Coh-Metrix: Capturing Linguistic Features of Cohesion

Danielle S. McNamara

Max M. Louwerse

Philip M. McCarthy

Arthur C. Graesser

University of Memphis

Contact Information:

Danielle S. McNamara, Ph.D.
202 Psychology Building
400 Innovation Drive
The University of Memphis
Memphis, TN 38152-3230
d.mcnamara@mail.psyc.memphis.edu

Abstract

This study addresses the need in discourse psychology for computational techniques that analyze text on multiple levels of cohesion and text difficulty. Discourse psychologists often investigate phenomena related to discourse processing using lengthy texts containing multiple paragraphs, as opposed to single word and sentence stimuli. Characterizing such texts in terms of cohesion and coherence is challenging. Some computational tools are available, but they are either fragmented over different databases, or they assess single, specific features of text. Coh-Metrix is a computational linguistic tool that measures text cohesion and text difficulty on a range of word, sentence, paragraph, and discourse dimensions. The current study investigated the validity of Coh-Metrix as a measure of cohesion in text using stimuli from published discourse psychology studies as a benchmark. Results showed that Coh-Metrix indices of cohesion (individually and combined) significantly distinguished the high versus low cohesion versions of these texts. The results also showed that commonly used readability indices (e.g., Flesch-Kincaid) inappropriately distinguished between low and high cohesion texts. These results provide a validation of Coh-Metrix, thereby paving the way for its use by researchers in cognitive science, discourse processes, and education, as well as for textbook writers, professionals in instructional design, and instructors.

Coh-Metrix: Capturing Linguistic Features of Cohesion

Traditionally, discourse psychologists investigated the effects of cohesion on online and offline cognitive processes by manipulating cohesive cues in the text (Gernsbacher, 1990; Zwaan & Radvansky, 1998). For instance, referential cohesion, the relatedness between persons and objects, has been manipulated by replacing definite articles by indefinite articles (Gernbacher & Robertson, 2002; Yekovich & Walker, 1978); temporal cohesion has been manipulated by changing the order of events (Ohtsuka & Brewer, 1992) or the scope of a time shift (Anderson, Garrod, & Sanford, 1983); spatial cohesion has been manipulated by changing the points of view of the protagonist (Black, Turner, & Bower, 1979); and causal cohesion has been manipulated by varying the degree of causal relatedness (Myers, Shinjo, & Duffy, 1987). The manipulation of text variables in these studies was possible because sentence pairs, or very small text fragments, were used. By contrast, when research questions call for longer texts, it becomes more difficult to keep track of the various sources of cohesion. This challenge is exacerbated for naturalistic texts such as newspaper articles or textbooks (Van Oostendorp, Otero, & Campanario, 2002).

At the same time, investigating the effects and sources of text cohesion and text difficulty remains an important objective in discourse psychology research. Text difficulty is also an important and common concern in education. One source of this concern is related to students' ability, or inability, to comprehend text (Snow, 2002). For example, according to the 2005 National Assessment of Educational Progress (NAEP) report, more than a quarter of U.S. students scored below a basic level of proficiency in reading at grades 4, 8, and 12 (National Center for Education Statistics, 2005). Such statistics are particularly relevant in light of studies investigating interactions between the effects of cohesion and readers' abilities. Cohesive cues have been found to be particularly beneficial for low-knowledge readers (Loxterman, Beck, & McKeown, 1994; McNamara et al., 1996; McNamara & Kintsch, 1996). Cohesion gaps in text force the reader to generate inferences to bridge those gaps. The lack of referential or causal cohesion forces the reader to infer ideas, relationships, or events. While this induced active processing can be beneficial for high-knowledge readers, low-knowledge readers often lack the world knowledge needed to make the inferences necessary to meaningfully connect constituents in the text (McNamara, 2001; McNamara et al., 1996). Thus, low-knowledge readers are more likely to benefit from increased cohesion, whereas high-knowledge readers may not as much (O'Reilly & McNamara, 2007).

In sum, cohesion and text difficulty have been shown to be important factors in discourse psychology, but as longer texts are increasingly used in discourse psychology research, there is a growing need to provide computational measures that can more easily and reliably track various aspects of language and cohesion. Although there are several computational linguistic databases and techniques available to analyze text, they are distributed across various sources (e.g., CELEX, Baayen, Piepenbrock, & van Rijn, 1993; LSA, Landauer & Dumais, 1997; MRC Psycholinguistic database, Coltheart, 1981; syntactic parsing, Charniak, 2000; WordNet, Fellbaum, 1998). The lack of a one-stop, computational linguistic department store has rendered it difficult, and sometimes impossible, for researchers to obtain an overview of computational linguistic indices on texts.

Perhaps the best approximate of a readily available technique to assess text difficulty is to assess text *readability*. Readability formulas became popular in the 1950s with the research of scholars such as Flesch (1948) and Dale and Chall (1949). By the 1980s, over 200 readability measures had been developed, with over a 1000 supporting studies (Chall & Dale, 1995; Dubay, 2004). Although readability measures have not been without their critics (Connaster, 1999; Duffy 1985; Manzo, 1970; Maxwell, 1978), such measures remain commonplace today in such examples as Flesch Reading Ease, Flesch-Kincaid Grade Level (Klare, 1974-5), Degrees of Reading Power (DRP; Koslin, Zeno, & Koslin, 1987), and Lexile scores (Stenner, 1996). Despite the seeming diversity of readability formulas, these measures are all based on, or highly correlated with, two variables: the frequency or familiarity of the words, and the length of the sentences. Word length has a strong negative correlation with word frequency (Haberlandt & Graesser, 1985; Zipf, 1949), so number of letters or syllables provides an excellent proxy for word frequency or familiarity. Thus, many measures are based simply on the length of the words and sentences. These components of readability formulae certainly have validity as indices of text difficulty. However, word length and sentence length alone explain only a part of text comprehension. More extensive computational measures may be desirable. One purpose of this study is to examine how a common and readily available readability index (Flesch-Kincaid) fares in distinguishing between high and low cohesion texts. If such indices appropriately distinguish between high cohesion (easy) and low cohesion (difficult) texts, then other indices would not be necessary beyond what is already available. However, if they do not, then indices of cohesion would augment our understanding of and ability to measure text difficulty.

Coh-Metrix is a tool that provides a wide range of computational linguistic indices to meet the growing need for comprehensive and automatic text analyses. Coh-Metrix uses lexicons, a syntactic parser, latent semantic analysis (LSA), and several other components that are widely used in computational linguistics (Graesser et al., 2004). For example, the MRC database is used for psycholinguistic information about words (Coltheart, 1981). Syntax and parts-of-speech are analyzed using Charniak's syntactic parser (Charniak, 2000) and WordNet provides linguistic and semantic features of words and relations between them (Fellbaum, 1998; Miller, Beckwith, Fellbaum, Gross, & Miller, 1990). LSA is used to compute the semantic similarities between words, sentences, and paragraphs by applying statistical computations, including Singular Value Decomposition, to a large corpus of text (Landauer & Dumais, 1997; Landauer, McNamara, Dennis, & Kintsch, 2007).

Coh-Metrix thus seems to fulfill the need of discourse psychologists and other researchers to have access to one computational linguistic tool that analyzes texts on various linguistic features. Indeed, Coh-Metrix has been used to detect a wide variety of differences in text and discourse. For instance, several studies have identified differences between spoken discourse and written text (Graesser, Jeon, Yang, & Cai, 2007; Louwerse, McCarthy, McNamara, & Graesser, 2004), as well as differences between different sources, purposes, and even the specific writers of written text (Crossley, Louwerse, McCarthy, & McNamara, 2007; Graesser, Jeon, Cai, & McNamara, 2008; Graesser & Morgan, 2008; Hall, Lewis, McCarthy, Lee, & McNamara, 2007; Lightman, McCarthy, Dufty, & McNamara, 2007; McCarthy, Briner, Rus, & McNamara, 2007; McCarthy, Lewis, Dufty, & McNamara, 2006). Collectively, these studies demonstrate that Coh-Metrix provides a powerful text analysis tool, capable of assessing and differentiating a wide variety of text types from the chapter level to the sentence level.

Even though these studies have shown that Coh-Metrix can successfully be used to answer various research questions concerning distributions of linguistic features across text types, the validity of Coh-Metrix has not yet been tested on a benchmark of cohesion. That is, the question can be raised whether Coh-Metrix can reliably discriminate between high and low cohesion texts. To address this question, we use texts manipulated by discourse psychologists to be higher or lower in cohesion. If Coh-Metrix indices can distinguish between high and low cohesion texts in available stimuli used in published discourse psychology studies, it will have passed a critical benchmark test.

The purpose of the current study is to assess the validity of Coh-Metrix in assessing cohesion, and in so doing, we also compare the outcome of Coh-Metrix indices with two common readability formulas (i.e., Flesch Reading Ease and Flesch-Kincaid Grade Level). For this purpose, we analyzed 19 samples of pairs of texts with high versus low-cohesion versions from 12 published experimental studies. The number of texts included in this study was constrained first by the number of studies published when we collected the texts, and second by the availability of the texts examined in those studies. Published studies on text cohesion for which the texts could not be accessed could not be included in the current study. We provide a review of the included studies in the following section and provide details on their results and effect sizes in Appendix B. Although the reported effect sizes constitute clear trends, the methods used to obtain these results differ from experiment to experiment, so some caution is advised when comparing results across studies.

Corpora: Studies on Text Cohesion

Two criteria determined the selection of the texts for the current study, First, the studies investigated the comprehension of multi-paragraph texts and had different versions prepared by experimenters that manipulated cohesion. The original versions of the texts came from a variety of sources. Some were created by the researchers, but most were culled from books, textbooks or encyclopedia articles. All of them were modified in some way by the researchers to create high and low cohesion versions of the texts. We were solely interested in extended text that mimicked natural text, as opposed to short sentence-pairs or sentence-triplets. Secondly, texts used in the selected studies had to be available in the published study or be made available by the author. With these criteria, we obtained 19 texts from 12 published studies for the current analysis (see Appendix A).

A starting point in our literature review was the Britton, Gulgoz, and Glynn (1993) review which identified 15 studies that fit our criteria, all published prior to 1989 (see Appendix A). To identify the studies published between 1989 and 2003, we searched major journals with articles on reading comprehension and text processing, including the *Cognition and Instruction; Discourse Processes; Journal of Educational Psychology; Journal of Experimental Psychology: General; Journal of Experimental Psychology: Learning, Memory, and Cognition; Cognitive Psychology; Journal of Memory and Language; Memory and Cognition,* and *Reading Research Quarterly*. This process led to the identification of 14 additional studies. Thus, 29 studies of text revision and comprehension were identified that met our criteria. These studies are listed in Appendix A.

This study was limited in the sense that only those for which we were able to obtain the full texts could be included. Sample size was thus dictated by studies in discourse psychology using full texts and the availability of those texts. Texts were obtained from the articles

themselves, from books the articles referred to, or from the authors of the published articles. For older publications, this proved to be more challenging. We were able to obtain the texts for 15 of the 29 studies from the article, the internet, or by contacting the author(s) of the studies. Two studies were redundant, however, because they used the same texts (i.e., Britton & Gulgoz, 1991; McNamara, & Kintsch, 1996). Also, texts used in two other studies (Lorch & Lorch, 1996; Meyer & Poon, 2001) were excluded from the analysis. The texts from Lorch and Lorch (1996) were excluded because the manipulation was on organizational signaling devices, such as headings, topical overviews, and topical summary. These features involved formatting and structuring the text (e.g., paragraph indentation, numbered subheadings), which were beyond the scope of the Coh-Metrix tool. The texts from Meyer and Poon (2001) were excluded because the text manipulation was strictly limited to the addition of very specific structural markers (e.g., *for example*, *because*).

