Title

Quantifying Text Difficulty with Automated Indices of Cohesion and Semantics

Permalink

https://escholarship.org/uc/item/62114025

Journal

Proceedings of the Annual Meeting of the Cognitive Science Society, 29(29)

ISSN

1069-7977

Authors

Duran, Nicholas D.
Bellissens, Cedrick
Taylor, Roger S.
et al.

Publication Date

2007

Peer reviewed

Quantifying Text Difficulty with Automated Indices of Cohesion and Semantics

Nicholas D. Duran (nduran@memphis.edu)
Cedrick Bellissens (cbellissens@memphis.edu)
Roger S. Taylor (rstaylor@memphis.edu)
Danielle S. McNamara (dsmcnamr@memphis.edu)

Institute for Intelligent Systems
Department of Psychology
202 Psychology Building
Memphis, TN 38152 USA

Abstract

We evaluated the effectiveness of new indices of text comprehension in measuring relative text difficulty. Specifically, we examined the efficacy of automated indices produced by the web-based computational tool Coh-Metrix. In an analysis of 60 instructional science texts, we divided texts into groups that were considered to be more or less difficult to comprehend. The defining criteria were based on Coh-Metrix indices that measure independent factors underlying text coherence: *referential overlap* and *vocabulary accessibility*. In order to validate the text difficulty groups, participants read and recalled two “difficult” and two “easy” texts that were similar in topic and length. Easier texts facilitated faster reading times and better recall compared to difficult texts. We discuss the implications of these results in the context of theoretically motivated techniques for improving textbook selection.

Keywords: text difficulty; readability; textbooks; natural language processing; Coh-Metrix; cohesion; semantics.

Introduction

For many students, learning largely depends on information acquired from textbooks. Consequently, educators are often faced with the daunting challenge of selecting texts that are at the appropriate level for their student’s learning ability. A text that is either too difficult or too easy can adversely affect comprehension and hinder academic progress. The challenges of selecting appropriate texts are also compounded by the vast amount of material available to educators, thus making a thorough assessment of each text virtually impossible (Shnick, 2000). Fortunately, educators have long had the assistance of standardized formulas that assess the “readability” (or difficulty) level of instructional texts (Hiebert, 2002). There are nearly 200 *readability formulas* available, all of which track simple linguistic features that serve as proxies of syntactic and semantic difficulty. One of the best-known readability formulas, the Flesch-Kincaid Reading Ease formula (Klare, 1974, 1975), provides a simple technique that is based on the average number of syllables per text, as well as the average length of all sentences. The result is a single index of difficulty, expressed on a straightforward scale of 1-100 (higher scores indicating easier texts).

While readability formulas are easy to use, they lack the sophistication of current theories of cognition. Indeed, readability formulas have remained tied to superficial aspects of language, despite reading comprehension research that has demonstrated that text difficulty is more aptly gauged by deep-structure features related to text cohesion and semantic information (Davison & Kantor, 1982). Both cohesion and semantic features are important textual constructs that positively correlate with the psychological constructs of text coherence and difficulty (McNamara & Kintsch, 1996; Stahl et al., 1989). A cohesive text passage, for example, explicitly links linguistic elements (e.g., constituents, propositions) that help readers in generating inferences and bridge conceptual gaps, thereby improving text comprehension (McNamara, 2001). Additionally, the semantic information in a text (e.g., ambiguity, word frequency) also facilitates comprehension by activating reader’s prior knowledge of the text topic (Graves et al., 1991).

The purpose of this study is to manipulate groups of mutually exclusive features of cohesion and semantics to create an automated technique for identifying levels of text difficulty. We focus on factors of cohesion and semantics that are hypothesized to underlie the difficulty of a text, namely, referential overlap and vocabulary accessibility. (Freebody & Anderson, 1983; Stahl et al., 1989). The *referential overlap* factor of cohesion is an approximation of conceptual redundancy that increases relatedness between sentences. The presence of referential overlap is typically established by the repetition of lexical items, such as pronouns, common nouns, and noun-phrases. Ideally, texts with high referential overlap allow readers to easily integrate content into a coherent mental representation (McNamara & Kintsch, 1996). As a result, an integrated mental representation influences long-term retention and recall of the text (van Oostendorp et al., 1999).

The *vocabulary accessibility* factor of semantic information is the word-level information that varies in familiarity, ambiguity, and abstractness. These word characteristics are particularly important in influencing the activation of concepts from memory while reading (Paivio, 1969). Accordingly, texts with high vocabulary accessibility are usually easier to process and understand because the text vocabulary is easily retrievable and therefore more apparent (Freebody & Anderson, 1983).

