When Do Standard Approaches for Measuring Vocabulary Difficulty, Syntactic Complexity and Referential Cohesion Yield Biased Estimates of Text Difficulty?

Kathleen M. Sheehan (ksheehan@ets.org)
Irene Kostin (ikostin@ets.org)
Yoko Futagi (yfutagi@ets.org)
Educational Testing Service, MS 12-R, Rosedale Road
Princeton, NJ 08541 USA

Abstract

Many widely-used approaches for assessing text difficulty provide a single prediction equation that is assumed to hold for texts belonging to a variety of different genres. This paper demonstrates that such models tend to overpredict the difficulty of informational texts while simultaneously underpredicting the difficulty of literary texts. Mechanisms that may account for these effects are examined. Results suggest that the estimated effects can be traced to certain frequently used measures of vocabulary difficulty, syntactic complexity and referential cohesion.

Keywords: genre; readability; referential cohesion; syntactic complexity; vocabulary difficulty.

Introduction

Several recent studies have examined differences in the processing requirements of informational and literary texts. Differences have been reported in word reading rates (Zabrucky & Moore, 1999); in vocabulary usage (Lee, 2001), in the way that cohesion relations are expressed (McCarthy, Graesser & McNamara, 2006); in the types of inferences generated during reading (van den Broek et al., 2002); in readers' use of comprehension strategies (Kukan & Beck, 1997) and in the type of prior knowledge accessed during inference generation (Best, Floyd & McNamara, 2004).

Several explanations for these findings have been proposed. In one view, literary texts (e.g., narratives and memoirs) are said to require different processing strategies because they deal with more familiar concepts and ideas (Graesser, McNamara & Louwerse, 2003). For example, while many literary texts employ familiar story grammars that are known to even extremely young children, informational texts tend to employ less well known structures such as cause-effect, comparison-contrast, and problem-resolution.

Genre-specific processing differences have also been attributed to differences in the types of vocabularies employed. To document this, Lee (2001) examined differences in the use of "core" vocabulary within a corpus of informational and literary texts that included over one million words downloaded from the British National Corpus. Core vocabulary was defined in terms of a list of 2000 common words classified as appropriate for use in the dictionary definitions presented in the Longman Dictionary of Contemporary English. The analyses demonstrated that core vocabulary usage was higher in literary texts than in informational texts. For example, when literary texts such as fiction, poetry and drama were considered, the percent of total words classified as "core" vocabulary ranged from 81% to 84%. By contrast, when informational texts such as science and social studies texts were considered, the percent of total words classified as "core" vocabulary ranged from 66% to 71%. In interpreting these results Lee suggested that the creativity and imaginativeness typically associated with literary writing may be less closely tied to the type or level of vocabulary employed and more closely tied to the way that core words are used and combined. Note that this implies that an individual word detected in a literary text may not be indicative of the same level of processing challenge as that same word detected in an informational text.

Significant genre-related differences were also detected in the corpus-based analysis reported in McCarthy et al. (2006). That analysis detected higher levels of referential cohesion in expository texts as compared to narratives even though the two corpora studied were targeted at equivalent populations of readers, i.e., students in grades kindergarten through college. These results suggest that it may also be the case that a particular level of referential cohesion detected in an expository text may not necessarily be indicative of the same type of processing challenge as that same level detected in a narrative text.

While few would dispute the informational/literary distinctions noted above, text difficulty models that account for such differences are rare. In particular, widely-used approaches such as the Flesch-Kincaid Grade Level Score (Flesch, 1948); the Dale-Chall Readability formula (Chall & Dale, 1995); and the Lexile Framework (Stenner, et al., 1988) each provide a single prediction equation that is assumed to hold for both informational and literary texts. The tendency to ignore genre effects is also present in more recent work. For example, Crossley, Dufty, McCarthy & McNamara (2007) present a text difficulty model composed of three features: average sentence length, average word frequency and referential cohesion. Even though the model was estimated from a corpus comprised of both informational and literary passages, genre effects were not examined.