Our sampling procedure resulted in a total of two versions of each of the 19 texts from 12 published studies. Several of these studies included more than two versions of the texts. The present study was limited to the examination of the highest and lowest cohesion text pairs from those studies. Appendix B presents a summary of the 12 studies included in the analyses, including the authors, participants, text titles, a brief summary of the text revisions, and the results for which effect sizes could be computed.

The effect sizes presented in Appendix B were computed using Effect Size Analysis Software, Version 1.0 (Shadish, Robinson, & Lu, 1999). Effect sizes reflect the size of the difference between the means being compared in relation to the variance of the means. Cohen (1988) uses the following standards for effect sizes: d = .2-.4 is a small, d = .5-.7 is a medium, and d > .8 is a large effect size.

Many of the studies included in this analysis examined interactions between the effect of text manipulation and an experimental manipulation or participants' individual differences. Thus, we computed effect size of the text manipulation separately for each group of participants whenever possible, with the following two exceptions. If the effect size could not be computed separately for a given group of participants due to the lack of necessary statistical values (e.g., F, SD, MSE, etc.), we computed the effect size based on all participants across different groups. Also, when separate standard deviations for each subset were missing, effect sizes for a subset of participants were computed using the standard deviation of all samples in the study. Thus, the effect sizes reported in Appendix B are informative but should be treated with some caution.

The following are brief descriptions of each of the 12 studies and their texts that were included in our corpora. The studies are described in chronological order.

1. Beck, McKeown, Omanson, and Pople (1984) included a narrative text titled "The Raccoon and Mrs. McGinnis." Although they used two texts in this study, only the Raccoon text was available. It was obtained from a second grade text book of the Reading 720 series. The goal of the Beck et al. study was to examine whether revision of a narrative passage improved third-grade students' comprehension of the story. Their revisions were aimed to alleviate three problems in the text: 1) *surface problems*, including syntactic complexity, unclear relations between reference and referent in the text, inappropriate use of conjunctions, and awkward descriptions of events and states; 2) *knowledge problems,* involving readers' lack of familiarity with the meaning and significance of events, and the relations between the events; and 3) *content*

problems, due to ambiguous, irrelevant, or confusing content. The authors identified 116 problems in the text and repaired the problems in the revision process. The children's comprehension was measured with recall and multiple choice questions. Their results showed that the text revisions improved comprehension, with skilled readers showing greater benefits than less skilled readers for text recall.

2. E. Kintsch (1990) applied the Van Dijk and W. Kintsch (1983) model of text comprehension for the revision of expository texts for students of varying ages (6th grade, 10th grade, and college). Two expository passages were developed, one of which (*Peru and Argentina*) was available for the current analysis. The texts were written at a reading level appropriate for sixth grade students and had a "compare-contrast" rhetorical structure in which the two countries were compared on the basis of three attributes: geography, economy, and society/culture. The manipulations to create the low cohesion version included inserting topic shifts, more difficult words, longer and more complex sentence patterns, and fewer connectives to signal relations between ideas. The dependent measure of the study was summary writing quality. Whereas older students could compensate for cohesion gaps by generating inferences to mentally re-ordering the content, younger students' comprehension was significantly disrupted by the lack of cohesion in the text; their summaries tended to become less general and more detail oriented. However, effect sizes were only computable for the entire set of participants. There was more reordering of ideas in the summaries for the low cohesion texts, and a greater use of connectives after reading the high cohesion text.

3. Beck, McKeown, Sinatra, and Loxterman (1991) examined the effects of text revision for elementary school students' comprehension of social studies texts. They included a sequence of four passages (*The French and Indian War, No Taxation without Representation, The Boston Tea Party,* and *The Intolerable Acts*) about the American Revolution obtained from a fifth grade social studies text book. The goal of the study was to examine whether systematic revision based on cognitive processing theory of comprehension improved children's understanding of social studies texts. Text comprehension was measured with recall and open-ended questions on the information common to both versions of the texts. The revisions were designed to provide additional causal connections between the ideas, concepts, and events, including making connections explicit as well as clarifying, elaborating, explaining, and motivating important information. The revisions were also designed to minimize the need for knowledge-based inferences. The results confirmed that revisions improved the students' recall as well as their performance on open-ended question.

4. Britton and Gulgoz (1991) approached the effect of text revision on comprehension from a more specific theoretical perspective based on the Van Dijk and Kintsch (1983) model of text processing (see also Kintsch & Van Dijk, 1978; Miller & Kintsch, 1980). Britton and Gulgoz specifically focused on how textual cues influence how readers understand the nature of connections between ideas presented in differ. Having this perspective in mind, they revised a text taken from the US Air Power Key to deterrence (*Air War in the North*; US Air Force Reserve Office Training Corpse, 1985). In the "principled revision", they first identified "coherence breaks" based on Van Dijk and Kintsch's model of comprehension. A coherence break was a location in the text in which there was no explicit cue on how the new information was linked to prior text. They found 40 coherence breaks in the text. Britton and Gulgoz applied three principles to repair these breaks. Principle 1 was to add argument overlap such that a

sentence repeated an idea stated in the previous sentence. Principle 2 was to rearrange part of each sentence so that readers first received old information (i.e., an idea presented previously in the text) and then the new information. Principle 3 was to make explicit any implicit references that did not have clear referent. Their method of revising text differed from that of Beck and colleagues (1984, 1991) in that it did not involve adding extra information (e.g., elaboration, explanation) to scaffold readers' knowledge deficits. College students read either the original or revised version of the text and comprehension was measured with free recall, multiple choice questions, and a keyword association task (effects sizes could not be computed for the latter). The results showed that the principled revision improved comprehension according to all three measures. Further, their efficiency measure for recall (the number of propositions recalled per minute of reading time) indicated that the revision made the comprehension process more efficient. McNamara and Kintsch (1996) later replicated the Britton and Gulgoz findings with this text, showing advantages for the principled revision, and further showing that only low-knowledge readers showed benefits from the revision.

5. Loxterman, Beck, and McKeown (1994) conducted a study similar to the Beck et al. (1991) study with a text (*El Niño*) extracted from a sixth grade social studies textbook. This was a shorter text (167 words) than used in the Beck et al. (1991) study because the students were instructed to think aloud while reading the text. The revision of the text followed the same method used in the 1991 study. The revisions made the relationships between concepts in the text more explicit and scaffolded the readers in integrating that information with knowledge to develop a more coherent text representation. Sixth grade students read either the original or revised text in either a silent reading or think-aloud reading condition. The dependent measures were recall and open-ended questions. In both the silent reading and think-aloud conditions, students' comprehension was better with the revised text. In the second experiment, two additional factors were included: reading skill level (high, intermediate) and delay after the reading (immediate, 1 week). The results indicated that participants benefited from the text revisions regardless of their reading skill level both immediately as well as after a one-week delay.

6. McNamara, E. Kintsch, Songer, and W. Kintsch (1996) examined the effects of text revision on comprehension of biology texts. They conducted two experiments with sixth to eighth grade students using two different biology texts. The text used in the first experiment (*Traits of Mammals*) was excerpted from a biology text targeting sixth to eighth grade students. The original text was locally coherent but lacked global coherence. Thus, the revisions for the high cohesion version focused on increasing links between subtopics and the main topic. The results indicated that the children comprehended the revised text better than the original text across the measures (recall, open-ended questions, keyword sorting). In the second experiment, they used a passage on *heart disease*, which was based on an entry in a science encyclopedia for school aged students. They manipulated coherence orthogonally at the local and global levels by adding or deleting cohesive cues. Maximizing local coherence of a text involved: 1) replacing pronouns with noun phrases when the referent was ambiguous; 2) adding descriptive elaborations to link unfamiliar concepts with familiar ones; 3) adding sentence connectives to specify the relations between ideas; and 4) replacing words to increase argument overlap. Global coherence was increased by adding topic headers and topic sentences to link each paragraph to the rest of the text and to the overall topic. In creating minimally cohesive texts, the above

processes were reversed. Comprehension and learning was measured with several dependent measures including recall, open-ended questions, and card sorting tasks. The results indicated that students' text recall was improved with the maximally coherent text as compared to minimally coherent text. However, according to the sorting task and inference questions, only the low-knowledge students benefited from reading the high cohesion text; high-knowledge students' comprehension benefited from reading the low cohesion text.

7. Voss and Silfies (1996) examined the effect of text revision on comprehension and its relation to prior knowledge and reading skill. Two pairs of texts were developed with each pair describing two different fictional countries (*Anchad* and *Padria*). The texts were designed to include descriptions of series of events with causal connections. Specifically, a series of causal events led to the initiation of hostilities between the two countries described in the text. Voss and Silfies chose history texts because they provide narrative accounts of how social, political and economic conditions and events interact and depict complex causal relations between the events, people, and concepts. The text manipulations included adding elaborations of causal factors related to the account of original text such that causal relations of how a given set of events led to other events were "unpacked." This approach to text revision resembles the Beck et al. (1991) approach, but more specifically emphasizes the causal cohesion of the text. The dependent measures included answering comprehension questions and writing an essay. The results indicated that comprehension of the unexpanded text was correlated with readers' prior knowledge on history, whereas comprehension of the expanded (more cohesive) text was correlated with readers' level of reading skill.

8. Cataldo and Oakhill (2000) explored whether skilled readers have an advantage in locating information in the text, and whether that advantage is related to their ability to form a more organized representation of the text content. To explore these questions, they asked skilled and less-skilled elementary school readers (5th grade) to read original (well-organized) and scrambled version of stories. The dependent measures included recall, question answering, search time for answers, spatial memory for keywords in the text, and sequential memory for the keywords. They used two texts (*The Demon Barber* and *The Return of Martin Guerre*) which were at suitable levels for the children. There are two versions for each story: original (cohesive) version and a scrambled version. The scrambled versions were obtained by randomly reordering the order of the original sentences. Skilled-readers' were better than less skilled readers at searching for information in the text. This difference was most apparent for the scrambled version of the text. Both skilled and less skilled readers' comprehension performance declined and search time increased when reading the scrambled version as compared to the unscrambled version.

9. Linderholm, Everson, van den Broek, Mischinski, Crittenden, and Samuels (2000) examined how repairing the causal structure of relatively easy and difficult texts can influence comprehension of more and less skilled readers. Two texts *Mademoiselle Germaine* (easy text) and *Project X-Ray* (difficult text) were both social studies texts describing little known events during World War II. According to the Fry (1975) readability scale, Mademoiselle Germaine was at the eleventh grade level whereas Project X-Ray was at the ninth grade level. The authors' classification of difficulty level of the texts, which contradicted grade level estimates of difficulty, was judged based on the number of causal and referential connections per text (Trabasso, Secco, & van den Broek, 1984), the explicitness of the goals, and schema

familiarity. The revision applied to the texts was based on the causal network theory of comprehension proposed by Trabasso et al. (1984) and aimed at repairing the causal structure/organization of the text. The specific repairs applied were: 1) arranging text events in temporal order; 2) making implicit goals of the character explicit; and 3) repairing cohesion breaks caused by inadequate explanation, multiple causality, or distant causal relations. Thus, the revision was somewhat similar to the Voss and Silfies (1996) and Beck et al. (1991) approaches, which focused on the causal relations between the events. Participants in the experiment were college students, whereas comprehension was measured with recall and comprehension questions. Overall the results indicated that both less and more skilled readers benefited from the revision of the difficult text, but the revision of the easy text did not affect performance.

