About the *TextEvaluator*® Technology

The *TextEvaluator*® technology provides a fully-automated approach for obtaining valid and reliable feedback about the complexity characteristics of reading passages selected for use in instruction and assessment. Teachers and other educators can use this technology to:

- determine an appropriate grade-level placement for a text, and
- determine which of eight possible sources of comprehension difficulty are likely to be most challenging within any specified text.

History

The *TextEvaluator* scoring engine was originally designed to help ETS test developers work more efficiently when searching for passages for use on assessments targeted at students with varying reading proficiency profiles. An early version of the engine was released in 2010 under the name *SourceRater*. In 2013, the name was changed from *SourceRater* to *TextEvaluator*®.

Overview

The *TextEvaluator* engine is designed to score any professionally edited text that is formatted as continuous prose, including informational text, literary text, and text that incorporates a mixture of informational and literary elements. The engine is not designed to score discontinuous or oddly-formatted texts such as menus, tables, recipes, lists and poetry. The engine is also not designed to score texts that include syntax, spelling or word choice errors (e.g., student essays).

This document provides additional information about the *TextEvaluator*® tool including:

- guidelines for ensuring that texts are properly formatted (p. 2);
- documentation of the *TextEvaluator* measurement approach (p. 3), including detailed descriptions of the *TextEvaluator* component scores (p. 6), types of feedback (p. 8), and genre categories (p. 10);
- a concordance table for use when assessing alignment with the Common Core text complexity guidelines (p. 12); and
- a summary of key validity evidence (p. 13).

TextEvaluator® Formatting Guidelines

TextEvaluator analyses will be most accurate when texts are formatted according to the following guidelines:

- Manually remove non-text elements such as figures, tables, equations, footnotes/endnotes, diagrams, pronunciation guides, author by-lines and non-standard characters.
- Since the *TextEvaluator* scoring engine considers paragraph structure as part of its feature extraction process, it is recommended that at least one hard return be inserted between each paragraph. Also make sure that Hard Returns are not included in the middle of paragraphs. (Note: some documents are formatted such that a Hard Return is placed at the end of each line. Texts with this type of formatting may not be accurately analyzed by the *TextEvaluator* tool since each line will then be interpreted as a separate paragraph.)
- If the text has headers, they should be formatted as the first sentence of the paragraph. That is, add a period, question mark or other end-of-sentence marker, and remove the closing Hard Return. This will instruct the text analysis module to interpret the header as the initial sentence of the paragraph, rather than as a paragraph unto itself.
- Check that the text is primarily prose, as opposed to non-prose formats such as poetry, drama or graphic novels. Although the *TextEvaluator* tool is not structured to score poetry, a few lines of poetry within a larger passage are OK, as long as the lines are punctuated properly, i.e., retain all commas, colons, semicolons, question marks and periods, but remove extra line breaks so that the text reads like prose instead of poetry. Also, start and end the poem with end-of-paragraph markers so that the poem will be interpreted as a separate paragraph.
- Store each text as a separate .txt file with either of two possible encodings: ASCII or UTF-8.
- Documents stored as doc, docx, PDF, RTF, or HTML cannot be processed by the *TextEvaluator* technology as they require additional modules that are not currently available within the *TextEvaluator* on-line portal.

The *TextEvaluator*® Measurement Approach

Each new version of the *TextEvaluator*® scoring engine is implemented in five steps, as follows:

- Step 1. In this initial step, two corpora are assembled: one for use in model development, and one for use in model validation. Each corpus is designed to represent the types of texts considered by students at successive points in the progression from beginning reader to proficient, college-ready reader. Only passages with grade level (GL) or grade band (GB) classifications assigned by trained human experts are included, as these assignments are needed at subsequent stages of the model development and model validation processes.

- Step 2. A cognitive model of the processes engaged by readers during comprehension is specified, linguistic features that may facilitate or hinder the successful completion of each process are proposed, and a vector of cognitively-based feature scores is extracted from each text. The cognitive model underlying the current scoring engine is described in Sheehan (2016). The model specifies four processes that are critically involved in comprehending complex text:

  - inferring the meaning of individual words and phrases;
  - assembling words into sentences and then inferring the meaning of individual sentences;
  - retrieving information from previously processed sentences, and then inferring how each new sentence relates to the set of all sentences that have already been read; and
  - using knowledge of more or less familiar discourse structures to generate the additional inferences needed to form a coherent mental representation of the information, argument or story presented within the text.