Both referential overlap and vocabulary accessibility have been used extensively to improve the comprehensibility of instructional texts (e.g., Graves et al., 1991). By adding or deleting the corresponding linguistic features to a text, difficult and easy versions can be validated with human comprehension norms. Unfortunately, these text revisions are usually time-consuming and require a great deal of experimenter training. The additional constraints imposed by text revisions are particularly disadvantageous considering the popularity of readability formulas. Educators who use readability formulas may do so because they prefer the practical advantage of quick and easy techniques to evaluate text difficulty. However, these same educators may risk an evaluation that is tangentially related to factors underlying text coherence. In this study, we attempt to balance the trade-off between established theory of text difficulty and automaticity of evaluation. In order to do so, we take advantage of advances in computational linguistics that reliably generate comprehensive profiles of language and cohesion. At the forefront of the computational techniques is a web-based software tool called Coh-Metrix (Graesser et al., 2004). Coh-Metrix is particularly useful as it provides a multivariate analysis of the linguistic features that index referential overlap and vocabulary accessibility. Using a subset of these indices, we attempt to uncover difficult and easy texts within a large corpus of naturalistic science texts.

Using Coh-Metrix

Coh-Metrix harnesses sophisticated developments in computational linguistics and discourse processing, featuring advanced syntactic parsers, part-of-speech taggers, distributional models, and psycholinguistic databases. These modules are integrated into the automated Coh-Metrix tool and used to generate over 400 indices of language, text, and readability. Coh-Metrix has been involved in many research endeavors, ranging from learning assessment (Best et al., 2005) to text classification (Louwerse et al., 2004). These successful applications allow us to proceed with confidence in our current analysis of using linguistic features in identifying psychological differences of text difficulty.

Two sets of Coh-Metrix indices were selected that capture the text difficulty dimensions of referential overlap and vocabulary accessibility. A summary of the Coh-Metrix technique for computing these text difficulty variables is provided in the following sections.

Indices of Coreference Coh-Metrix tracks four major types of lexical co-reference: *common noun overlap*, *pronoun overlap*, *argument overlap*, and *stem overlap*. Common noun and pronoun overlap is a proportion of all sentence pairs that share one or more common nouns or pronouns. Argument overlap is a proportion of all sentence pairs that share one or more nouns with a common stem, whereas stem overlap is the proportion of sentence pairs that share one or more words (of any grammatical category) with a common stem.

Coh-Metrix also assesses the contextual similarity between sentences by adopting a computational model called Latent Semantic Analysis (LSA; Landauer & Dumais, 1997). For LSA, similarity is defined as the likelihood that any group of words will occur in the same context in the language environment. Contexts are derived from a large corpus of texts, and each context can range from the sentence, paragraph, or document level. LSA computes word meaning by populating a large word X context co-occurrence matrix based on the number of times word W_i appears in context C_j. Words, now reinterpreted as vector representations, are projected into high-dimensional space and compared using the cosine between the vectors.

In this study, we used a combination of 30 LSA and overlap measures (as calculated by Coh-Metrix) to represent the various aspects of referential overlap.

Indices of Vocabulary Accessibility Coh-Metrix computes word-level information that varies on four conceptual dimensions: *meaningfulness*, *concreteness*, *imagability*, and *familiarity*. These indices are based on human ratings of over 150,000 words compiled in the MRC database (Coltheart, 1981). Coh-Metrix also assesses word properties that affect the accessibility of a word from memory, such as abstractness and ambiguity. Coh-Metrix computes abstractness and ambiguity scores by incorporating a module based upon WordNet (Miller, 1995). WordNet is an online lexicon tool that groups words into sets of synonyms that are connected by semantic relations. One such relationship, the *hypernym value*, refers to the number of levels a word has above it in a conceptual, taxonomic hierarchy. A low hypernym value is a proxy for word abstractness because the word has few distinctive features. Ambiguity, on the other hand, is inferred by the number of senses, or *polysemy value*, of a word. A polysemy value is simply a function of the number of synonym sets a word is assigned to. Coh-Metrix translates the hypernym value, as well as the polysemy value, into a mean composite score for any text.

In this study, we used a combination of 23 MRC database and WordNet measures (as calculated by Coh-Metrix) to represent the various aspects of vocabulary accessibility.