10. Vidal-Abarca, Martinez, and Gilabert (2000) compared the effects of two types of text revisions on text comprehension of a history text, *Russian Revolution*, obtained from an eighth grade history textbook (Anaya Publishers, 1987). The first approach was to follow the Britton and Gulgoz (1991) procedure and increase referential links between the sentences by increasing argument overlap. Specifically they identified eight locations of coherence breaks using a program based on the rules proposed by Miller and Kintsch (1980). They used Principles 1 and 3 (i.e., increase argument overlap and make explicit any implicit references) from Britton and Gulgoz (1991) to repair these breaks, creating the Argument Overlap revision of the text. The second approach was based on causal constructionist model of narrative comprehension (e.g., Graesser, Singer, & Trabasso, 1994; Trabasso, Secco, & van den Broek, 1984). According to this model, breaks in coherence occur when the reader needs to make inferences to causally connect two ideas. They identified the causal cohesion breaks by using causal-network analyses created by Trabasso et al. (1984). This method involves repairing breaks in causal-time sequences by adding (1) information to trigger the readers' causal antecedents, and (2) super-ordinate goal references. The third version, the most coherent version, was created by implementing the changes made in both the argument overlap and causal constructionist revisions. Participants were eighth grade students. The effect of text revision on comprehension was measured with immediate and delayed tests of free-recall and a delayed test with open-ended inference questions (where students could refer to the text that they had read). The results indicated that the causal constructionist revision helped readers form a good situation model as indicated by better performance in inference question answering, less erroneous recall, and greater focus on main ideas in the recall. The argument overlap revision in itself did not improve students' deep level comprehension, but the revision with both modifications (examined in the current study) yielded the largest and most consistent benefits for comprehension.

11. McNamara (2001) examined the effect of text revision on comprehension of biology text with college level students. This study addressed the issue of how reading the same text twice versus both versions of the text (high and low cohesion) affects readers' comprehension of the text. The text topic was *cell division* and was obtained from a middle school textbook. The manipulations to increase cohesion were: 1) replacing ambiguous pronouns with nouns; 2) adding descriptive elaborations to link unfamiliar concepts with familiar concepts; 3) adding connectives to specify the relationships between sentences or ideas; 4) replacing or inserting words to increase the conceptual overlap between adjacent sentences; 5) adding topic headers; and 6) adding thematic sentences that serve to link each paragraph to the rest of the text and overall topic. These modifications repaired conceptual gaps in the text without adding additional

information to the text. Thus, the modifications included increasing referential cohesion (similar to Britton & Gulgoz, 1991) and what might be called explanatory cohesion (similar to Beck et al., 1991). Results of the experiment indicated that low-knowledge readers benefited from reading the high-cohesion text, whereas high-knowledge readers benefited from reading the low-cohesion text. The effects of text revision were observed only with text-based questions that probed for the readers' understanding of the basic information stated within a single sentence; text revision did not affect performance on inference questions. O'Reilly and McNamara (2007) replicated these findings with the same text, and also showed that the benefit of reading the low-cohesion text was only found for less skilled, high-knowledge readers. This finding supports the assumption that more skilled readers generate inferences regardless of the cohesion level of the text, and only less skilled (high-knowledge) readers require the cohesion gaps to induce active processing.

12. Lehman and Schraw (2002) examined the effects of local and global cohesion on comprehension and also how relevance (manipulated with instructions) interacts with the text cohesion in moderating the comprehension. The history text was *The Quest for Northwest Passage*. In Experiment 1, low-cohesion text was created by reducing local cohesion by altering the order of the sentences. The sentences that were moved if they 1) promoted referential or causal coherence in the original location and 2) could be relocated within the paragraph without altering the overall meaning of the paragraph. Reading instructions in a high relevance condition were to attend to specific aspects of the story (e.g., the main theme) whereas a low relevance condition included the general instruction to read the text carefully. Comprehension of the passage was assessed using four measures: (1) recognition (i.e., using multiple-choice questions), (2) free recall; (3) an essay (including a holistic situation model score reflecting global understanding, and a total claim score based on the participant's causal arguments and supporting evidence, and (4) an ease of comprehension (text coherence) evaluation based on ratings between 1 and 5. In Experiment 2, temporal flow of the story was interrupted to reduce global coherence and the story was reorganized thematically rather than chronologically. The thematic organization of the passage resulted in four themes with seven breaks in the temporal order of the story. The results indicated that the reduction of local coherence (Experiment 1) did not affect comprehension even though it affected participants' coherence rating. On the other hand, lower global coherence (Experiment 2) impaired college student readers' ability to recall the text content but did not affect their ability to recognize facts in multiple-choice questions. Together the findings further suggest that local and global coherence of text have different effects on comprehension. Type of instructions did not have an effect in Experiment 1, whereas the effects of revision were larger in Experiment 2 when the readers were given the *High Relevance* instructions to attend to specific aspects of the story.

Summary of Corpus Studies

The published studies that we were able to include here showed effects of cohesion across a variety of text genres, text manipulation methods, and target participants. On the one hand, the results are quite complex in that there are differences depending on the experimental measures used; for example, recall or simple questions did not always show the same effects as inference questions and key word sorting. On the other hand, the findings are consistent in many respects. For the most part, cohesion tends to improve comprehension across a wide range of circumstances. Among those studies that have examined the effects of prior knowledge, low-

knowledge readers benefit more from added cohesion than do high-knowledge readers. In contrast, the majority of the studies that have investigated the interactive effects of reading skill indicate that cohesion results in benefits regardless of reading skill. Five of the 12 studies included reading skill measures. Four of these five studies (Beck et al., 1984; Cataldo & Oakhill, 2000; Linderholm et al., 2000; Loxterman et al., 1994) reported significant effects of cohesion for both skilled and less skilled readers. Although Voss and Silfies (1996) found that comprehension of the expanded, higher cohesion showed higher correlations with reading skill than with knowledge, this is difficult to interpret because they did not report how effects varied as a function of reading skill. Thus, this result does not necessarily imply that less skilled readers did not gain from the cohesion manipulations. The results collectively indicated that cohesion is likely to benefit both more and less skilled readers. If the studies had found reduced effects of cohesion for less skilled readers, such a result could be interpreted as indicating that the added cohesive elements increase the demands of the text by increasing its length. As such, it appears that additional cohesive elements do not increase the processing demands of the text, and moreover, they tend to improve comprehension across a wide range of circumstance.

Coh-Metrix Analyses

The purpose of the following analyses is to examine whether Coh-Metrix indices afford distinguishing between high and low cohesion text. Such an analysis serves two purposes. First, it validates Coh-Metrix indices of text cohesion. That is, if the cohesion indices from Coh-Metrix discriminate between the high and low cohesion versions created by researchers in previous studies, it confirms that Coh-Metrix is able to accurately assess text cohesion. Second, this analysis provides a better understanding of what text features characterize text that are associated with improved comprehension. Here, we focus on a subset of those indices that are related to cohesion: LSA, coreference, connectives, and indices related to causality.

Several studies have used LSA to measure differences in text cohesion (Foltz, Kintsch, & Landauer, 1998; McNamara, Cai, & Louwerse, 2007). LSA reflects human knowledge in a variety of ways. For example, LSA correlates highly with humans' scores on standard vocabulary and subject matter tests; it mimics human word sorting and category judgments; it simulates word–word and passage–word lexical priming data; it has been used to estimate passage coherence; and it grades essays as well as experts in English composition (see Landauer et al., 2007). Here, we use LSA to assess cohesion at local and global levels by considering conceptual similarity between adjacent sentences, between all sentences, between each sentence and the paragraph, between each sentence and the text, between paragraphs, and between paragraphs to the text.

In addition to indices of conceptual cohesion provided by LSA, Coh-Metrix provides indices of coreference, connectives, and causal cohesion. Coreference occurs when a noun, pronoun, or noun phrase refers to another constituent in the text. This study focuses on three sources of coreference. First, *noun overlap* is the overlap of nouns between two sentences, with no deviation in the morphological forms of the nouns. That is, *mother*/*mother* would be overlapping nouns, but *mother*/*mothers* would not be. Second, there is *argument overlap*, where the head nouns or pronouns of noun-phrases overlap between two sentences (e.g., *mother/mother; mother/mothers; she/she*). The term argument is used here in line with Kintsch and Van Dijk (1978), with noun/pronoun arguments being contrasted with verb/adjective

predicates. Argument overlap occurs when there is overlap between a noun in one sentence and the same noun (in singular or plural form) in another sentence; it also occurs when there is a matching pronoun between two sentences. Third, *stem overlap* occurs when any content word or pronoun in one sentence refers to a word in another sentence with the same lemma (i.e., core morphological element, be it a noun, verb, adjective, or adverb). Thus, lemma overlap could include overlap between *giver* in one sentence and *giver*, *giving, gives, or gave* in another sentence.

The third index of cohesion we assess here is the use of connectives. Connectives provide explicit cues to the types of relationships between ideas in a text (Halliday & Hasan, 1976; Louwerse, 2002). Coh-Metrix provides an incidence score (occurrence per 1000 words) for all connectives, as well as these connectives broken down into four general types: causal, additive, temporal, and clarification. Causal connectives cue the reader that there is a causal relation between two text segments, such as *because* and *therefore*. Additive connectives cue the reader that two text segments need to be tied together, as in the case of *also, as well*, and *further*. Temporal connectives cue the reader that there is a temporal relation between segments, such as *before* and *after*. Clarification connectives cue the reader that the writer is restating previous text in different words or providing examples to illustrate a concept, such as *for example*.

The fourth index assesses causal cohesion by measuring the ratio of the incidence of causal connectives to *change-of-state* verbs (i.e., *causal ratio*). According to WordNet, these verbs refer to changes of state (*break, freeze*), actions (*impact, hit)*, or events (*move*), rather than states. The necessity of connectives in text will depend on the number of events expressed in the text. A text is judged as more causally cohesive to the extent that there are proportionally more causal connectives that relate actions and events in the text. If there are numerous action and event verbs without causal connectives to aid the reader, then the reader may be more likely to be forced to generate inferences to understand the relationships between the actions and events in the sentences.

Each of the 19 pairs of high and low cohesion texts from the 12 studies were analyzed by Coh-Metrix version 1.4, available at coh-metrix.memphis.edu. The output of the analysis was analyzed treating text pair as the random variable (i.e., cohesion was a within-text variable). The dependent variables were the scores on the indices provided by the Coh-Metrix output.

Insert Table 1

Descriptive and Readability Statistics

Table 1 shows the descriptive and traditional readability statistics for the low and high cohesion texts. These statistics show that the high cohesion texts tend to include more words and sentences, and more words per sentence. The results also indicate that adding cohesion to a text requires adding words to fill in the conceptual gaps. Because this generally lengthens the number of words within a sentence, grade level indices such as Flesch-Kincaid increase because they are partially driven by the number of words per sentence.