- Step 3. In this step, feature weights determined via a Principal Components Analysis of a large corpus of texts (Sheehan, Kostin, Napolitano & Flor, 2014) are used to translate each text's vector of observed feature scores into a profile of eight component scores defined such that each component is focused on a single, construct-relevant dimension of text variation. Individual components include the classic readability dimensions of Vocabulary Difficulty and Syntactic Complexity (i.e., understanding words and sentences), as well as additional, cognitively-relevant dimensions such as Cohesion (understanding connections across sentences), and Degree of Narrativity (the extent to which the text follows a familiar narrative structure, see Sheehan, Kostin, Futagi and Flor, 2010).

- Step 4. In this step, analyses implemented only on the training data are used to model
- the GL classifications collected for each text conditional on the component scores estimated at Step 3. Since many important complexity measures are known to function differently within texts from different genres (Sheehan, Flor & Napolitano, 2013; Sheehan, Kostin & Futagi, 2008; Sheehan, Kostin, Futagi & Flor, 2010) three distinct prediction models are estimated: one optimized for application to informational texts, one optimized for application to literary texts, and one optimized for application to mixed texts, i.e., texts that incorporate a mixture of informational and literary elements. A more complete description of each genre category is provided below. As is illustrated in Sheehan (2016), Sheehan, Flor and Napolitano (2013), and Sheehan, Kostin, Napolitano & Flor (2014) this approach yields text complexity scores that exhibit little or no genre bias.
- Step 5. Each estimated prediction model is validated by examining the agreement between text complexity scores generated via the proposed prediction model, and text complexity classifications provided by human experts. Key results are summarized below, and in Sheehan (2016).

Additional information about the features, components and models described above is provided in a series of three different U. S. patents. The following section provides instructions for accessing relevant patent documents.

Relevant Text Analysis Patents

The innovative measurement approach incorporated within the *TextEvaluator* technology has been recognized by the U.S. Patent Office on three separate occasions. Resulting patents are listed in Table 1. One pending patent application is also listed.

Detailed descriptions of each awarded patent are available from the U.S. Patent Full Text Database. Individual documents may be accessed as follows:

- Go to patft.uspto.gov
- Select “Quick Search”
- Enter the Patent Number from Table 1 into the box labeled “Term 1”
- Select “Patent Number” in the box labeled “Field 1”
- Click “Search”

Table 1. List of *TextEvaluator*® Patents

(Both Awarded and Pending)

Patent Number	Date Awarded (or submitted if still pending)	Focus
8,517,738	9/27/2013	Measuring Overall Text Complexity when Genre DFF is present
8,888,493	11/18/2014	Measuring Overall Text Complexity when Genre DFF is present
8,892,421	11/18/2014	Measuring Cohesion
Pending	12/22/2016	Matching Readers to Texts

Note. DFF = Differential Feature Functioning. Approaches for detecting and addressing genre DFF are described in Sheehan (2016), Sheehan, Kostin, Futagi and Flor (2010), Sheehan, Flor and Napolitano (2013) and Sheehan, Kostin, Napolitano and Flor (2014).

TextEvaluator® Component Scores

The *TextEvaluator* tool employs a variety of natural language processing techniques to extract evidence of text standing relative to eight construct-relevant components of text variation. Each component focuses on one or another of the following types of cognitive processes: (1) understanding words, (2) understanding sentences, (3) inferring connections across sentences, and (4) using knowledge of discourse organization to generate the additional inferences needed to form a coherent mental representation of the text. Additional information about the specific aspects of text variation included within these four types of cognitive processes, and within each of eight component scores, is provided below.

Process #1: Understanding Words

Vocabulary difficulty has long been recognized as a key predictor of reading comprehension success or failure. A reader is more likely to build an accurate mental representation of the situation presented in a text if that situation is described via words that are already a part of her receptive vocabulary (Cohen & Steinberg, 1983). The *TextEvaluator* engine evaluates three components focused on this dimension of text variation. These three components are described below.