Investigations focused on the variation captured by particular proposed features have also tended to ignore genre effects. For example, McNamara et al. (2006)

evaluated alternative approaches for assessing variation due to differences in text cohesion. The approaches differed in terms of (a) whether lexical overlap was calculated using all content words or just nouns; (b) whether overlap was calculated with or without stemming (i.e., with inflected forms of a word, such as *read*, *reads* and *reading,* treated as equivalent); (c) whether the calculation considered one previous sentence, two previous sentences, three previous sentences or all previous sentences; and (d) whether or not indices included distance-based weighting, i.e. weights that favored referents detected in nearby sentences over those detected in more distant sentences. The evaluation suggested that the most useful indices were those that focused on noun overlap only, and that considered either two or three previous sentences. Note however, even though the corpus considered in the analyses included both informational and literary texts, genre effects were not examined.

In a rare exception to this trend, Sheehan, Kostin & Futagi (2008) provide two independent text difficulty models: one optimized for application to informational texts, and one optimized for application to literary texts. The focus of this paper, however, is not on the details of any particular prediction model. Rather, our goal is to document the inferential errors that are likely to arise when genre effects are ignored. By documenting these errors, we hope to underscore the importance of accounting for genre effects in future investigation of text difficulty, including both feature evaluation studies, and model development studies.

Our paper is structured as follows. First, the cognitive model adopted as a basis for the investigation is summarized. Next, we present our methodology, results and conclusions.

Text Comprehension

Reading is a complex activity that requires the coordinated use of a large number of component processes. Kintsch (1998) characterizes these processes as operating at three separable, yet interacting levels. First, word recognition and decoding processes are used to translate the written code into meaningful language units called propositions. Next, interrelationships among the propositions are clarified. Depending on the characteristics of the text and the reader's goals, this processing could involve reader-generated bridging inferences designed to fill in gaps and establish coherence. Kintsch argues that this process culminates in the development of a network representation of the text called the *textbase*. While only text-based inferences are generated during the construction of the textbase, knowledge-based inferences may also be needed to completely satisfy a reader's goals. Consequently, a third level of processing is also frequently implemented. This third level involves reconciling the current text with relevant prior knowledge and experience to provide a more complete, more integrated model of the situation presented in the text, i.e., what Kintsch calls *the situation model.*

Best, Floyd & McNamara (2004) discuss differences in the processes engaged when developing situation models for expository vs. narrative texts. For expository texts, situation model processing involves integrating the textbase with readers' prior knowledge of the subject matter. Since a given reader's prior knowledge may not always be sufficient, resulting situation models may fail to maintain the author's intended meaning. For narrative texts, by contrast, situation model processing involves generating inferences about the characters, settings, actions and events in the reader's mental representation of the story, an activity that is much less likely to be affected by deficiencies in required prior knowledge. These findings further underscore the importance of attending to genre effects when developing models of text difficulty and when evaluating alternative approaches for measuring specific components of text difficulty.

Method

As is indicated in the above review, many existing approaches for assessing text difficulty provide a single prediction equation that is assumed to hold for texts belonging to a variety of different genres. Because our literature review suggested that the processes engaged when reading informational texts differ substantially from those engaged when reading literary texts, this study presents a detailed analysis of the prediction errors obtained when difficulty estimates are generated *without* considering variation due to differences in text genre. Key aspects of the methodology are described below.

Defining "True" Text difficulty

Every modeling application requires a start up set of observations for which the "true" values of the dependent variables are known. Previous text modeling applications have considered a variety of different approaches for specifying the "true" difficulty level of a text. In one approach, "true" text difficulty estimates are developed from cloze fill-in rates (e,g., Chall and Dale, 1995; Crossley et al. 2007, Stenner, et al., 1988). Shanahan, Kamil, and Tobin (1983) evaluated this approach by comparing students' performances on cloze items administered under four different passage conditions: (a) intact passages, (b) scrambled passages (with sentences randomly reordered); (c) intermingled passages (with sentences from different passages interspersed); and (d) eclectic passages (collections of unrelated sentences). After observing similar cloze fill-in rates under all four conditions, Shanahan et al. (1983) concluded that cloze fill-in rates do not provide useful information about "intersentential" comprehension, that is, comprehension that requires integrating information across sentence boundaries.