Word frequency is an important measure because high frequency words are normally read more quickly and are more easily understood than are infrequent words. Researchers have investigated the impact of frequency on word processing in great depth. One finding is that word

processing time tends to increase linearly with the logarithm of word frequency rather than with the raw word frequency. This is because some words have extremely high frequencies (such as *the* and *is*), with minimal incremental facilitation in processing time over words that are common but not nearly as frequent. The logarithmic transformation makes the distribution of word frequencies better fit a normal distribution and have a closer fit with reading times (Haberlandt & Graesser, 1985; Just & Carpenter, 1980). Table 2 presents the average word frequency according to the Celex written corpus (Baayen, Piepenbrock, & van Rijn, 1993), using the mean logarithm of word frequency for the lowest frequency content word per sentence. The underlying theoretical foundation of this measure is that sentence comprehension is most constrained by the rarest words in a sentence. In essence, a rare word in a sentence can create comprehension difficulties for the entire sentence. Using this measure of word frequency, we see that the low-cohesion texts tended to contain higher frequency (hence more familiar) content words compared to the high-cohesion texts. Words of lower word frequency are indicative of a text being more knowledge-demanding. Likewise, word concreteness (for content words) mimics that result indicating that high-cohesion texts have lower word concreteness than do the low-cohesion texts. These results support the claim that the high cohesion texts tend to increase processing demands on the reader at the lexical level even though they are less demanding due to the added referential cohesion. Thus, this could suggest that there is a tradeoff between difficulty at the lexical level and referential cohesion level, as we will discuss next.

Latent Semantic Analysis (LSA) Indices

The results in this section examine whether LSA detected differences between the high and low cohesion versions of the texts. Coh-Metrix includes the six types of LSA indices discussed earlier. As shown in the second section of Table 1, four of the six LSA indices showed significantly higher cohesion scores for the high cohesion text versions compared to the low cohesion versions. The two that did not were indices of global cohesion (paragraph to paragraph, paragraph to text). The finding that the global cohesion scores were not significant is compatible with the conclusion that the local cohesion manipulations were effective in that the comprehension advantages for the high-cohesion texts could not be attributed to differences between texts at more global levels.

Insert Table 2

Coreference

Table 2 presents the mean coreference scores for the corpus of texts as a function of the particular measure of word overlap. Coreference occurs when a noun, pronoun, or noun phrase refers to another constituent in the text. Coh-Metrix provides several classes of coreference, but three of these are reported in the present study (i.e., noun, argument, and stem overlap, defined earlier).

Coreference indices vary by distance between the target sentence and the previous sentences in the text. Adjacent overlap includes only the two adjacent sentences. Distal indices include more than two adjacent sentences. A distance of two sentences includes the target sentence and the two previous sentences. A distance of three sentences consists of overlap between a target sentence and a coreferent among any of the three previous sentences. The *all*

distances index includes the overlap between each sentence and all other sentences in the text. This is intended as a more global index of cohesion.

All of the indices represent average overlap over selected sentence pairs. The overlap for each sentence pair is either 0 (not overlapped) or 1 (overlapped). For unweighted versions, the indices are the simple average overlap over the sentence pairs. For weighted versions, the indices are a weighted average overlap that adjusts for the distance between sentences. The weight for each sentence pair is the inverse of the distance between two sentences (e.g., 1/2, 1/3), with adjacent sentences having a distance of 1.

The results presented in Table 2 indicate that all of the indices revealed significantly higher cohesion values for the high-cohesion than the low-cohesion texts. This result validates the Coh-Metrix tool with respect to the coreference indices. One additional question is whether there were differences in detecting high versus low cohesion as a function of index type, distance, and weighting. Therefore, a mixed ANOVA was conducted including the within-text factors of cohesion (high, low), index type (noun, argument, stem), distance (all distances, 2 sentences, 3 sentences) and weight (unweighted, weighted). The results for the adjacent indices are presented in Table 2 but could not be included in the ANOVA because differential weights in a weighted version cannot exist for adjacent sentences. There were main effects of cohesion ($F(1,18)=23.36$, MSe=0.136, $p<.001$, $M_{high}=0.480$, $M_{low}=0.344$), index type ($F(1,18)=36.32$, MSe=0.789, $p<.001$, $M_{noun}=0.364$, $M_{arg}=0.417$, $M_{stem}=0.455$), distance ($F(1,18)=45.78$, MSe=0.016, $p<.001$, $M_{2sent}=0.451$, $M_{3sent}=0.433$, $M_{all}=0.352$), and weight ($F(1,18)=41.84$, MSe=0.004, $p<.001$, $M_{unwtd}=0.396$, $M_{wtd}=0.428$), all in the predicted direction yielding higher scores for high cohesion than for low cohesion texts. There was an interaction between distance and weighting ($F(1,18)=42.69$, MSe=0.001, $p<.001$), indicating that weighting affected the all-distance indices ($M_{unwtd}= 0.32$ $M_{wtd}=0.38$), but had little effect on two-sentence ($M_{unwtd}= 0.44$, $M_{wtd}=0.46$) and three-sentence ($M_{unwtd}= 0.42$ $M_{wtd}=0.44$) distances. Thus, there was a significant difference between unweighted and weighted values for the all-distance indices, but little difference when the distance being weighted in value was only two to three sentences.

Cohesion interacted with weighting ($F(1,18)=9.11$, MSe=0.001, $p<.01$) such that the weighted algorithms yielded larger differences between text versions (Diff = 0.144) than did the unweighted algorithms (Diff = 0.128). Cohesion also interacted with distance ($F(1,18)=11.25$, MSe=0.003, $p<.01$), such that the more local indices of coreference yielded larger differences between text versions ($Diff_{2sent} = 0.152$; $Diff_{3sent} = 0.144$) than did the all-distances algorithms ($Diff_{all} = 0.111$). Although cohesion did not interact with index type, $F(1,18)=2.28$, $p>.10$, the Cohen's *d* effect sizes (in the last column of Table 2) indicate that noun overlap and argument overlap indices yielded slightly more robust differences between text versions (Cohen's *d* M=1.02; Cohen's *d* M=1.00; respectively) compared to the stem overlap measures (Cohen's *d* M=0.78). However, all of the coreference indices successfully detected the differences between the high and low cohesion versions.

To further test our conclusions, we examined whether each of the 19 text pairs showed significant differences in cohesion according to the coreference indices. For each text, index was treated as the random variable, and cohesion as the within-text variable. These analyses indicated for each text whether there were reliable differences between the high and low cohesion versions according to the coreference indices. All but one of the texts showed significant differences

between the high and low cohesion versions. Interestingly, the text that did not show a significant difference (i.e., *The Quest for Northwest Passage* in Lehman & Schraw, 2002; $F(1,20)=3.66$, $p=.07$), also did not yield comprehension differences as a function of cohesion (see Appendix B).

In summary, the more global indices (i.e., all distances) and unweighted algorithms tended to yield smaller, though significant, differences between the two text versions. However, all of the coreference indices successfully detected the differences between the high versus low cohesion versions.

One potential concern for this study is that low and high-cohesion texts were not of equal length, as indicated in Table 1. As stated earlier, increasing the cohesion of a text necessarily requires adding words; thus this has been a confounding variable in most studies of cohesion. To somewhat alleviate that concern here, we truncated the high-cohesion texts to be equal in length to the low-cohesion texts. We found that the results and trends were equivalent to those that we reported here. We also conducted the analyses including number of words in the text as a covariate and the differences between high and low cohesion texts remained significant. Thus, the number of words is not driving our reported differences between high and low cohesion texts.

Connectives

Coh-Metrix provides an incidence score (occurrence per 1000 words) for four types of connectives: causal, additive, temporal, and clarification. Coh-Metrix also provides an incidence score for all of these connectives combined. The results in Table 1 indicate that the higher cohesion texts contained more causal connectives. However, there were no differences between the texts in terms of the other types of connectives. This might be expected because causal connectives are more often considered to be signatures of textual cohesion (Gernsbacher, 1990; Sanders, Spooren & Noordman, 1992; Zwaan & Radvansky, 1998), and the researchers (i.e., the authors of the 12 studies) would be more likely to add those types of connectives. Many of the researchers intentionally attempted to increase causal cohesion, and thus the causal connectives prevailed.

Causal Indices

The Coh-Metrix index of causal cohesion investigated here is the causal ratio, which is the ratio of the incidence of causal connectives to change-of-state verbs. The results in Table 1 indicate that the higher cohesion texts contained more causal connectives and that the ratio of causal particles to change-of-state verbs was greater. Thus, there were more connectives, and they were needed to express the relationships between actions and events expressed in the texts.

Combined Predictors of Cohesion: Towards a Computational Model

Multicollinearity Analysis

The previous analyses indicated which variables significantly distinguished between high-cohesion and low-cohesion texts. However, they do not indicate which variables would collectively predict cohesion levels. A first step toward that goal is to assess multicollinearity, or the extent to which the indices account for unique variance associated with cohesion levels. One statistical technique to assess multicollinearity of variables is *tolerance*, which varies between 0

and 1, indicating the percentage of variance that cannot be explained by other variables. If tolerance values are low, this is in an indication that the variable is redundant relative to the other variables; in other words, a variable may account for variance that another variable, or a combination of other variables, explains equally well or better. What exactly constitutes *low tolerance* depends very much on the data under analysis, and its associated hypotheses, assumptions, and body of support. As a rule of thumb, Allison (1999) suggests that tolerance values under .40 may indicate redundancy. A value of .40 indicates that about 40% of the variance explained by the variable is not accounted for by other variables in the model. Even though our dependent variable is binary (low versus high cohesion), it can be viewed as *presence of high cohesion* or *presence of low cohesion*, and therefore functions well for the purposes of assessing tolerance (Leech, Barret, & Morgan, 2008). Therefore, we conducted a series of tolerance analyses using the variables reporting significant differences (see Tables 1 and 2) as independent variables and low/high cohesion as the dependent variable.

Among descriptive indices, two variables, *log word frequency* (.680) and *content word concreteness* (.527) reported tolerance values above .40. For LSA, all variables tolerance values were less than .15, suggesting that all of the variables were explaining similar variance. As such, we retained the variable with the highest effect size: *LSA adjacent sentence to sentence*. For coreference variables, no index registered a tolerance value above .002. Hence, we again selected the variable with the highest effect size to represent coreference, *noun overlap adjacent unweighted*. Among the connectives and the causal variables, only one index was significant for each set, so no tolerance analysis was necessary. Thus, we retained *causal connectives* and *causal ratio*. The six retained variables might be described as *representative* of their respective discourse function as it pertains to cohesion. However, one further tolerance analysis is necessary to ascertain whether there is redundancy across these discourse functions. In this analysis, causal connectives fell below the tolerance threshold (.346), presumably because of variance explained by the causal ratio (.426), which includes causal connectives in its calculation. The LSA index (.429) and the coreference index (.590) recorded moderate but acceptable tolerance values. The word frequency tolerance value (.853) and the word concreteness value (.829), suggest that both variables are providing unique contributions. Based on this analysis, we excluded only the causal connectives index because of its low tolerance value and also because its effect size was lower than that of the similar variable causal ratio.

Discriminant Analysis

The above analysis indicates that five Coh-Metrix variables may explain differences in low and high cohesion text. To supply further evidence to this conclusion, we conducted a discriminant function analysis to test the relative contribution of the combined variables. In this analysis, low/high cohesion was again the dependent variable and the five retained Coh-Metrix variables were the predictor variables. Prior to the analysis, we followed common practice and centered the predictors for ease of interpreting the results (Kinnear & Gray, 2008; Kreft, de Leeuw, &Aiken, 1995).

The results of the discriminant analysis were significant (Wilks' Lambda (5) = .616, p = .006), and the model successfully predicted 76.3% of the items (baseline = 50%, see Table 3). The Fisher's coeffecients (see Table 4) were nearly all in directions reflecting the means reported in Tables 1 and 2. Word frequency was higher for low cohesion texts; LSA values, coreference

values, and causal ratio values were lower for low cohesion texts. While the content word concreteness Fisher's coefficients were in the opposite direction as the means, this reversal in sign is not meaningful because this variable was also not significant.