Academic Vocabulary. This component measures whether the words in each new text are more characteristic of academic texts than of nonacademic texts such as fiction or transcripts of conversations (Biber, et al., 2004; Coxhead, 2000).

Word Unfamiliarity. This component summarizes variation detected via two different word frequency indices: one developed at ETS from a corpus of 400 million words, and one developed by an outside firm from a corpus of 17 million words (Zeno, et al., 1995).

Concreteness. Words that are more concrete are more likely to evoke meaningful mental images, a response that has been shown to facilitate comprehension. The *TextEvaluator* tool addresses this particular component of text variation via a database of human concreteness ratings developed by Coltheart (1981). Resulting scores are expressed on a 1 to 100 scale structured such that texts with higher scores contain a higher proportion of more concrete words, and thus, are likely to present a *less difficult* comprehension problem, and texts with lower scores contain a higher proportion of more abstract words, and thus, are likely to present a *more difficult* comprehension problem.

Process #2: Understanding Sentences

The Understanding Sentences process is measured via a single component called the syntactic complexity component.

Syntactic Complexity. This component incorporates a variety of sentence-level features including the following: average sentence length, average number of modifiers per noun phrase, average number of dependent clauses per sentence, and an automated sentence "depth" measure similar to that introduced in Yngve (1960).

Process #3: Inferring Connections Across Sentences.

In addition to understanding the individual words and sentences comprising a text, successful readers must also determine how each new sentence relates to the set of all sentences that have already been read. This task will be more difficult under some conditions, and less difficult under others. The *TextEvaluator* system includes two components designed to characterize the ease or difficulty of inferring connections across sentences. These are summarized below.

Lexical Cohesion. Cohesion is that property of a text that enables it to be interpreted as a "coherent message" rather than a collection of unrelated clauses and sentences (Sheehan, 2013). Halliday and Hasan (1976) argued that readers are more likely to interpret a text as a "coherent message" when certain observable features are present. These include repeated instances of the same word stem (e.g., *read*, *reads* and *reading*), and explicit connectives (e.g., *consequently*, *as a result*, etc.). Lexical Cohesion is assessed via two features that measure the frequency of content word repetition across adjacent sentences within paragraphs. These measures differ from a similar set of measures described in Graesser et al., (2004), Pitler and Nenkova (2008) and Tierney and Mosenthal (1983) in that each is reported on an equated scale defined such that differences in genre and sentence length are accounted for (see Sheehan, 2013).

Level of Argumentation. This component measures the ease or difficulty of inferring connections across sentences when the underlying format of a text is argumentative, i.e., when the process of inferring needed connections involves following an author's line of reasoning. Texts with high Argumentation scores may be more difficult for readers who are not familiar with common argumentation strategies.

Process #4: Using Prior Knowledge about How Texts are Organized to Develop a More Complete, More Integrated Mental Representation of the Text

The manner in which texts are organized has frequently been linked to variation in comprehension ease or difficulty (Meyer, Brandt & Bluth, 1980). The *TextEvaluator* tool currently includes two components focused on this dimension.

Degree of Narrativity. This component characterizes the extent to which a given text exhibits features that are more characteristic of narrative text than of nonnarrative text.

Interactive/Conversational Style. This component measures the extent to which a text exhibits an interactive/conversational style as opposed to a non-interactive, non-conversational style. It includes a variety of indicators of conversation such as those described in Biber, et al. (2004).

Types of *TextEvaluator*® Feedback

A key advantage of the *TextEvaluator* measurement approach is that feedback about the sources of comprehension difficulty detected in each new text can be communicated to users at two levels of granularity: as a single, overall measure of text complexity, and as a profile of eight component scores. Additional information about each type is summarized below.

- **Overall Text Complexity Scores**: The *TextEvaluator* capability provides a single, overall measure of text complexity for each text. These are expressed on a numeric scale that ranges from 100 (appropriate for extremely young readers) to 2000 (appropriate for college graduates). Since this new scale has been linked to the Common Core Text Complexity Scale, publishers, teachers and test developers can use this feedback to determine an appropriate GL classification for any candidate reading passage. A Concordance Table for use in translating *TextEvaluator* Complexity Scores into Common Core GL classifications is provided below. Sheehan (2015) provides a detailed description of the methods used to develop and validate the *TextEvaluator*/Common Core alignment table.