Method

The primary goal of this study was to provide a theoretically grounded and automated technique that extends traditional metrics of text difficulty. In doing so, we also wanted to demonstrate that groups of more or less difficult texts could be identified without manipulating or revising texts. To this end, it was necessary to establish groups of naturalistic text that were distinguishable only in terms of referential overlap and vocabulary accessibility. Using the Coh-Metrix indices that measure our two factors of text difficulty, a corpus of science texts were categorized into groups considered to be "difficult to understand" or "easy to understand". We hypothesized that the difficult texts would have lower scores in both referential overlap and vocabulary accessibility than that of easy texts. To ensure that our groups of difficulty were truly different, we evaluated comprehension of the texts by using sentence reading times and content recall. Based on the goals of our study, the method section that follows is divided into two parts: (a) Creating Groups of Text Difficulty and (b) Validating Groups of Text Difficulty.

Creating Groups of Text Difficulty

Corpus Selection In order to provide a diverse source of expository science texts, we collected an initial corpus of 161 candidate texts, compiled from 23 different science textbooks. The textbooks were from three different levels - junior high school (6-8th grade), high school (9-12th grade), and college (introductory undergraduate courses).

We initially examined two science domains: *physical science* and *life science*. Each domain consisted of 10 subtopics that were specifically chosen to align with national science education standards (National Research Council [NRC], 1996). For the physical science domain, there were 9 textbooks from 9 different publishers (2 junior high school textbooks, 2 high school textbooks, and 3 undergraduate level textbooks). For the life science domain, there were 14 textbooks from 11 different publishers (2 junior high school textbooks, 7 high school textbooks – of which four were from different publishers, and 5 undergraduate level textbooks).

From this initial corpus of 161 candidate texts, a subset of 60 texts was chosen. This subset consisted of 10 physical science subtopics and 10 life science subtopics selected from all three grade levels. This selection process was an iterative process of seeking to satisfy multiple constraint criteria. The first two criteria were concept-oriented while the last two were text-oriented. The *concept-oriented* criteria were focused on higher level factors of the text selection: (a) Maintaining topic alignment with national science education standards, and (b) Ensuring that the subtopics were taught at three different education levels – junior high school, high school, and college. The *text-oriented criteria* were focused on lower level, pragmatic constraints in text selection: (a) Excluding subtopics that were reliant on images or complex formula, and (b) Obtaining text passages that were of approximately the same length (400 – 500 words) that still formed complete conceptual units.

For instance, "the biological cell" is one of the main life science content standards. Within this content standard, "photosynthesis" is an important subtopic; thus, it meets the first criteria for inclusion in the corpus. The subtopic of photosynthesis is also covered in junior high, high school, and college classes, satisfying the second criteria of being taught at three different educational levels. The subtopic of photosynthesis is not dependent on images or complex formula, thus satisfying criterion three. Lastly, meeting the fourth criterion required obtaining three specific textbook passages (one from junior high school, high school, and college textbooks) that were of the correct length (i.e., 400-500 words) while accurately presenting the complete set of concepts and principles contained in this topic.

Data Reduction The large numbers of Coh-Metrix indices that measure referential overlap and vocabulary accessibility were reduced to six indices, three for each group. Typically, one would reduce a set of independent variables based on how well each independent variable differentiates the levels of a dependent variable (e.g., text difficulty). Because we do not have an *a priori* dependent variable, we used a Principle Components Analysis (PCA), with varimax rotation. A PCA is appropriate for our purposes because it is a mathematical technique that reduces a large number of observations (or indices) to N components. Each component is composed of observations that capture as much of the information from the original set of observations as possible. The final N components are rank-ordered according to the total variance explained. In turn, the observations in each component are rank-ordered according to how well they load onto their respective component.

The PCA reduction for the 30 Coh-Metrix referential overlap indices and 26 Coh-Metrix vocabulary accessibility indices were conducted within the sample space of the 60-text corpus. We maintained a 2:1 ratio of data points (i.e., science texts) to observations (i.e., Coh-Metrix index scores) in order to avoid spurious variance or "over-fitting" of the data (Witten & Frank, 2005). Because the indices would eventually be used as classification variables in distinguishing difficult and easy texts (see *clustering technique* section below), we selected three of the most representative indices from the entirety of the referential overlap indices, as well as three of the most representative indices from the entirety of the vocabulary accessibility indices. In the PCA, an index is considered most representative if it has the highest factor loading score in the principal component that accounts for the most overall unique variance.