Results for tests of equality of group means (see Table 4) show that noun overlap and causal ratio are significant contributors to the model, and that word frequency and LSA are marginally significant, whereas word concreteness was not. These results likely emerged because these texts were manipulated by discourse researchers whose explicit intentions were primarily to modify referential and causal cohesion. These results would thus indicate that Coh-Metrix provides an objective measure of those researchers' intentions. Notably, a corpus comprised of texts that naturally vary in difficulty may yield somewhat different results. Also, the results need be interpreted with caution due to the limited sample size.

Qualitative Classification Analysis

In this section, we explore qualitatively which texts are misclassified by the discriminant analysis. As shown in Table 3, 29 of the 38 texts were correctly classified and 9 texts were misclassified based on the discriminant analysis. For 5 of the 9 misclassified texts (i.e., the low cohesion versions of Air War in the North, Quest for the Northwest Passage, Traits of Mammals, Padria, and El Niño), the model predicted that both versions of the texts were high in cohesion. For the remaining 4 misclassified texts (i.e., high cohesion versions of Mademoiselle Germaine, Project X-Ray, the Return of Martin Guerre, and Peru and Argentina), the model predicted that both versions of the texts were low in cohesion.

These results provide three major points worth noting. First, there were no cases where the text cohesion prediction and the items were both reversed such that a low cohesion text was predicted as high and the high cohesion version was predicted as low. Second, the potentially problematic *narrative* texts in this study (the Raccoon and Mrs. McGinnis) were both classified correctly. And, third, the probabilities generated from the model for 8 of the 9 misclassifications were in the correct direction. For example, although the texts for *Air War in the North* were both classified as high cohesion, the probability for the genuinely high cohesion text (.944) was higher than the probability for the genuinely low cohesion text (.644).

We performed additional analyses to examine the misclassifications. The *Air War in the North* and the *Quest for the Northwest Passage* texts were manipulated for coreference overlap, but not for causal information. For these, the indices of noun coreference and LSA were in the correct direction, whereas causal ratio was lower for the high cohesion text. Thus, the misclassification is likely due to relatively low causal cohesion in the high cohesion texts. For *Padria* and *El Niño*, the four variables were in the predicted direction. The misclassification may be because the original texts were already high in cohesion relative to other texts in the corpus.

The misclassification of *Traits of Mammals* is perhaps the most interesting in the classification analysis. All four variables were in the wrong direction, giving the low cohesion version of the text a higher probability of being the high cohesion text than the high cohesion text itself. The reason for the misclassification appears to be the difference between the experimental manipulation and the selected variables. As previously mentioned, the original text was already locally coherent, but lacked global cohesion. Thus, revisions focused on making explicit the links between subtopics and the main topic. However, we presume that none of the

four predictor variables in our model adequately assess global cohesion, presumably resulting in the misclassification.

Turning to misclassifications where both texts were assessed as low cohesion, variables for *Mademoiselle Germaine* were all in the predicted direction and the probability for the genuinely low cohesion text was higher than that for the misclassified high cohesion text. The result suggests that the researchers had greater room to play with in their manipulations. Indeed, given that only causal and temporal aspects were modified, the model's coreference variables were unlikely to have been greatly affected.

Project X-Ray, from the same experiment as *Mademoiselle Germaine* (above), featured word frequency results that were not in the predicted direction. Again, the focus on causal and temporal features rather than content and coreference presumably led to the misclassification.

For *The Return of Martin Guerre*, the original version had its sentence order scrambled. The results were in the predicted order with local coreference indices' values lower for the low-cohesion version. However, also predictable, causal and word frequency values did not and would not have been changed. As such, the misclassification can be attributed to a relatively low in cohesion text being modified to be even lower in cohesion.

For *Peru and Argentina* the variables are in the predicted directions with the exception of causal ratio. Like *The Return of Martin Guerre*, the misclassification appears to be mainly the result of the original text being relatively low in cohesion before it was made even lower.

Discussion

There is a need in discourse psychology for computational techniques to analyze text on levels of cohesion and text difficulty, particularly because discourse psychologists increasingly use longer, naturalistic texts from real-world sources. This need is also increasing as large electronic corpora become more readily available and of interest to researchers. Computational tools have become more available over the last decade, but they are either fragmented over different databases, or they investigate individual features of text. Coh-Metrix offers the ability to automatically assess a wide range of linguistic features in text, and also offers new indices of text cohesion.

Numerous studies have been conducted with Coh-Metrix to identify linguistic features that allow discriminating between various types of text and discourse. However, none of these studies have validated the primary application of Coh-Metrix – to assess the cohesion of text. It is important to confirm that Coh-Metrix significantly distinguishes between high and low cohesion texts (an important benchmark), and it is also important to compare these indices to the more standard readability measures that are more often used to assess text difficulty.

To address the first objective, we reviewed a subset of the studies that have empirically investigated text cohesion (see Appendix A), and conducted a computational linguistic analysis of the texts used in these published studies. We confirmed that the coreference indices provided in Coh-Metrix revealed significant differences between the high and low cohesion texts, showing higher cohesion scores for the high-cohesion than low-cohesion texts. Analyses further indicated that two indices, noun coreference and causal cohesion, were most discriminative between high and low cohesion texts.

At a more detailed level, we examined which of the indices provided greater distinction between the high and low cohesion texts. One of the factors examined was whether the distal coreference measures (beyond adjacent sentences) were affected by weighting the overlap as a function of the distance from the target sentence. We found that the weighted algorithms yielded more discriminative indices than did the unweighted algorithms. These results are compatible with our understanding of limitations of working memory capacity (Just & Carpenter, 1980; Kintsch, 1998). Namely, if distance is to be considered in the computation, it should be weighted to accommodate the constraints of working memory functioning. It is quite possible that the researchers who manipulated the texts did this intentionally or implicitly, based on what they know about working memory and contiguity. Manipulations of coreferential cohesion were intentionally made by most of the researchers who prepared the texts. Thus, the researchers' epistemology on cognitive mechanisms may be reflected in these results. It is noteworthy that the Coh-Metrix indices are sufficiently sensitive to detect this.

LSA indices followed the patterns found for coreference, apart from the paragraph-to-paragraph and paragraph-to-text. Although the LSA indices distinguished between the text versions, the differences were smaller compared to the coreference indices. This difference is most evident in the effect sizes in the final columns of Tables 2 and 3. The average Cohen's *d* effect size for the coreference indices was 0.98 compared to the largest LSA effect size of 0.59. One difference between the coreference indices and the LSA indices is that LSA is more generous in its determination of overlap and the coreference indices are more stringent semantically. Using LSA, a word in a sentence is more likely to overlap with a word in another sentence because LSA overlap is not as strict in determining conceptual overlap, or similarity, as compared to the coreference indices (i.e., word to word). This result is similar to that found for stem overlap, which showed smaller effect sizes (M=0.78) as compared to noun and argument overlap (M=1.01). Stem overlap, like LSA, is more generous because it counts more words as overlapping, and thus there too, the differences between the versions are less distinguishable. Thus, the indices with the strictest indices of overlap tend to show greater differences between versions within this text corpus.

Another explanation for the reduced cohesion effects for the LSA and stem overlap indices may be because the compared texts were on the same topics. That is, the texts were high and low-cohesion versions of the same text. Given that LSA is designed to represent semantic similarity, the smaller differences shown by LSA may reflect the fact that the texts were highly similar semantically. Whereas LSA is apparently not as effective in measuring explicit cohesion differences, it may be more effective in picking up implicit differences in conceptual cohesion.

The text versions also differed in terms of causal connectives, and the ratio of causal connectives to change-of-state verbs. Of particular interest to us was the causal ratio index which is indicative of causal cohesion. This study indicated that the ratio successfully distinguished between high and low-cohesion texts.

In sum, regarding the first purpose of this study, we found that indices of cohesion such as coreference and connectives showed differences between texts in line with expectations. For the second purpose of this paper, comparing Coh-Metrix with readability formulas, we found that the Flesch-Kincaid, a traditional measure of text readability, indicated that lower cohesion texts would be less difficult than the higher cohesion texts, a prediction that was not supported by the

results of these published studies. At the same time, the conflict between cohesion and traditional readability formulas does not come as a surprise. Adding cohesion generally results in longer sentences and consequently higher grade level texts according to Flesch-Kincaid Grade Level indices. Thus, the Flesch-Kincaid values indicating that the higher cohesion texts should be more challenging calls into question one of the core assumptions of readability measures, namely that lengthening the texts makes the texts more difficult. It also indicates that text difficulty may be a result of trade-offs between various aspects of text difficulty. In some cases for the 19 pairs of high and low cohesion texts examined here, specific changes were made intentionally by discourse researchers. In other cases, the differences in indices reflected side-effects or tradeoffs of cohesion manipulations that may or may not have been expected theoretically. Thus, this study brings us closer to better understanding characteristics of text associated with text difficulty.

Our multivariate analysis suggested that Coh-Metrix produces five unique variables that capture the differences between the high and low cohesion texts: coreferential noun overlap, LSA sentence to sentence, causal ratio, word concreteness, and word frequency. Of the these variables, the coreference, LSA, and causal ratio measures are more likely, in terms of face validity, to be considered direct indices of cohesion, whereas word concreteness and word frequency are indices likely related to the side-effects of manipulating cohesion. These latter variables are also important considerations in terms of text difficulty in general.

Noun overlap and LSA are certainly related variables; however, noun overlap appears to capture unique cohesion at the explicitly marked noun level, whereas LSA captures cohesion at the semantic or implicit level. The causal variable appears to capture unique cohesion as a relationship between verbs and particles that explicitly link clauses. And word concreteness and word frequency, although not indices of cohesion *per se*, appear to play a crucial role in the lexical distribution of low and high cohesion texts.

The discriminant analysis informs us of two main points. First, there is an important difference between *low*-cohesion and *lower*-cohesion on the one hand and *high*-cohesion and *higher*-cohesion on the other. That is, all text manipulations in this study are relative to their original, rather than related to a gold standard. We attempt to illustrate what this gold standard might look like via the discriminant analysis; though we have stressed that the restricted number of available texts means that results should be interpreted with caution. Of course, our ultimate goal is to discover the parameters of such a gold standard and the analysis presented here with its related model should help in this endeavor. Second, our analysis also informs us that our model is lacking in at least one critical component: global cohesion. Although Coh-Metrix provides numerous indices that are theoretically global, the constructs may be too closely related to their local counterparts. We are currently developing and testing new measures to address this issue.

A potential weakness of the discriminant analysis is the low level of power (19 pairs of texts). More texts would have certainly afforded firmer conclusions. However, the 19 pairs of texts included in this study reflect the number of studies conducted and the number of texts available to us. In future research, it will be important to verify that Coh-Metrix successfully discriminates variations in cohesion, particularly in comparison to other types of differences between texts. Related to this issue, having established a benchmark test showing that Coh-Metrix can reliably distinguish between low and high cohesion texts, future studies should investigate variations in cohesion in what might be called naturally occurring texts (for which cohesion differences are natural, and not manipulated). At this time, however, there is not sufficient research investigating comprehension differences between high and low cohesion texts where the cohesion differences are *graded* and not manipulated (as we investigated here), and thus there are not sufficient corpora available to conduct such corpora analyses.

The results and analyses offered in this study add compelling evidence to the validation of Coh-Metrix as a tool that assesses cohesion in text. The purpose of this study was not to present a simple “one-stop solution” to cohesion assessment, and neither was Coh-Metrix designed to do so. Coh-Metrix is designed to provide *a vehicle* for cohesion assessment, but not its *chauffeur*. The results presented in this paper provide a validation of Coh-Metrix thereby paving the way for its use by chauffeurs in cognitive science, discourse processes, and education, as well as for textbook writers, professionals in instructional design, and instructors.