- **A Profile of Eight Component Scores**. This profile is designed to help users understand *why* a text is rated as being more or less complex. Each component characterizes text standing relative to a single, construct-relevant source of comprehension ease or difficulty. The aspects of text variation measured by each component are described below in the section called "*TextEvaluator* Component Scores." All scores are expressed on a standardized scale that ranges from 1 to 100. In some cases, 1 indicates an extremely non-complex text, and 100 indicates an extremely complex text. In other cases the reverse is true. The specific scaling option employed for each component is indicated on the output display by an up arrow or a down arrow. An up arrow indicates that *higher* component score values are indicative of *higher levels of text* complexity, and a down arrow indicates that higher component score values are indicative of *lower levels of text* complexity.

Additional feedback is available for users who have purchased a *TextEvaluator* client code.

To learn more about purchasing a *TextEvaluator* client code, send an e-mail to TextEvalSupport@ets.org.

The additional feedback provided to users who have purchased a client code is described below in the section headed "Color Coded Component Score Classifications."

Color-Coded Component Score Classifications

Additional analyses are available for users who have purchased a client code. This additional feedback is designed to help users interpret the GL implications of the numeric values returned for each individual component score.

Since many important indicators of comprehension difficulty have been shown to function differently within informational and literary texts (Sheehan, 2016; Sheehan, et al., 2010; Sheehan, et al., 2013; Sheehan, et al., 2014), different classification rules have been developed for texts in each genre.

This additional feedback capability is implemented as follows. First, the user inputs a text and selects a target GL. Next, the text is assigned to one or another of two genres (informational or literary) and the component score values generated for the specified text are compared to the range of values expected for texts in that genre, at that targeted GL. Results are then classified into one or another of three possible categories (Green, Yellow or Red) as follows:

- A **Green Classification** implies that the value of the component is consistent with the range of variation typically observed among texts at the targeted grade level. Thus, a Green Classification suggests that the referenced component presents an appropriate level of complexity for readers at the targeted GL.

- A **Yellow Classification** implies that the value of the component is near the high end of the range of variation typically observed at the targeted GL. Thus, a Yellow Classification signals that the specified aspect of text variation is likely to be moderately challenging for readers at the targeted GL.

- A **Red Classification** implies that the value of the component is outside the range of variation expected at the targeted GL. Thus, a Red Classification signals that the specified aspect of text variation is likely to be extremely challenging for readers at the targeted GL.

Publishers, teachers and test developers can use these classifications to determine which of eight possible sources of comprehension difficulty are likely to be most challenging within any candidate reading passage.

The *TextEvaluator*® Genre Categories

The *TextEvaluato*r scoring engine classifies each text into one of three possible genre categories: informational, literary or mixed. Category definitions are based on the classification guidelines given in the 2009 NAEP Reading Framework (see American Institutes for Research, 2008).

Table 2 provides a brief description of each category.

Table 2. *TextEvaluator* Genre Categories

Category	Purpose of Text	Examples
Informational	To inform or persuade	excerpts from science and social studies textbooks, historical documents, newspaper editorials
Literary	To provide a rewarding literary experience	narratives, short stories, memoirs, essays with strong literary characteristics
Mixed	Both to inform or persuade and to document personal experiences	blogs, informative magazine articles written from a personal perspective

Using *TextEvaluator*® to Evaluate Book-Length Texts

The *TextEvaluator* engine evaluates book-length texts on a chapter-by-chapter basis. The following guidelines should be followed whenever book-length texts are evaluated:

- First, separate the book into individual chapters; format each chapter according to the general formatting guidelines specified above; and store each chapter in a separate .txt file (using either ASCII or UTF-8 encoding).
- Second, submit each chapter to the *TextEvaluator* tool as a separate text, and take note of the text complexity scores returned for each chapter.
- Finally, generate a text complexity score for the text as a whole by taking the median of all available chapter scores.

Figure 1 illustrates this procedure for two books: *I am Malala* by Malala Yousafzai (left), and *Living History* by Hillary Clinton (right). A dashed line shows the median text complexity score calculated for each book.