For referential overlap, the PCA generated four significant principal components, with the first component explaining 68% of the overall variance. The three referential overlap indices selected were (a) unweighted proportional score of content words across adjacent sentences, (b) weighted proportional score of content words across two-sentence windows, and (c) weighted proportional score of content words across three-sentence windows. For vocabulary accessibility, the PCA generated six significant principal components, with the first component explaining 37% of the overall variance. The three vocabulary accessibility indices selected were (a) average of content word concreteness, (b) average of all words concreteness and (c) average of content word imagability.

For each respective group, we found the intercorrelations between the three indices to be statistically significant. The correlations between each group (taking the mean of each group) and the Flesh-Kincaid Reading Ease score were also significant. However, there was no significant correlation when groups were compared to each other (see Table 1).

Table 1. Pearson correlations between Flesch-Kincaid Readability Ease index, combined mean for referential overlap, and combined mean for vocabulary accessibility.

Indices of text difficulty	1	2	3
1. Reading ease	-	.54**	.32**
2. Referential overlap		-	.01
3. Vocabulary accessibility			-

**Correlation significant at p < .001.

Clustering Technique The PCA-selected Coh-Metrix indices of referential overlap and concept accessibility were used in an unsupervised cluster analysis to identify groups of text difficulty. A two-step clustering algorithm with the Akaike Information Criterion (AIC) computed Euclidean distances between data points in the 60-text corpus using the referential overlap or concept accessibility scores. Within each of these groups, the algorithm converged on two distinct text clusters by partitioning the variance so as to maximize the within-cluster variation and minimize the between-cluster variation (Kaufman & Rousseeuw, 1990). In order to classify the emergent text clusters as containing "difficult to understand" or "easy to understand" texts (per referential overlap and vocabulary accessibility scores), we took the mean differences of the combined indices in each group as a defining criterion. As such, a text was considered difficult if it had been assigned to clusters with the lowest mean scores for referential overlap and vocabulary accessibility. Conversely, a text was considered easy if it had been assigned to clusters with the highest mean scores for both referential overlap and vocabulary accessibility.

For the referential overlap clusters, the cluster with the highest Coh-Metrix mean score was 0.217, whereas the cluster with the lowest Coh-Metrix mean score was 0.122. For the vocabulary accessibility clusters, the cluster with the highest Coh-Metrix mean score was 0.416, whereas the cluster with the lowest Coh-Metrix mean score was 0.323. In the end, we selected 4 topics for which we could obtain 4 difficult and 4 easy texts.

Validating Groups of Text Difficulty

Participants Twenty-four undergraduates enrolled in an introductory psychology course participated for course credit.

Materials The materials consisted of eight texts that were classified as either difficult or easy in terms of the selected Coh-Metrix indices of referential overlap and vocabulary accessibility. Topic was also held constant between levels of difficulty to ensure that comprehension differences were not confounded with topic. Of the 20 topics that were involved in the original 60-text corpus, four emerged that contained a difficult and easy text version. These topics fall under the classification of *Life Sciences*, and describe the function of (a) *The Mammalian Eye*, (b) *The Biological Cell*, (c) *Photosynthesis*, and (d) *Chemistry of Life* (e.g., proteins, carbohydrates, and lipids).

Experiment Procedure Participants were tested in small groups of 2 to 4 participants. Prior to the experiment, the participants were informed that the goal of the study was to assess reading comprehension. As such, participants were expected to read a short passage from a science textbook and recall everything they could after reading each passage. The texts were presented on a computer monitor, with only a single sentence displayed on the monitor at any time. Participants advanced at their own pace by pressing the spacebar on the keyboard, thereby removing the currently displayed sentence and replacing it with the subsequent sentence. When the last sentence of the text had been read, participants were automatically instructed via the computer to type their recall in a text box. The dependent variables of recall and reading time per sentence were recorded.

Participants read four texts, one in each topic, and two at each level of text difficulty. We combined two counterbalancing methods to control for topic at each level of text difficulty. First, the order of topic presentation was counterbalanced by a four-order Latin square. Next, the order of text difficulty was counterbalanced by blocked randomization, resulting in six possible orders. Finally, the six blocked orders were mapped onto each row of the Latin square, thus resulting in 24 unique orders of topic combined with difficulty.