References

Allison, P.D. (1999). *Multiple regression*. Thousand Oaks, CA: Pine Forge Press.

Anderson, A., Garrod, S.C.. & Sanford, A.J. (1983). The accessibility of pronominal antecedents as a function of episode shifts in narrative texts. *Quarterly Journal of Experimental Psychology, 35*, 427-440.

Baayen, R.H., Piepenbrock, R., & van Rijn, H. (Eds.) (1993). *The CELEX lexical database*. Philadelphia, PA: University of Pennsylvania Press.

Beck, I.L, McKeown, M.G, Omanson R., & Pople, M. (1984). Improving the comprehensibility of stories: The effects of revisions that improve coherence. *Reading Research Quarterly, 19*, 263-277.

Beck, I. L., McKeown, M.G., Sinatra, G. M., & Loxterman, J.A. (1991). Revising social studies text from a text-processing perspective: Evidence of improved comprehensibility. *Reading Research Quarterly, 26*, 251-276.

Black, J.B., Turner, T.J. & Bower, G.H. (1979). Point of view in narrative comprehension, memory and production. *Journal of Verbal Learning and Verbal Behavior, 18*, 187-198.

Britton, B.K., & Gulgoz, S. (1991). Using Kintsch's computational model to improve instructional text: Effects of repairing inference calls on recall and cognitive structures. *Journal of Educational Psychology, 83*, 329-345.

Britton, B.K., Gulgoz, S., & Glynn, S. (1993). Impact of good and poor writing on learners: Research and theory. In B.K. Britton, A. Woodward, & M.R. Binkley (Eds.), *Learning from textbooks: Theory and practice* (pp. 1-46). Hillsdale, NJ: Erlbaum.

Cataldo, M.G., & Oakhill, J. (2000). Why are poor comprehenders inefficient searchers? An investigation into the effects of text representation and spatial memory on the ability to locate information in text. *Journal of Educational Psychology, 92*, 791-799.

Charniak, E. (2000). A maximum-entropy-inspired parser. *Proceedings of the First Conference on North American Chapter of the Association for Computational Linguistics* (pp. 132-139). San Francisco, CA: Morgan Kaufmann Publishers.

Chall, J.S., & Dale, E. (1995). *Readability revisited.* Cambridge, MA: Brookline.

Coltheart, M. (1981). The MRC psycholinguistic database quarterly. *Journal of Experimental Psychology*, *33*, 497-505.

Connaster, B. F. (1999). Last rites for readability formulas in technical communication. *Journal of technical writing and communication*, *29*, 271-287.

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

Dale, E. & Chall, J.S. (1949). The concept of readability. *Elementary English*, *26*, 19-26.

Dubay, W.H. (2004). The principles of readability. Costa Mesa, CA: Impact Information.

Duffy, T. M. (1985). Readability formulas: What's the use? In T.M. Duffy & R.M. Waller (Eds.), *Designing usable texts* (pp. 113-143). New York, NY: Academic Press.

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

Flesch, R. (1948). *A new readability yardstick*, *Journal of Applied Psychology, 32*, 221-233.

Foltz, P.W., Kintsch, W., & Landauer, T.K. (1998). The measurement of textual coherence with Latent Semantic Analysis. *Discourse Processes, 25,* 285-307.

Fry, E. (1975). *Reading drills for speed and comprehension (2nd ed.)*. Providence, RI: Jamestown Publishers.

Gernsbacher, M.A. (1990). *Language comprehension as structure building*. Hillsdale, NJ: Erlbaum.

Gernsbacher, M.A., & Robertson, R.R.W. (2002). The definite article 'the' as a cue to map thematic information. In W. van Peer, & M. Louwerse (Eds.), *Thematics: Interdisciplinary studies* (pp. 119-136). Amsterdam, Netherlands: Benjamins.

Graesser, A.C., Jeon, M., Cai, Z., & McNamara, D. S. (2008). Automatic analyses of language, discourse, and situation models. In J. Auracher & W. van Peer (Eds.), *New beginnings in literary studies* (pp. 72–88). Cambridge, UK: Cambridge Scholars Publishing.

Graesser, A.C., Jeon, M., Yang, Y., & Cai, Z. (2007). Discourse cohesion in text and tutorial dialogue. *Information Design Journa1, 15*, 199-213.

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

Graesser, A.C., & Morgan, B. (2008). An analysis of Will van Peer's scholarly contributions with an automated text analysis tool called Coh-Metrix. In S. Zyngier, M. Bortolussi, A. Chesnokova, & J. Auracher (Eds.), *Directions in empirical literary studies: Essays in honor of Willie van Peer*. Amsterdam, Netherlands: Benjamins.

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

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

Hall, C., McCarthy, P. M., Lewis, G. A., Lee, D. S., & McNamara, D. S. (2007). A Coh-Metrix assessment of American and English/Welsh Legal English. *Coyote Papers: Psycholinguistic and Computational Perspectives. University of Arizona Working Papers in Linguistics*, *15,* 40–54.

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

Just, M.A. & Carpenter, P.A. (1980). A theory of reading: From eye fixations to comprehension. *Psychological Review*, *87*, 329-354.

Kinnear, P.R. & Gray, C.D. (2008). *SPSS 15 made simple*. New York, NY: Psychology Press.

Kintsch, E. (1990). Macroprocesses and microprocesses in the development of summarization skill. *Cognition and Instruction, 7*, 161-195.

Kintsch, W. (1998). *Comprehension: A paradigm for cognition*. Cambridge, MA: Cambridge University Press.

Kintsch, W., & van Dijk, T.A. (1978). Toward a model of text comprehension and production. *Psychological Review, 85,* 363-394.

Klare, G. R. (1974-1975). Assessing readability. *Reading Research Quarterly, 10,* 62-102.

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

Kreft, I., de Leeuw, J., & Aiken, L.S. (1995). The effect of different forms of centering in hierarchical linear models. *Multivariate Behavioral Research 30,* 1-21.

Landauer, T.K., & Dumais, S.T. (1997). A solution to Plato's problem: The Latent Semantic Analysis theory of the acquisition, induction, and representation of knowledge. *Psychological Review, 104*, 211-240.

Landauer, T.K., McNamara, D.S, Dennis, S., & Kintsch, W. (Eds.). (2007). *LSA: A road to meaning.* Mahwah, NJ: Erlbaum.

Leech, N.L., Barrett, K.C., & Morgan, G.A. (2008). *SPSS for intermediate statistics: Use and interpretation*. Mahwah, NJ: Erlbaum.

Lehman, S., & Schraw, G. (2002). Effects of coherence and relevance on shallow and deep text processing. *Journal of Educational Psychology, 94*, 738-750.

Lightman, E. J., McCarthy, P. M., Dufty, D. F., & McNamara, D. S. (2007). The structural organization of high school educational texts. In D. Wilson & G. Sutcliffe (Eds.), *Proceedings of the Twentieth International Florida Artificial Intelligence Research Society (FLAIRS) Conference* (pp. 235–240). Menlo Park, CA: The AAAI Press.

Linderholm, T. Everson, M.G., van den Broek, P., Mischinski, M., Crittenden, A., & Samuels, J. (2000). Effects of causal text revisions on more and less skilled readers? Comprehension of easy and difficult text. *Cognition and Instruction, 18*, 525-556.

Lorch, R.F. & Lorch, E.P. (1996). Effects of headings on text recall and summarization. *Contemporary Educational Psychology, 21*, 261-278.

Louwerse, M. (2002). An analytic and cognitive parameterization of coherence relations. *Cognitive Linguistics*, *12*, 291–315.

Louwerse, M. M., McCarthy, P. M., McNamara, D. S., & Graesser, A. C. (2004). Variation in language and cohesion across written and spoken registers. In K. Forbus, D. Gentner, & T. Regier (Eds.), *Proceedings of the Twenty-Sixth Annual Conference of the Cognitive Science Society* (pp. 843–848). Mahwah, NJ: Erlbaum.

Loxterman, J.A., Beck, I.L., & McKeown, M.G. (1994). The effects of thinking aloud during reading on students' comprehension of more or less coherent text. *Reading Research Quarterly, 29*, 353-367.

Manzo, A. (1970). Readability: A postscript. *Elementary English, 47*, 962-965.

Maxwell, M. (1978). Readability: Have we gone too far? *Journal of reading, 21*, 525-530.

McCarthy, P. M., Briner, S. W., Rus, V., & McNamara, D.S. (2007). Textual signatures: Identifying text-types using latent semantic analysis to measure the cohesion of text structures. In A. Kao & S. Poteet (Eds.), *Natural language processing and text mining* (pp. 107–122). London: Springer-Verlag UK.

McCarthy, P. M., Lewis, G. A., Dufty, D. F., & **McNamara**, D. S. (2006). Analyzing writing styles with Coh-Metrix. In G.C.J. Sutcliffe & R.G. Goebel (Eds.), *Proceedings of the 19th Annual Florida Artificial Intelligence Research Society International Conference (FLAIRS)* (pp. 764–770). Melbourne Beach, FL: AAAI Press.

McNamara, D.S. (2001). Reading both high and low coherence texts: Effects of text sequence and prior knowledge. *Canadian Journal of Experimental Psychology, 55*, 51-62.

McNamara, D.S., Cai, Z., & Louwerse, M.M. (2007). Optimizing LSA measures of cohesion. In T.K. Landauer, D.S. McNamara, S. Dennis, & W. Kintsch (Eds.), *Handbook of latent semantic analysis* (pp. 379-400). Mahwah, NJ: Erlbaum.

McNamara, D.S., & Kintsch, W. (1996). Learning from text: Effects of prior knowledge and text coherence. *Discourse Processes, 22*, 247-288.

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

Meyer, B.J.F., & Poon, L.W. (2001). Effects of structure strategy training and signaling on recall of text. *Journal of Educational Psychology, 93*, 141-159.

Miller, G.A., Beckwith, R., Fellbaum, C., Gross, D. & Miller, K. (1990). *Five papers on WordNet*. Princeton, NJ: Princeton University Press.

Miller, J. & Kintsch, W. (1980). Readability and recall of short prose passages: A theoretical analysis. *Journal of Experimental Psychology: Human Learning and Memory, 6*, 335-354.

Myers, J.L., Shinjo, M. & Duffy, S.A. (1987). Degrees of causal relatedness in memory. *Journal of Verbal Learning and Verbal Behavior, 26*, 453-465.

National Center for Education Statistics (2005). National assessment of educational progress: The nation's report card. Retrieved September 2006 from http://nces.ed.gov/nationsreportcard/.

Ohtsuka, K. & Brewer, W.F. (1992). Discourse organization of temporal order in narrative texts. *Discourse Processes, 15*, 317-336.

O'Reilly, T., & McNamara, D.S. (2007). Reversing the reverse cohesion effect: good texts can be better for strategic, high-knowledge readers. *Discourse Processes*, *43*, 121-152.

Sanders, T. J. M., Spooren, W. P. M., & Noordman, L. G. M. (1992). Toward a taxonomy of coherence relations. *Discourse Processes, 15*, 1-35.

Shadish, W.R., Robinson, L., & Lu, C. (1999). *ES: A computer program and manual for effect size calculation.* St. Paul, Minnesota: Assessment Systems Corporation.

Stenner, A. J. (1996). *Measuring reading comprehension with the lexile framework*. Durham, NC: MetaMetrics, Inc.

Snow, C. (Ed.). (2002). *Reading for understanding*. Santa Monica, CA: RAND.