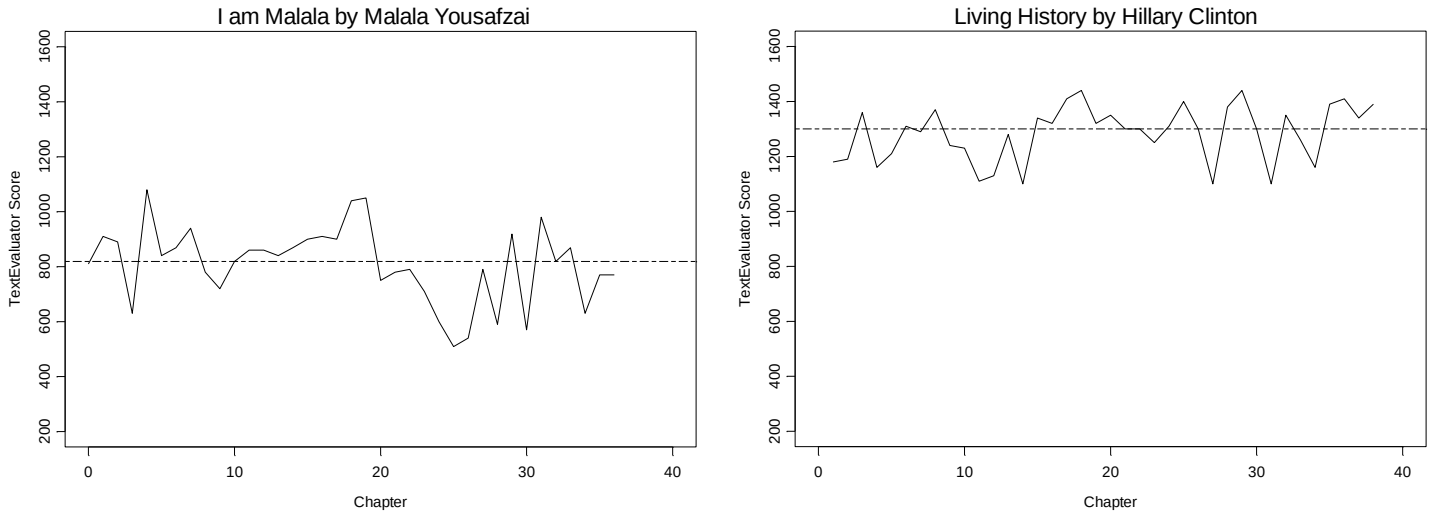

Figure 1. *TextEvaluator* scores, by chapter, and book-level median text complexity scores for two books: *I am Malala*, by Malala Yousafzai (left), and *Living History* by Hilary Clinton (right). Texts for these analyses were obtained from a corpus prepared by Dr. David Kaufer, Mellon Distinguished Professor of English, Carnegie Mellon University.

TextEvaluator®/Common Core Concordance Table

The overall text complexity scores obtained via the *TextEvaluator* tool are expressed on a quantitative scale that ranges from 100 to 2000. This new scale has been linked to the complexity scale described in the new Common Core State Standards (Common Core State Standards Initiative, 2010). Alignment results are summarized below. The table shows the range of *TextEvaluator* scores associated with texts at successive Common Core grade levels in the range from Grade 2 to Grade 12 (also called College and Career Ready or CCR).

Table 3. *TextEvaluator* to Common Core Concordance

Common Core Grade Level	*TextEvaluator* Score Range (100 - 2000 Scale)
2	100 – 525
3	310 - 590
4	405 - 655
5	480 - 720
6	550 - 790
7	615 - 860
8	685 - 940
9	750 - 1025
10	820 - 1125
11	890 - 1245
12	970 - 1360

Note. The method used to generate these score ranges is documented in Sheehan (2015).

Validity Evidence

Evidence about the validity of *TextEvaluator* scores has been reported in a number of different studies. The following table presents correlations between *TextEvaluator* scores and GL classifications provided by human experts. Note that relatively high correlations were obtained in each of three independent validation datasets.

The lower correlations observed for the Common Core texts are due to the fact that the human-generated GL classifications provided for these texts are reported on a five-point scale, as opposed to the 10 or 12 point scales employed in the other two validation datasets.