Responses to multiple-choice reading comprehension items have also been used to measure text difficulty. In this approach, the "true" difficulty level of a text is estimated from the average difficulty of its associated items. While item difficulty is surely related to passage difficulty, several previous studies have suggested that item difficulty estimates also incorporate variation due to non-passage factors such as distractor plausibility (Embretson & Wetzel, 1987; Freedle & Kostin, 1991; Gorin & Embretson, 2006.)

Researchers have also generated "true" text difficulty estimates from grade level (GL) classifications provided by textbook publishers or Web content providers (e.g., Heilman, et al. 2007). While this approach may occasionally provide useful information about gradations of text difficulty, in many cases, classification guidelines are not published, so the specific factors considered during text classification may be difficult to determine. An additional problem is that procedures for detecting and correcting misclassifications are not typically provided.

The "true" text difficulty estimates considered in this study are designed to reflect the aspects of text variation deemed appropriate for readers at successive grade-levels (GLs) in the range from 3rd to 12th grade. Estimates were developed as follows. First, we assembled a balanced corpus of informational and literary passages that had each been previously administered on a high-stakes accountability assessment such as those mandated by the federal No Child Left Behind legislation. Next, we employed an "inheritance principle" to assign a GL classification to each passage. That is, we assumed that the GL classification of each passage could be inherited from the GL classification of its parent test form. Note that the resulting classifications offer a number of advantages: (1) classifications are developed and reviewed in a high-stakes environment; (2) classifications are based on published guidelines that have also been reviewed by large numbers of concerned stake-holders; (3) since classifications are specified in advance of item administration, they are not overly influenced by non-passage factors such as option plausibility; (4) classification errors (i.e., passages that are either too easy or too hard for the targeted population of readers) are typically detected and corrected during the pretesting phase of test development; and (5) classifications are designed to capture both intersentential and intrasentential effects.

Corpus

The corpus consisted of 374 informational and literary texts selected to represent the range of linguistic variation typically observed in reading passages developed for reading assessments targeted at students in grades 3 through 12. Individual passages were selected from three different state assessments: California, Michigan and Maine. GL and genre classifications (informational vs. literary) were obtained for all passages. The resulting corpus included a total of 171 informational passages drawn from the content areas of science, social studies and humanities, and 203 literary passages that were developed from either fiction or literary memoirs.

Features

Eight widely-used measures of text difficulty were evaluated: average word length measured in log syllables; average word frequency determined from the TASA Index (Touchstone Applied Science Associates, Zeno, et al. 1995); average word frequency determined from a second word frequency index developed from a version of the Lexile Corpus provided by the Metametrics Corporation; average sentence length measured in log words, and four unweighted measures of referential cohesion. Like the referential cohesion measures described in McNamara et al. (2006), each cohesion measure provides an estimate of the average number of sentences classified as exhibiting noun overlap with preceding sentences. The measures differ with respect to whether stemming is included (Yes or No) and whether overlap detection considers 2 or 3 preceding sentences.

Genre Effects

Genre effects are evaluated by fitting the following linear model to the full corpus of 374 passages:

$$y_i = \beta_0 + \beta_{0,\inf} x_{0i} + \beta_1 x_{1i} + \beta_{1,\inf}(x_{0i} * x_{1i}) + \varepsilon_i$$

(1)

where y_i is the observed GL of the i^{th} text, x_{0i} is an indicator variable coded as 1 for informational texts and 0 for literary texts, and x_{1i} represents one or another of the eight candidate features discussed above (i.e., a measure of vocabulary difficulty, syntactic complexity or referential cohesion.) Note that this model permits evaluation of two different types of genre effects: effects that function independently of x_1 and effects that vary linearly with x_1.

