh-Metrix*. At the same time, we will compare the predictive ability of *Coh-Metrix* to other methods of scoring readability such as the Flesch-Kincaid and to LSA. Our overarching goal is to create an automated cohesion metric that predicts comprehension more accurately than other readability scores.

Compare readability and cohesion metrics. Our second approach is to examine and compare readability scores and cohesion metrics for an extensive sample of basal readers and instructional texts from selected K-14 grades. Our goal is to determine patterns of text characteristics across all of the elementary grade levels and for a subset of secondary grade levels. The sample will include frequently used narrative and expository texts used for educational purposes in grades K through 12 and introductory college level courses. We will digitally scan (with required permissions) or obtain digital copies of the texts from the textbook publishers. [Some publishing houses have already expressed their interest in participation]. Analyses will be conducted to determine and compare the following measures across grade levels: (a) readability scores, (b) cohesion scores, and (c) differences between separate cohesion metrics. This research will provide a more complete description of the characteristics of the texts that children are exposed to during the K-14 instructional years. It will also allow us to determine the amount of variance explained by textual characteristics in the *Coh-Metrix* versus traditional readability measures.

Empirical examination of *Coh-GIT*. The purpose of *Coh-GIT* is to identify the locations and specific types of cohesion gaps within texts. Our goal here is to verify that *Coh-GIT* reliably identifies cohesion gaps. Preliminary experiments will be conducted to examine the face validity of *Coh-GIT* by comparing its output to reading experts' and readers' judgments of points of difficulty in texts. For

our subsequent experiments, we will take advantage of eye-tracking technology that we have available to verify that *Coh-GIT* reliably predicts participants' fixation times. McNamara and Kintsch (1996) found longer reading times for less cohesive expository texts and Zwaan et al. (1995) reported that reading times for sentences in narrative texts increase robustly with the number of coherence categories that have breaks in continuity. Therefore, gaze durations should be longer for words that are in sentences associated with coherence gaps than words in other sentences. Our first experiment will include college students; our second will include children in grades 3, 4, and 5.

Experiment 7: College student participants (n=80) will read the same passages as described in Experiment 6 (i.e., 12 passages from three genres and four cohesion levels). Measures of reader aptitude will include knowledge, reading ability, and motivation. Participants will be asked to recall the passages to ensure reading for comprehension. Each passage will be divided into four text screens to allow for a font size large enough to ensure eye-tracking accuracy. An ASL 501 eye tracker will measure the participant's eye movements and record data on total fixations per word, first pass fixation times, and number of regressions. Participants will be calibrated before reading each passage. As in Experiment 6, participants will read the passages and type their recall of the passage either after reading each one (immediate recall) or after reading all of the texts (delayed recall). Analyses are similar to Experiment 3, except that the locus of processing load will also be identified.

Our hypotheses are identical to those formulated in Experiment 6. We hypothesize that passage recall and total fixation time will vary as a function of cohesion for each of the three text genres. Effects of vocabulary-driven cohesion (referential relations in particular) will be more pronounced for the expository (historical, scientific) texts than for the narrative texts because the former contain fewer familiar concepts. We expect that fixation times for vocabulary-driven cohesion will vary as a function of readers' knowledge, while grammar-driven cohesion will vary as a function of reading

ability. We will confirm that locations in the texts with cohesion gaps identified by *Coh-GIT* are associated with longer fixation time and more regressions than locations without cohesion gaps (adjusting for reader aptitudes). Finally, the recall coherence matrix should be predicted by the coherence matrices of particular coherence categories that were found to be significant in the previous studies.

Experiment 8: Children in grades 3, 4, and 5 (n=40 per grade level) will read four expository and four narrative passages that vary linearly in cohesion and that have been estimated by readability formulas to be appropriate for grades 2-6. Measures of reader aptitude will include those indicated in Experiments 1-3 to be most predictive of comprehension and coherence effects. Participants will be asked to verbally recall the passages to ensure that they are reading to comprehend the material. [Recall protocols will be recorded and analyzed.] We will examine eye-tracking patterns to verify that *Coh-GIT* reliably predicts participants' eye-tracking behaviors. In contrast with other eye trackers (e.g., the Purkinje eye tracking systems), the ASL Model 501 is very light and non-intrusive. The participant can wear lightweight, head-mounted optics and have unrestricted freedom of movement, which makes this eye tracker suitable for children.

We have similar expectations here as in Experiment 7. In addition, we hypothesize that the children's sensitivity to coherence gaps (in terms of fixation times and regressions) will vary linearly as a function of grade level because children in later grades (i.e., 4-6) are increasingly more likely to make inferences than children in earlier grades (i.e., 1-3; Paris, 1991; Van den Broek, 1989).

Summary

The following table summarizes the approach, time frame, and additional personnel required to achieve the three research goals of this project. The three PIs (McNamara, Louwerse, Graesser) will play significant roles within all phases of the proposed research, so the table lists only additional personnel.

Research Goal	**Research Approach**	**Time Frame**	**Additional Personnel**
A. Develop automated Cohesion metrics (*Coh-Metrix*) and the Cohesion Gap Identification Tool (*Coh-GIT*)	Computational simulation	2 yrs (Y1-2)	**Team A:** Computational Linguist (full-time), Computer Programmer (part-time), Computer Scientist (Faculty), Mathematical Psychologist (Faculty), 2 Graduate RAs (CS; Psych), Undergrad. RA
B. Empirically examine interactive effects of text cohesion and reader aptitudes	Experimental, Quasi-Experimental	2 yrs (Y1-2)	**Team B:** Experimental Psychologist (full-time), School Psychologist (Faculty), 2 Graduate RAs (Psych), Undergrad. RA
C. Fine-tune and establish validity of *Coh-Metrix* and *Coh-GIT*	Experimental, Archival, Computational simulation	2 yrs (Y2-3)	**Teams A and B**

Conclusion

Our project will make important theoretical and applied contributions. This research will further our understanding of comprehension processes for young readers (McNamara, 2000) and effects of cohesion across a wide developmental range. Moreover, the potential applied contributions of the *Coh-Metrix* and *Coh-GIT* tools are innumerable. These tools will allow readers, writers, editors, educators, researchers, and policy makers to more accurately estimate the appropriateness of a text for their audience, predict comprehension, and pinpoint specific problems with text. Thus, this project will be of benefit to both practitioners and policy makers and make a substantial contribution toward OERI's goal to improve students' ability to comprehend and learn from text.