Table 4. Correlation Between *TextEvaluator* Scores, and Grade Level Classifications Provided by Human Experts

Source	No of Texts	Included During Model Training?	Spearman Correlation
State, National & College Admissions Assessments	941	Yes	0.83
Stanford Achievement Test, Version 9	59	No	0.89
Exemplar Passages from Appendix B, CCSS	128	No	0.72
Exemplar Passages from Chall, et al. (1996)	52	No	0.93

Note: Separate correlations are reported for each of three different validation datasets because the human generated text complexity classifications obtained for these texts are not necessarily expressed on a common scale.

References

American Institutes for Research (2008). *Reading framework for the 2009 National Assessment of Educational Progress.* Washington, DC: National Assessment Governing Board.

Biber, D., Conrad, S., Reppen, R., Byrd, P., Helt, M., Clark, V., et al., (2004). *Representing Language Use in the University: Analysis of the TOEFL 2000 Spoken and Written Academic Language Corpus*. TOEFL Monograph Series, MS-25, January 2004. Princeton, NJ: ETS.

Chall, J. S., Bissex, G. L., Conrad, S. S., & Harris-Sharples, S. (1996). Qualitative assessment of text difficulty: A practical guide for teachers and writers. Cambridge, MA: Brookline Books.Chall,

Cohen, S. A. & Steinberg, J. E. (1983). Effects of three types of vocabulary on readability of intermediate grade science textbooks: An application of Finn's transfer feature theory. *Reading Research Quarterly, 19*(1), 86-101.

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

Common Core State Standards Initiative (2010, June). *Common core state standards for English language arts & literacy in history/social studies, science & technical subjects*. Washington, DC: Council of Chief State School Officers & National Governors Association.

Coxhead, A. (2000). A new academic word list. *TESOL Quarterly, 34*, 213-238.

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

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

Meyer, B.J.F., Brandt, D.M., & Bluth, G.J. (1980). Use of top-level structure in text: Key for reading comprehension of ninth-grade students, *Reading Research Quarterly, 16*(*1*), 72-103.

Pitler, E. & Nenkova, A. (2008). Revisiting readability: A unified framework for predicting text quality. In *Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, 186-195.

Sheehan, K. M. (2013). Measuring cohesion: An approach that accounts for differences in the degree of integration challenge presented by different types of sentences. *Educational Measurement: Issues and Practice,* 32(4), 28-37.

Sheehan, K. M. (2015). *Aligning TextEvaluator scores with the accelerated text complexity guidelines specified in the Common Core State Standards*. (ETS Research Report No. RR-15-21). Princeton, NJ: Educational Testing Service.

Sheehan, K. M. (2016). *A review of evidence presented in support of three key claims in the validity argument for the TextEvaluator text analysis tool*. (ETS Research Report No. RR-16-12). Princeton, NJ: Educational Testing Service.

Sheehan, K. M., Flor, M. & Napolitano, D. (2013). A two-stage approach for generating unbiased estimates of text complexity. In the Proceedings of the 2nd *Workshop on Natural Language Processing for Improving Textual Accessibility (NLP4ITA), Annual Conference of the Association for Computational Lingusitics,* Atlanta, GA, pp. 49-58.

Sheehan, K. M., Kostin, I. & Futagi, Y. (2008). When do standard approaches for measuring vocabulary difficulty, syntactic complexity and referential cohesion yield biased estimates of text difficulty? In B.C. Love, K. McRae & V.M. Sloutsky (Eds.), *Proceedings of the 30th Annual Meeting of the Cognitive Science Society*, Washington D.C., pp. 1978-1983.

Sheehan, K. M., Kostin, I, Futagi, Y. & Flor, M. (2010). *Generating automated text complexity classifications that are aligned with published text complexity standards*. (ETS Research Report No. RR-10-28). Princeton, NJ: Educational Testing Service.

Sheehan, K. M., Kostin, I., Napolitano, D. & Flor, M. (2014). The TextEvaluator Tool: Helping teachers and test developers select texts for use in instruction and assessment. *Elementary School Journal*, 115(2), 184-209.

Tierney, R. J., & Mosenthal, J. H. (1983). Cohesion and textual coherence. *Research in the Teaching of English*, 17, 215-229.

Yngve, V.H. (1960). A model and hypothesis for language structure. *Proceedings of the American Philosophical Society. 104*, 444-466.

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