The practical significance of failing to account for variation due to differences in text genre is also evaluated. This is accomplished by first estimating the above model with $\beta_{0,\inf}$ and $\beta_{1,\inf}$, *excluded* to obtain a non-genre-specific $\hat{y}_i$ for each text, and then calculating mean differences separately for informational and literary texts as follows:

$$Bias(Inf) = 1/171 \sum (\hat{y}_i - y_i), \text{ for } i = 1, \ldots, 171$$

$$Bias(Lit) = 1/203 \sum (\hat{y}_i - y_i), \text{ for } i = 172, \ldots, 374.$$

Note that a positive bias value is indicative of a prediction equation that systematically *overestimates* text difficulty, and a negative bias value is indicative of a prediction equation that systematically *underestimates* text difficulty.

each of the four measures of referential cohesion. Note that these results are consistent with the theoretical model of text comprehension summarized above.

Table 1. Model Coefficients, Significance Probabilities and Expected Genre Biases
For Selected Measures of Vocabulary Difficulty, Syntactic Complexity and Referential Cohesion

Model	β_1	$\beta_{0,\text{inf}}$	$\beta_{1,\text{inf}}$	Bias(Inf)	Bias(Lit)
Vocabulary Difficulty					
Avg. Word Length (log syllables)	61.58 ***	-10.24 **	21.93 *	+ 0.72	- 0.60
Avg. TASA Word Frequency	-0.57 ***	13.48 *	-0.24 *	+ 0.54	- 0.45
Avg. Lexile Word Frequency	-0.55 ***	1.42 *	-0.05	+ 0.73	- 0.61
Syntactic Complexity					
Avg. Sentence Length (log words)	12.07 ***	-6.21 ***	4.68 **	+ 0.30	- 0.26
Referential Cohesion					
V1: Stemming=No, Span=2	-3.56 ***	-1.12 *	3.52 ***	- 0.14	+ 0.12
V2: Stemming=No, Span=3	-2.77 ***	-1.07 *	2.70 ***	- 0.13	+ 0.11
V3: Stemming=Yes, Span=2	-3.06 ***	-1.04 *	3.04 **	- 0.11	+ 0.10
V4: Stemming=Yes, Span=3	-2.43 ***	-1.03 *	2.41 **	- 0.11	+ 0.10

Notes: * $p < .10$, ** $p < .05$, *** $p < .01$. Biases are expressed on a GL scale.

Results and Discussion

Model coefficients estimated via equation (1) are summarized in Table 1. The column labeled β_1 confirms that, as expected, each of the selected features varies linearly with GL in the expected direction. That is, on average, GL *increases* with average word length and average sentence length, and *decreases* with average word frequency as determined from either the TASA Word Frequency (WF) Index or the Lexile WF Index, and with each of the four measures of referential cohesion. Note that these results are consistent with the theoretical model of text comprehension summarized above.

Table 1 also shows that significant interactions with text genre were detected for each of the eight features. In particular, $\beta_{0,\text{inf}}$ is significant for all eight features, and $\beta_{1,\text{inf}}$ is significant for each of the features except the Lexile WF feature. (The fact that $\beta_{1,\text{inf}}$ is not significant for the Lexile WF feature merely indicates that, for this feature, the magnitude of the bias is invariant across the entire observed range of feature values.) These results confirm that models of text difficulty that include any of these eight features without also accounting for variation due to differences in text genre run the risk of yielding predictions of text GL that incorporate significant genre biases.