Trabasso, T., Secco, T., & van den Broek, P. (1984). Causal cohesion and story coherence. In H. Mandl, N.L. Stein, & T. Trabasso (Eds.), *Learning and comprehension of text* (pp. 83-111). Hillsdale, NJ: Erlbaum.

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

van Oostendorp, H., Otero, J., & Campanario, J.M. (2002). Conditions of updating. In M. Louwerse & W. van Peer (Eds.), *Thematics. Interdisciplinary studies* (pp. 55-76). Philadelphia, PA: John Benjamins Publishing Company.

Vidal-Abarca, E., Martinez, G., & Gilabert, R. (2000). Two procedures to improve instructional text: Effects on memory and learning. *Journal of Educational Psychology, 92,* 107-116.

Voss, J.F., & Silfies, N.L. (1996). Learning from history text: The interaction of knowledge and comprehension skill with text structure. *Cognition and Instruction, 14,* 45-68.

Yekovich, F.R. & Walker, C. (1978). Identifying and using referents in sentence comprehension. *Journal of Verbal Learning and Verbal Behavior, 17*, 265-277.

Zipf, G. K. (1949). *Human behavior and the principle of least-effort*. Cambridge, MA: Addison-Wesley.

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

Appendix A. List of 29 published studies on cohesion. The 12 studies included in the Experiment 1 corpus analysis are marked with an asterisk.

1. Marshal and Glock (1978-1979)
2. Reder and Anderson (1982)
3. Schwartz and Flammer (1981)
4. Pepper (1981)
5. Swaney, Janik, Bond, and Hayes (1991; original report in 1981)
6. Phiefer, McNickle, Ronning, and Glover (1983)
7. Loman and Mayer (1983)
8. Roen and Piche (1984)
9. **Beck, McKeown, Omanson, and Pople (1984)***
10. Slater (1985)
11. Baumann (1986)
12. Brennan, Bridge, and Winograd (1986)
13. Britton, Van Dusen, Gulgoz, and Glynn (1989)
14. Hidi and Baird (1988)
15. Duffy, Higgins, Mehlenbacher, Cochran, Wallace, Hill, Haugen, McCaffery, Burnett, Sloane, and Smith (1989)
16. **E. Kintsch (1990)***
17. **Britton and Gulgoz (1991)***
18. **Beck, McKeown, Sinatra, and Loxterman (1991)***
19. **Loxterman, Beck, and McKeown (1994)***
20. McNamara and W. Kintsch (1996)
21. **McNamara, E. Kintsch, Songer, and W. Kintsch (1996)***
22. **Voss and Silfies (1996)***
23. R. Lorch and E. Lorch (1996)
24. **Cataldo and Oakhill (2000)***
25. **Vidal-Abarca, Martinez, and Gilabert (2000)***
26. **Linderholm, Everson, van den Broek, Mischinski, Crittenden, and Samuels (2000)***
27. **McNamara (2001)***
28. Meyer and Poon (2001)
29. **Lehman and Schraw (2002)***

Notes: * The bolded studies with a * indicates that the text(s) used in the study are included in the analysis. Study 20 (McNamara & Kintsch, 1996) was excluded because it used the text used in the Britton and Gulgoz (1991). Study 23 (R. Lorch and E. Lorch, 1996) and study 28 (Meyer & Poon, 2001) were excluded because the text revisions were limited in nature (changes in format, adding headers).

Appendix B. Summaries of the 12 studies, and 19 text pairs analyzed, including, authors, participants, text titles, a brief summary of the text revisions and the results for which effect sizes could be computed.

Authors	Participants	Text Title(s)	Cohesion Revisions	Results
Beck, McKeown, Omason, and Pople (1984)	Students Grade 3	The Raccoon and Mrs. McGinnis	Alleviated surface, knowledge, and content problems.	**Sub-Group** / **Measure** / **Effect Size**: Skilled Readers — Recall 1.56; Multiple Choice 0.58. Less Skilled Readers — Recall 0.63; Multiple Choice 0.60
E. Kintsch (1990)	College Students; Students Grade 6 and 10	Peru and Argentina	The originally high cohesion text (MM) is compared to the revision (mm) with frequent topic shifts, more difficult words, longer and more complex sentences, and fewer connectives.	**Sub-Group** / **MeasureS: Summary Quality** / **Effect Size**: All Participants — # Propositions -0.14; Generalizations 0.04; Elaborations 0.00; Re-ordering -0.46; Connectives 0.77
Beck, McKeown, Sinatra, and Loxterman (1991)	Students Grade 4-5	1) The French and Indian War 2) Boston Tea Party 3) Intolerable Acts 4) No Taxation without Representation	Minimized need for knowledge based inferences and increased causal cohesion.	**Text** / **Measure** / **Effect Size**: Text 1 (War) — Recall 0.32; Open-Ended Q 0.75. Text 2 (Tea) — Recall 0.49; Open-Ended Q 0.63. Text 3 (Acts) — Recall 0.41; Open-Ended Q 0.72. Text 4 (Tax) — Recall 0.35; Open-Ended Q 0.68

<table>
<tr>
<td>Britton and Gulgoz (1991)</td>
<td>College Students</td>
<td>Air War in the North</td>
<td>In principled revision, cohesion breaks identified, then added argument overlap, rearranged clauses (old ideas first), and made implicit references explicit.</td>
<td>
<table>
<tr><th>Sub-Group</th><th>Measure:</th><th>Effect Size</th></tr>
<tr><td>All Participants</td><td>Free Recall</td><td>1.07</td></tr>
<tr><td></td><td>Recall Efficiency*</td><td>0.61</td></tr>
<tr><td></td><td>MC Accuracy</td><td>0.77</td></tr>
<tr><td></td><td>MC Efficiency*</td><td>- 0.10</td></tr>
<tr><td></td><td>Connectives</td><td>1.07</td></tr>
</table>
Notes: MC is Multiple-Choice; *efficiency = performance/RT
</td>
</tr>
<tr>
<td>Loxterman, Beck, and McKeown (1994)</td>
<td>Students Grade 3</td>
<td>El Nino</td>
<td>Increased causal coherence and explicit connections between events.</td>
<td>
Experiment 1
<table>
<tr><th>Sub-Group</th><th>Measure</th><th>Effect Size</th></tr>
<tr><td>Silent</td><td>Recall</td><td>0.81</td></tr>
<tr><td></td><td>Open-Ended Q</td><td>1.20</td></tr>
<tr><td>Think Aloud</td><td>Recall</td><td>0.57</td></tr>
<tr><td></td><td>Open-Ended Q</td><td>1.21</td></tr>
</table>
Experiment 2
<table>
<tr><th>Sub-Group</th><th>Measure</th><th>Imd Effect Size</th><th>Del Effect Size</th></tr>
<tr><td>High Skill Silent</td><td>Recall</td><td>0.92</td><td>1.35</td></tr>
<tr><td></td><td>Open-Ended Q</td><td>0.91</td><td>1.39</td></tr>
<tr><td>High Skill Think Aloud</td><td>Recall</td><td>0.28</td><td>1.03</td></tr>
<tr><td></td><td>Open-Ended Q</td><td>1.66</td><td>2.07</td></tr>
<tr><td>Intermediate Skill Silent</td><td>Recall</td><td>0.91</td><td>1.28</td></tr>
<tr><td></td><td>Open-Ended Q</td><td>1.14</td><td>1.27</td></tr>
<tr><td>Intermediate Skill Think Aloud</td><td>Recall</td><td>0.84</td><td>1.19</td></tr>
<tr><td></td><td>Open-Ended Q</td><td>1.71</td><td>1.53</td></tr>
</table>
Notes: Imd is immediate test, Del is 1 week delayed test; Skill refers to reading skill
</td>
</tr>
</table>

<table>
<tr>
<td>McNamara, E. Kintsch, Songer, and W. Kintsch (1996)</td>
<td>Students Grade 7-9</td>
<td>Trait of Mammals</td>
<td>Principled revision made links explicit between subtopics and main topic.</td>
<td>
<table>
<tr><th>Sub-Group</th><th>Measure:</th><th>Effect Size</th></tr>
<tr><td>All Participants</td><td>Overall recall</td><td>0.77</td></tr>
<tr><td></td><td>Macro recall</td><td>1.52</td></tr>
<tr><td></td><td>Micro recall</td><td>0.24</td></tr>
<tr><td></td><td>Open-ended Q</td><td>1.14</td></tr>
<tr><td></td><td>Keyword Sorting</td><td>1.24</td></tr>
</table>
</td>
</tr>
<tr>
<td></td>
<td>Students Grade 7-10</td>
<td>Heart Disease</td>
<td>The high cohesion text included micro and macro level changes, including reducing anaphor, increasing argument overlap, adding descriptive elaborations; adding sentence connectives, adding topic headers, and topic sentences.</td>
<td>
<table>
<tr><th>Sub-Group</th><th>Measure</th><th>Effect Size</th></tr>
<tr><td>High Knowledge</td><td>Recall</td><td>0.48</td></tr>
<tr><td></td><td>Open-Ended Q (all)</td><td>-0.40</td></tr>
<tr><td></td><td>Open-Ended P.Solving</td><td>-0.63</td></tr>
<tr><td></td><td>Open-Ended Bridging</td><td>-0.83</td></tr>
<tr><td></td><td>Keyword Sorting</td><td>-1.00</td></tr>
<tr><td>Low Knowledge</td><td>Recall</td><td>0.49</td></tr>
<tr><td></td><td>Open-Ended Q (all)</td><td>0.93</td></tr>
<tr><td></td><td>Open-Ended P.Solving</td><td>0.55</td></tr>
<tr><td></td><td>Open-Ended Bridging</td><td>0.37</td></tr>
<tr><td></td><td>Keyword Sorting</td><td>1.33</td></tr>
</table>
</td>
</tr>
<tr>
<td>Voss and Silfies (1996)</td>
<td>College Students</td>
<td>1) Anchad
2) Padria</td>
<td>Increased elaboration of causal factors related to events described in the texts.</td>
<td>Could not be computed because data provided in the article were all correlational.</td>
</tr>
</table>

<table>
<tr>
<td>Cataldo and Oakhill (2000)</td>
<td>Students Grade 5</td>
<td>1) The Demon Barber
2) The Return of Martin Guerre</td>
<td>Reordered the original sentences to create a scrambled version of the text. Reduced local cohesion.</td>
<td>
<table>
<tr><th>Sub-Group</th><th>Measure</th><th>Effect Size</th></tr>
<tr><td rowspan="5">Skilled Readers</td><td>Recall</td><td>1.02</td></tr>
<tr><td>Open-Ended Q (bf srch)</td><td>1.53</td></tr>
<tr><td>Open-Ended Q (aft srch)</td><td>0.87</td></tr>
<tr><td>Search time</td><td>-1.16</td></tr>
<tr><td>Keyword Spatial Acc</td><td>-0.08</td></tr>
<tr><td></td><td>Keyword Seq. Acc</td><td>1.06</td></tr>
<tr><td rowspan="6">Less Skilled Readers</td><td>Recall</td><td>1.23</td></tr>
<tr><td>Open-Ended Q (bf srch)</td><td>0.29</td></tr>
<tr><td>Open-Ended Q (aft srch)</td><td>1.45</td></tr>
<tr><td>Search time</td><td>0.87</td></tr>
<tr><td>Keyword Spatial Acc</td><td>0.94</td></tr>
<tr><td>Keyword Seq. Acc</td><td>0.48</td></tr>
</table>
</td>
</tr>
<tr>
<td>Linderholm, Everson, van den Broek, Mischinski, Crittenden, and Samuels (2000)</td>
<td>College Students</td>
<td>1) Project X-Ray
2) Mademoiselle Germaine</td>
<td>Repaired causal structure of the text, including arranging text events in temporal order, making implicit goals of the character explicit, repairing cohesion breaks caused by inadequate explanation, multiple causality, or distant causal relations.</td>
<td>
<table>
<tr><th>Sub-Group</th><th>Measure</th><th>Easy Text Effect Size</th><th>Diff Text Effect Size</th></tr>
<tr><td rowspan="2">Skilled Readers</td><td>Recall</td><td>-1.00</td><td>0.67</td></tr>
<tr><td>Open-Ended Q Total</td><td>-0.59</td><td>1.46</td></tr>
<tr><td rowspan="2">Less Skilled Readers</td><td>Recall</td><td>0.01</td><td>0.33</td></tr>
<tr><td>Open-Ended Q Total</td><td>0.01</td><td>1.00</td></tr>
</table>
Notes: Diff Text is Difficult Text
</td>
</tr>
</table>