Scoring Procedure To score the free recall protocols, each sentence was divided into idea units by the Conceptual Unit Tagger, a web-based software developed at the University of Memphis (for additional information, visit http://141.225.14.229/cut/webform1.aspx). This tool systematically isolates idea units by analyzing the structural representation of a sentence in a syntactic parse tree. The syntactic tree, composed of an underlying formal grammar, is generated using the Charniak (1997) parser. The root of the tree (i.e., the sentence under analysis) is separated into intermediate branches that specify nodes that include noun phrases (NP), verb phrases (VB), prepositional phrases (PP), and embedded sentence constituents. The tool selects a node as a single coherent concept if it adheres to simple guidelines, such as containing a finite verb with related arguments (e.g., dependent and independent clausal phrases) or a prepositional phrase that contains a gerund. For example, here are two sentences (1) *The phase of a substance can be changed | by adding or removing heat* and (2) *It is not affected | in reproducing for the rest of it's lifespan.* Based upon the preceding guidelines each sentence would each be identified as having two distinct idea units (delineated by the "|" symbol).

Results

Free Recall

The mean number of idea units recalled for the difficult texts was compared with the mean number of idea units recalled for the easy texts. We conducted a one-way within-subject ANOVA to evaluate the differences in idea units recalled. There was a significant effect for type of text (difficult vs. easy), $F(1,22) = 24.59$, $p < .001$, $\eta^2 = .528$. Participants recalled more from the easy texts than from the difficult texts (see Table 2).

In addition to number of idea units recalled, we also computed the number of words recalled (see Table 2). A one-way within-subject ANOVA demonstrated a significant effect for type of text (difficult vs. easy), $F(1,22) = 41.80$, $p < .001$, $\eta^2 = .655$. Again, participants recalled more from the easy texts than from the difficult texts.

The last analysis involved the qualitative differences of recall for difficult and easy texts. We used LSA to assess the contextual similarity between the free recall and the text from which the free recall was generated. We used the TASA (general college) semantic space and "document x document" comparison metric. The LSA cosine scores for each of the four texts (2 difficult and 2 easy) that the participants read and recalled were submitted to a one-way within-subjects ANOVA. There was a significant effect for type of text (difficult vs. easy), $F(1,22) = 13.19$, $p < .001$, $\eta^2 = .528$. Participant's recall was more contextually similar to the text for the easy texts than for the difficult texts (see Table 2).

Table 2: Recall based on number of words, number of idea units, and LSA scores between text and recall.

	Level of text difficulty	
Unit of analysis	Easy texts Mean(SD)	Difficult texts Mean(SD)
Number of idea units	12.09(4.55)	8.37(2.92)
Number of words	87.67(31.62)	57.00(20.59
LSA	0.78(0.07)	0.69(0.11)

Reading Times

The reading times for each sentence were recorded for the difficult and easy texts. Before analyzing the data, it was necessary to normalize each sentence for differences in length. Four techniques were used: (a) reading time divided by number of characters, (b) reading time divided by number of syllables, (c) reading time divided by number of words, and (c) reading time divided by number of idea units. After normalizing for length, we also removed reading times that were possible outliers for each participant. A reading time was excluded if the time was two standard deviations above or below the mean of reading time for all sentences. Across all participants, we removed 1.36% of the reading times per character, 2.00% per syllable, 1.80% per word, and 1.12% per idea unit. The remaining normalized reading times for the difficult texts were compared against the normalized reading time scores for the easy texts. We used a one-way within-subjects ANOVA to determine if differences between levels of text difficulty were significant.

Table 3. Reading times for difficult and easy texts normalized by character, syllable, word, and idea units.

	Level of text difficulty	
Reading time:	Easy texts Mean(SD)	Difficult texts Mean(SD)
by character	61.84 (3.41)	68.74 (5.40)
by syllable	230.93 (12.15)	247.48 (19.62)
by word	363.23 (19.74)	426.63 (34.25
by idea units	3113.67 (162.62)	4014.96 (320.40)

There was a statistically significant effect when reading times were normalized by number of characters, $F(1,23) = 4.25$, $p < .05$, $\eta^2 = .162$, number of words $F(1,23) = 8.09$, $p < .01$, $\eta^2 = .269$, and number of idea units $F(1,23) = 18.58$, $p < .01$, $\eta^2 = .458$. Overall, these results suggest that participants spend more time (per sentence) reading the difficult texts (see Table 3). It should also be noted that the normalization by syllables was not significant. However, there was a trend of slower reading time when processing the difficult texts.