The direction and magnitude of the resulting biases are also shown in Table 1. Note that all three of the vocabulary features and the average sentence length feature yielded *positive* bias for informational texts and *negative* bias for literary texts. Figure 1 provides a graphical display designed to illuminate these results. The plot shows changes in text GL conditional on the Average Lexile Word WF feature. In order to highlight differences in the results obtained for informational and literary texts, a LOESS scatter plot smoother has been applied to the data (Cleveland, 1979). Results for informational texts are plotted with a solid line; results for literary texts are plotted with a dashed line. Note that the literary curve appears above the informational curve throughout the entire observed range of the data. This confirms that a given value of the Average Lexile WF feature is indicative of a *higher* average GL score if the text in question is a literary text and a *lower* average GL score if the text in question is an informational text. Since a model that includes this feature without also accounting for genre differences will tend to yield predictions that fall *between* the two curves, resulting GL predictions will tend to be too high for informational texts (positive bias) and too low for literary texts (negative bias). Although not shown here, similar plots prepared for each of the other two vocabulary features and the average sentence length feature yielded similar trends.

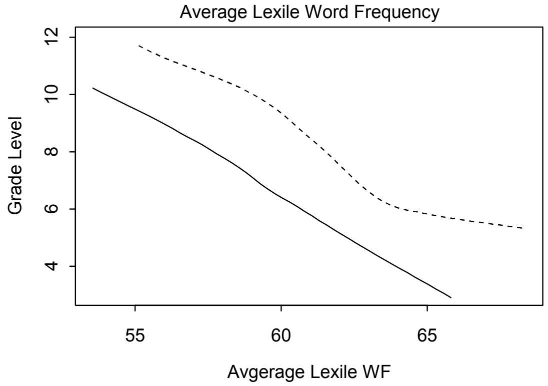

Figure 1. Trends in GL variation plotted conditional on the Average Lexile WF feature for informational texts (solid line) and literary texts (dashed line).

Table 1 also shows the bias expected for four common approaches for measuring referential cohesion. Note that the direction of the bias is *opposite* to that shown for the vocabulary and syntactic complexity measures.

Table 2. Model Coefficients, Significance Probabilities and Genre Biases for Models Involving Multiple Features

	Type of		Average			
	Vocabulary	Vocabulary	Sentence	Referential	GL Bias	GL Bias
Model	Measure	Difficulty	Length	Cohesion[a]	(Inf)	(Lit)
FK (Original)	AWL	11.80 ***	0.39 ***		+1.40	- 0.65
FK (Reestimated)	Log (AWL)	28.13 ***	11.31 **		+ 0.64	- 0.54
Lexile (Reestimated)	Avg. Lexile WF	- 0.09 **	12.09 ***		+ 0.45	- 0.38
Mixed, Version 1	Avg. Lexile WF	- 0.25 ***	11.51 ***	-3.38 ***	+ 0.19	- 0.16
Mixed, Version 1, Inf	Avg. Lexile WF	- 0.37 ***	12.06 ***	-1.71 ***	0.00	0.00
Mixed, Version 1, Lit	Avg. Lexile WF	- 0.22 ***	11.03 ***	-4.77 ***	0.00	0.00
Mixed, Version 2	Avg. TASA WF	- 0.37 ***	11.50 ***	-2.83 ***	+ 0.10	-0.10
Mixed, Version 2, Inf	Avg. TASA WF	- 0.50 ***	11.36 ***	-1.08 **	0.00	0.00
Mixed, Version 2, Lit	Avg. TASA WF	- 0.26 ***	11.56 ***	-4.70 ***	0.00	0.00

Notes: ** $p < .05$, *** $p < .01$. AWL = Average Word Length, WF = Word Frequency, Inf = Informational, Lit = Literary
a. Referential Cohesion is assessed via feature V1 in Table 1, i.e., Stemming = No, Span = 2.

This suggests that a model that includes all three types of measures (i.e., a measure of vocabulary difficulty, a measure of syntactic complexity and a measure of referential cohesion) may benefit from mutually compensating biases.