<table>
<tr>
<td>Vidal-Abarca, Martinez, and Gillabert (2000)</td>
<td>College Students</td>
<td>The Russian Revolution</td>
<td>Maximally cohesive version included argument overlap and repaired causal breaks by adding information to trigger the reader's causal antecedents and super-ordinate goal inferences.</td>
<td>
<table>
<tr><th>Sub-Group</th><th>Measure</th><th>Imd Effect Size</th><th>Del Effect Size</th></tr>
<tr><td>All Participants</td><td>Recall: Main Idea</td><td>0.90</td><td>0.92</td></tr>
<tr><td></td><td>Recall Supporting Inf</td><td>0.29</td><td>0.09</td></tr>
<tr><td></td><td>Inference Test</td><td></td><td>0.92</td></tr>
</table>
Notes: Inference test is open-ended inference questions answered while referring back to the text.
</td>
</tr>
<tr>
<td>McNamara (2001)</td>
<td>College Students</td>
<td>Cell Division</td>
<td>Reduced anaphor, increased argument overlap, added elaborations; added sentence connectives, added topic headers, and topic sentences.</td>
<td>
<table>
<tr><th>Sub-Group</th><th>Measure</th><th>Effect Size</th></tr>
<tr><td>High Knowledge</td><td>Textbased Q</td><td>-0.94</td></tr>
<tr><td></td><td>Bridging Inference Q</td><td>-0.19</td></tr>
<tr><td>Low Knowledge</td><td>Textbased Q</td><td>0.78</td></tr>
<tr><td></td><td>Bridging Inference Q</td><td>0.00</td></tr>
</table>
Note: Questions are Open-Ended.
</td>
</tr>
<tr>
<td>Lehman and Schraw (2002)</td>
<td>College Students</td>
<td>The Quest for Northwest Passage</td>
<td>Experiment 1
Altered order of sentences that promoted referential or causal coherence if they could be relocated without altering meaning.</td>
<td>
<table>
<tr><th>Sub-Group</th><th>Measure:</th><th>Effect Size</th></tr>
<tr><td>Low Relevance</td><td>Recognition</td><td>0.38</td></tr>
<tr><td></td><td>Recall</td><td>0.16</td></tr>
<tr><td></td><td>Essay Situation Model</td><td>0.13</td></tr>
<tr><td></td><td>Essay Claim</td><td>0.11</td></tr>
<tr><td></td><td>Ease of Comprehension</td><td>0.87</td></tr>
<tr><td>High Relevance</td><td>Recognition</td><td>-0.01</td></tr>
<tr><td></td><td>Recall</td><td>-0.05</td></tr>
<tr><td></td><td>Essay Situation Model</td><td>0.15</td></tr>
<tr><td></td><td>Essay Claim</td><td>0.16</td></tr>
<tr><td></td><td>Ease of Comprehension</td><td>0.88</td></tr>
</table>
</td>
</tr>
</table>

			Experiment 2 Interrupted the temporal flow of the story to reduce global coherence. Revision organized thematically rather than chronologically.	**Sub-Group**	**Measure:**	**Effect Size**
				Low Relevance	Recognition	0.13
					Recall	0.37
					Essay Situation Model	-0.20
					Essay Claim	-0.26
					Ease of Comprehension	0.66
				High Relevance	Recognition	0.53
					Recall	0.78
					Essay Situation Model	0.69
					Essay Claim	0.80
					Ease of Comprehension	1.15

Author Notes

The research was supported in part by the Institute for Education Sciences (IES R305G020018-02). Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the IES. We are grateful to Yasuhiro Ozuru, who collected the texts analyzed in this study, and to Zhiqiang Cai, who is the lead programmer on the Coh-Metrix project. Coh-Metrix is a collaborative endeavor; hence we are also grateful to all of the members of the Coh-Metrix project who contributed to this study in various ways.

Table 1
Coh-Metrix indices as a function of low and high cohesion text versions

	Low Cohesion	High Cohesion				
	Mean (SD)	Mean (SD)	Diff	*F*(1,18)	p	Effect
Descriptive Indices						
Number words	507.32 (326)	673.05 (424)	165	17.08	0.001	0.44
Number sentences	36.26(19.16)	41.68 (23.35)	5.42	7.09	0.016	0.25
Number paragraphs	10.84 (9.66)	10.58 (8.28)	-0.26	0.06	*ns*	0.03
Words per sentences	13.50 (3.97)	15.78 (3.73)	2.28	18.21	<.001	0.59
Sentences per paragraph	3.90 (1.23)	4.19 (1.14)	0.29	2.06	*ns*	0.24
Flesch-Kincaid Grade	7.76 (3.04)	8.35 (2.82)	0.59	4.64	0.045	0.20
Flesch Reading Ease	62.91 (16.81)	61.57 (16.86)	-1.33	0.64	*ns*	0.08
Syllables per word	1.54 (0.17)	1.53 (0.16)	-0.01	0.36	*ns*	0.06
Celex Log Word Frequency	1.01 (0.26)	0.87 (0.22)	-0.14	17.49	0.001	0.58
Content Word Concreteness	388.63 (21.20)	385.05 (22.58)	-3.58	4.49	0.048	0.16
LSA Indices						
Adjacent Sentence to Sentence	0.205 (0.105)	0.270 (0.116)	0.064	15.90	0.001	0.59
Sentence to All Sentences	0.186 (0.101)	0.239 (0.106)	0.052	9.24	0.007	0.51
Sentence to Paragraph	0.268 (0.136)	0.333 (0.121)	0.065	12.28	0.003	0.51
Sentence to Text	0.337 (0.127)	0.372 (0.131)	0.036	8.85	0.008	0.27
Paragraph to Paragraph	0.356 (0.197)	0.356 (0.192)	0.001	0.00	*ns*	0.00
Paragraph to Text	0.502 (0.189)	0.516 (0.193)	0.013	0.55	*ns*	0.07
Connectives Incidence						
Causal	21.40 (7.78)	28.57 (15.63)	7.17	5.60	0.029	0.61
Additive	39.64 (13.31)	36.25 (11.10)	-3.39	1.39	*ns*	0.28
Temporal	10.68 (6.65)	11.88 (5.14)	1.20	1.25	*ns*	0.20
Clarification	0	0.37 (0.99)	0.37	2.64	*ns*	0.75
All Connectives	69.29 (17.20)	73.26 (13.20)	3.97	0.86	*ns*	0.26
Causal Indices						
COS Verbs Incidence	25.21 (12.25)	24.10 (10.42)	-1.11	1.05	*ns*	0.10
Causal Ratio	0.87 (0.39)	1.14 (0.46)	0.27	10.69	0.004	0.64

Notes: standard deviations are in parentheses; Diff refers to the difference between the high and low cohesion texts; Effect refers to effect sizes using Cohen's *d*; COS verbs refers to change-of-state verbs, Causal Ratio refers to the Causal Connective to COS Verb Ratio

Table 2

Coreference indices by cohesion (*high, low) and by type of index as a function of the type of index* (*noun, argument, stem), distance* (*all distances, 2 sentences, 3 sentences) and weight* (*unweighted, weighted)*

			Cohesion				
			Low	High			
Type	Distance	Weight	Mean (SD)	Mean (SD)	Diff	*F*(1,18)	Effect
Noun	Adjacent	Unwtd	0.340 (0.194)	0.532 (0.160)	0.192	23.65	1.08
	2 Sent	Unwtd	0.316 (0.162)	0.473 (0.143)	0.158	27.06	1.03
		Wtd	0.324 (0.171)	0.493 (0.147)	0.169	26.31	1.06
	3 Sent	Unwtd	0.301 (0.151)	0.444 (0.133)	0.143	25.89	1.01
		Wtd	0.314 (0.163)	0.474 (0.140)	0.159	26.48	1.06
	All Dist	Unwtd	0.225 (0.098)	0.329 (0.134)	0.105	18.62	0.90
		Wtd	0.269 (0.124)	0.401 (0.133)	0.132	23.59	1.03
Argument	Adjacent	Unwtd	0.396 (0.197)	0.575 (0.148)	0.180	19.43	1.04
	2 Sent	Unwtd	0.375 (0.163)	0.525 (0.135)	0.150	20.59	1.01
		Wtd	0.382 (0.173)	0.542 (0.138)	0.160	20.70	1.03
	3 Sent	Unwtd	0.358 (0.152)	0.500 (0.124)	0.142	22.22	1.03
		Wtd	0.372 (0.164)	0.526 (0.131)	0.154	21.67	1.04
	All Dist	Unwtd	0.275 (0.102)	0.378 (0.138)	0.103	16.84	0.86
		Wtd	0.324 (0.128)	0.451 (0.127)	0.127	19.72	1.00
Stem	Adjacent	Unwtd	0.448 (0.221)	0.608 (0.160)	0.160	14.83	0.84
	2 Sent	Unwtd	0.421 (0.187)	0.556 (0.154)	0.134	18.03	0.79
		Wtd	0.431 (0.197)	0.573 (0.155)	0.143	17.11	0.81
	3 Sent	Unwtd	0.402 (0.174)	0.532 (0.147)	0.130	24.14	0.81
		Wtd	0.418 (0.186)	0.558 (0.150)	0.139	20.38	0.83
	All Dist	Unwtd	0.317 (0.126)	0.407 (0.154)	0.090	17.63	0.64
		Wtd	0.368 (0.152)	0.482 (0.148)	0.114	18.25	0.76

Notes: All $p < .001$; standard deviations are in parentheses; Diff refers to the difference between the high and low cohesion texts, with positive differences indicative of greater cohesion for the high-cohesion text version; Unwtd is unweighted and wtd is weighted; Effect refers to effect sizes using Cohen's *d*.

Table 3

Actual and predicted group membership based on the discriminant analysis

Actual		Predicted		
		Low Cohesion	High Cohesion	Total
Number	Low Cohesion	14 (0.74)	5 (0.26)	19
	High Cohesion	4 (0.21	15 (0.79)	19

Note: Proportion of the total is provided in parentheses.

Table 4

Tests of equality of group means for the five predictor variables and Fisher's coefficients for the dimensions of low and high cohesion

	Tests of Equality			Fisher's Coefficients	
Coh-Metrix Variable	Wilks' Lambda	*F(1,36)*	*P*	*Low Cohesion*	*High Cohesion*
Log Word Frequency	0.92	3.26	0.08	0.27	-0.27
Content Word Concreteness	0.99	.25	0.62	-0.19	0.19
LSA Adjacent Sentence to Sentence	0.92	3.21	0.08	-0.13	-0.13
Noun Coreference (Adjacent Unweighted)	0.76	11.10	< 0.01	-0.69	0.69
Causal Ratio	0.90	3.94	0.05	-0.52	0.52
Constant				-0.99	-0.99