Discussion

In this study, we addressed a challenge faced by many educators: Given a diverse set of instructional texts, how is text difficulty established? Using Coh-Metrix, a computational language processing tool, we demonstrated that two independent factors of cohesion and semantics could uncover divergent groups of text difficulty in a large corpus. Specifically, a subset of three indices for *referential overlap* (a factor of cohesion) and a subset of three indices for *vocabulary accessibility* (a factor of semantics) were used in identifying texts that were difficult or easy to understand. Texts that had high scores in referential overlap and vocabulary accessibility (i.e., easy texts) were read faster and recalled better than texts with low scores in referential overlap and vocabulary accessibility (i.e., difficult texts). Our results contribute to a large body of reading comprehension research that makes use of text-level features to vary text coherence. However, where previous research varied coherence by hand, we used an automated technique that allows natural differences between texts to emerge.

Our technique also has many of the advantages of traditional readability formulas. For example, the Flesch-Kincaid Reading Ease (FKRE) formula is widely used by educators because of its proven effectiveness in identifying text difficulty. As reported earlier, the correlations between Coh-Metrix indices and FKRE scores are statistically significant, thus suggesting the two techniques are on par with each other. In similar fashion, the Coh-Metrix indices and FKRE scores also provide text assessments that are reliable and automatic.

There are also notable discrepancies between the techniques that may favor one technique over the other. For example, educators and researchers can use Coh-Metrix to identify texts that vary along two independent dimensions of coherence (e.g., cohesion and semantic information). Moreover, future research will provide educators and researchers additional options by incorporating Coh-Metrix indices of temporal/causal, anaphor resolution, and syntactic complexity. The FKRE formula, in contrast, does not allow such an in depth analysis because the scores are based on shallow linguistic features that converge on a generalized index of difficulty.

A possible advantage for the FKRE formula, however, is the ability to identify texts on an absolute scale. At this point, the technique used in this study is based on relative text difficulty. Further analyses are required to determine if the difficulty thresholds reported here are reflections of the true population (i.e., junior high, high school, and college level instructional texts). If so, identifying text difficulty will not necessitate a cluster-like analysis for each evaluation.

While much work remains to be done, this initial investigation contributes to the field by demonstrating that Coh-Metrix derived indices accurately identify texts that have unique influences on human comprehension. In doing so, we hope to provide educators a simple and theoretically grounded technique to select appropriate texts that match their students' individual reading abilities.

Acknowledgements

The research was supported by the Institute for Educational Sciences (IES R305G020018-02, IES R3056040046). 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.

References

Best, R. M., Rowe, M., Ozuru, Y., & McNamara, D. S. (2005). Deep-level comprehension of science texts: The role of the reader and the text. *Topics in Language Disorders, 25*, 65-83.

Charniak, E. (1997). Statistical techniques for natural language processing. *AI Magazine, 18*, 33-44.

Coltheart, M. (1981). The MRC psycholinguistics database. *Quarterly Journal of Experimental Psychology, 33A*, 497-505.

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

Freebody, P., & Anderson, R. C. (1983). Effects of vocabulary difficulty, text cohesion, and schema activation on reading comprehension. *Reading Research Quarterly, 18*, 277-294.

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.

Graves, M. F., Prenn, M. C., Earle, J., Thompson, M., Johnson, V., & Slater, W. H. (1991). Improving instructional texts: Some lessons learned. *Reading Research Quarterly, 26*, 110-122.

Hiebert, E. H. (2002). Standards, assessment, and text difficulty. In A. E. Farstrup & S. J. Samuels (Eds.). *What research has to say about reading comprehension (3rd Ed.).* Newark, DE: International Reading Association.

Kaufman, L., & Rousseeuw, P. J. (1990). Finding groups in data: An introduction to cluster analysis NY: John Wiley & Sons.

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

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.

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 26th Annual Meeting of the Cognitive Science Society* (pp. 843-848). Mahwah, NJ: Erlbaum.

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

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

National Research Council (NRC). (1996). *National science education standards*. Washington, DC: National Academy Press.

Paivio, A (1969). Mental Imagery in associative learning and memory. *Psychological Review, 76*, 241-263.

Stahl, S. A., Jacobson, M. G., Davis, C. E., & Davis, R. L. (1989). Prior knowledge and difficult vocabulary in the comprehension of unfamiliar texts. *Reading Research Quarterly, 24*, 27-43.

Schnick, T. (2000). *The Lexile framework: An introduction for educators.* New York, NY: MetaMetrics.

van Oostendorp, H., & Goldman, S. R. (Eds.). (1999). The construction of mental representations during reading. Mahwah, NJ: Lawrence Erlbaum Associates, Inc.

Witten, I. H., & Frank, E. (2005). *Data mining: Practical machine learning tools and techniques.* San Francisco, CA: Morgan Kaufmann Publishers.