Table 2 presents results for a series of nine multiple-feature models. Both the original Flesch-Kincaid (FK) model and an updated FK model are included. The updated FK model differs from the original in that both feature values (i.e., average word length and average sentence length) are logged. This change was implemented because the non-logged values exhibited significant departures from linearity. The table also provides a reestimated version of the Lexile model. This model differs from the model described in Stenner et al. (1988) in that a slightly different version of the Lexile WF Index is employed. Results confirm that, as expected, all three models yield positive biases for informational texts and negative biases for literary texts. The table also shows that the biases are most extreme for the original FK model. This finding is most likely due to the extreme departures from linearity exhibited by that model.

Several Mixed Models are also shown. These employ the same three features as those reported in Crossley et al. (2007) with some slight modifications. For example, although Crossley et al. employed the CELEX WF Index, Version 1 of the Mixed Model employs the Lexile WF Index and Version 2 employs the TASA WF Index. Results confirm that the TASA index yields slightly less bias, both for informational texts and for literary texts. (To see this, note that GL Bias(Inf) = +0.10 when the TASA WF Index is used, and GL Bias(Inf).= +0.19 when the Lexile WF Index is used. Similarly, GL Bias(Lit) = -0.10 when the TASA WF Index is used and GL Bias(Lit) = -0.16 when the Lexile WF Index is used.) The results also show that, although inclusion of the referential cohesion measure has succeeded in reducing the overall bias it has also introduced an internal bias. That is, the effects due to differences in referential cohesion are either understated or overstated depending on whether the text in question is an informational text or a literary text. This can be seen by comparing the cohesion coefficients listed for the non-genre-specific Mixed Models to those listed for the corresponding genre-specific models. The comparison shows that when genre effects are properly accounted for, cohesion effects are much smaller among informational texts (i.e., -1.71 or -1.08 depending on the particular WF Index used) and much larger among literary texts (i.e., -4.77 or -4.70 depending on the particular WF Index used.) When genre effects are not accounted for, however, an average effect is estimated (i.e., -3.38 or -2.83 depending on the particular WF Index used.) This suggests that, no matter which WF Index is used, a model that does not account for variation due to differences in text genre will tend to the overstate the impact of referential cohesion among informational texts and understate the impact of referential cohesion among literary texts.

Conclusions

Ever since the first readability formulas were published in the late 1940s researchers have been investigating alternative approaches for modeling text difficulty. In almost every case, these investigations have culminated in the development of a single prediction equation that is assumed to hold for texts belonging to a variety of different genres. This paper has demonstrated that such models tend to generate predicted values of text difficulty that are consistently too low for some genres and consistently too high for other genres.

The findings reported above allow for strong conclusions regarding the direction and magnitude of the genre biases expected for models containing certain frequently used measures of text variation. For example, the analyses confirmed that several frequently used approaches for measuring vocabulary difficulty tend to be structured such that resulting text difficulty estimates overstate the difficulty of informational texts while simultaneously understating the difficulty of literary texts. These results can be explained in terms of the higher proportion of "core" vocabulary words typically found in literary texts as opposed to informational texts.

The analyses also demonstrated that it is possible to define multivariate prediction models that benefit from mutually compensating biases. Models that include both vocabulary effects and referential cohesion effects fall into this category. That is, when only informational texts are considered, the positive bias in the vocabulary effect is offset by the negative bias in the referential cohesion effect. Similarly, when only literary texts are considered, the negative bias in the vocabulary effect is offset by the positive bias in the referential cohesion effect. While subsequent estimates of text difficulty will tend to exhibit relatively small genre biases, conclusions regarding the contributions attributable to specific text features may be substantially less accurate. Understanding these biases is especially important when the goal of the analysis is not only to assess difficulty, but to also control it.

We are continuing to evaluate the biases associated with additional text features. Such research can facilitate the goal of providing models of text difficulty that provide more precise information about *why* individual texts are likely to be more or less challenging for particular populations of readers.

References

Best, R., Floyd, R.G, & McNamara, D. S. (2004, April). *Understanding the fourth-grade slump: Comprehension difficulties as a function of reader aptitudes and text genre.* Paper presented at the Annual Conference of the American Educational Research Association: SanDiego, CA.

Chall, J.S., & Dale, E. (1995). *Readability Revisited: The new Dale-Chall readability formula.* Cambridge: Brookline Books.

Cleveland, W.S. (1979). Robust locally weighted regression and smoothing scatterplots. *Journal of the American Statistical Association*, *74*, 829-836.

Crossley, S.A., Dufty, D.F., McCarthy, P.M., & McNamara, D.S. (2007). Toward a new readability: A mixed model approach. In D. S. McNamara and G. Trafton (Eds.), *Proceedings of the 29th Annual Conference of the Cognitive Science Society*. Nashville, TN:Cognitive Science Society.

Embretson, S.E. & Wetzel, C.D. (1987). Component latent trait models for paragraph comprehension tests. *Applied Psychological Measurement*, 11, 175-193.

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

Freedle, R. & Kostin, I. (1991). *The prediction of SAT reading comprehension item difficulty for expository prose passages.* ETS Research Report # RR-91-29. Princeton, NJ: Educational Testing Service.

Gorin, J. S., & Embretson, S. E. (2006). Item difficulty modeling of paragraph comprehension items. *Applied Psychological Measurement, 30*(5), 394-411.

Graesser, A.C., McNamara, D., & Louwerse, M. (2003). What readers need to learn in order to process coherence relations in narrative and expository text. In A.P. Sweet & VC. E. Snow (Eds.), *Rethinking reading comprehension* (pp. 82-98). New York: Guilford Press.

Heilman, M., Collins-Thompson, Callan, J. & Eskenazi, M. (2007) Combining lexical and grammatical features to improve readability measures for first and second language texts. *Proceedings of NAACL HLT 2007*, pages 460-467.

Kintsch, W. (1998). *Comprehension: A paradigm for cognition.* Oxford: Cambridge.

Kukan, L., & Beck, I.L., (1997) Thinking aloud and reading comprehension research: Inquiry, instruction, and social interaction. *Review of Educational Research*. 67, 271-299.

Lee, D. Y. W. (2001) Defining core vocabulary and tracking its distribution across spoken and written genres. *Journal of English Linguistics*. 29, 250-278.

McCarthy, P., Graesser, A.C., & McNamara, D.S. (2006, July). *Distinguishing genre using Coh-Metrix indices of cohesion*. Poster presented at the annual meetings of the Society for Text and Discourse, Minneapolis, MN.

McNamara, D., Ozuru, Y., Greasser, A., & Louwerse, M. (2006). Validating Coh-Metrix. In R. Sun & N. Miyake (Eds.), *Proceedings of the 28th Annual Conference of the Cognitive Science Society*, Mahwah, NJ:Erlbaum.

Shanahan, T., Kamil, M.L., & Tobin, A.W. (1983). Cloze as a measure of intersentential comprehension. *Reading Research Quarterly, 17*, 229-255.

Sheehan, K. M., Kostin, I., & Futagi, Y (2008). *Reading level assessment for high-stakes testing applications: A second look at variation due to differences in text genre.* ETS Research Report Princeton, NJ: Educational Testing Service.

Stenner, A.J. Horabin, I., Smith, D.R & Smith, M. (1988). Most comprehension test do measure reading comprehension: A response to McLean and Goldstein. *Phi Delta Kappan*, June, 1988, 765-767.

van den Broek, P., Everson, M., Virtue, S., Sung, Y., & Tzeng, Y. (2002). Comprehension and memory of science texts: Inferential processes and the construction of a mental representation. In J. Otero, J. Leon, & A. C. Graesser (Eds.), *The psychology of science text comprehension* (pp 131-154). Mahwah, NJ: Lawrence ErlbaumAssociates.

Zabrucky, K.M. & Moore, D. (1999). Influence of text genre on adults' monitoring of understanding and recall. *Educational Gerontology, 25*, 691-710.

Zeno, S.M., Ivens, S.H., Millard, R.T., Duvvuri, R. (1995). *The educator's word frequency guide.* Brewster, NY: Touchstone Applied Science Associates.