Supplemental Materials

Important Text Characteristics for Early-Grades Text Complexity

by J. Fitzgerald et al., 2014, *Journal of Educational Psychology*

http://dx.doi.org/10.1037/a0037289

Supplemental Material 1: Details About Substudies Conducted to Create the Text-Complexity Outcome Measure

Text-Complexity Level

The outcome variable was early-reader text-complexity level measured using a continuous, developmental scale, with scores ranging from 0 to 100. An overview of the scale-building procedures is as follows, with details in following paragraphs. Because text complexity was defined at the intersection of printed texts with students reading them for particular purposes and doing particular tasks, a multiple-perspective measure of text complexity was created using student responses during a reading task and teachers' ordering of texts according to complexity. In a first substudy, through Rasch modeling (Bond & Fox, 2007) a text-complexity logit scale was created from the interface of children reading texts. That is, complexity was in part defined according to children's responses while reading texts in our study. In a second substudy, also through Rasch modeling, a text-complexity logit scale was created from teachers' evaluations of texts' complexity. Then the magnitude and strength of the association between the two logit scales was examined, and to arrive at a single scale, a linear equating linking procedure (Kolen & Brennan, 2004) was used to bring the student results onto a common scale with the teacher results. Finally, for ease of interpretability, the logit scale was linearly transformed to a 0 to 100 scale.

For the first substudy 1,258 first and second graders from 10 U.S. states read texts from a subset of the 350 texts, and completed a maze task (Shin, Deno, & Espin, 2000). Of the students for whom ethnicity data were reported (n = 1,221), 15% were African American; 6% were Asian; 70% were Caucasian; 4% were Latino; and 5% were American Indian, Hawaiin/Pacific Islander, or mixed ethnicity. Of the students for whom English-language learner status was reported (n = 504), 19% were English-language learners. A random sample of 90 texts was selected from the 350 texts used in the present study, stratified by the six categories and by publisher-designated difficulty level. One passage was later rejected due to a printing error. Six test forms were created by randomly assigning eight (first grade) or seven (second grade) passages to a form. Passages were 75 (first grade) or 150 (second grade) words long, randomly generated through a computer program. Each form was replicated with a new item set to create 12 forms per grade. Maze items (a blank with a multiple choice for the removed word) for first and second grade, respectively, were inserted at seven-word and 10-word intervals, plus or minus one word randomly to avoid syncing exactly with seven- or 10-word interval repeated phrase/sentence patterns. Form administration was counterbalanced across students with teachers reading standardized directions to large groups of students. From the student responses, a logit scale was created, and the 89 texts were assigned logits for text-complexity level. Cronbach's alpha estimates of reliability for the forms ranged from .85 to .96. Also, using the student responses, dimensionality assessments for text genre and for differential text ordering according to student ethnicity, gender, or free-reduced-lunch status suggested no evidence of measurement multidimensionality.

For the teacher-judgment substudy, teachers were solicited through an existing nationwide e-mail listserv. Initially 250 early-grades educators expressed interest, and once given specific information about the purpose and task involved, as well as benefits of participating (a set of classroom books was given to each teacher), 90 teachers from 33 states and 75 school districts chose to participate. On the whole, the sample was experienced, and they taught in urban or suburban public schools. Slightly more than half came from schools where 50% or more of the students received free or reduced lunch. The texts (including images) were digitally scanned, and through computer programming, excerpts were randomly selected and positioned so that teachers could see two texts side by side on a computer screen. Pairs were randomly computer-generated in the moment so that teachers could receive different sets of pairs. On average, each teacher saw a total of 125 comparisons involving 35 books. After scrolling through each pair of texts, teachers clicked a button at the bottom of the screen to indicate that text they thought was more complex. From the teachers' responses, a Rasch-modeled logit scale was created, and the 350 texts were assigned logit scores for text-complexity level. Using the separation index method (Wright & Stone, 1999), measurement reliability was .99.

Next, the correlation between the two logit scales (N = 89 texts) was .79 ($p < .01$), suggesting that the texts ordered on text complexity similarly whether teachers or students were involved. The relatively high correlation was also evidence of concurrent validity in that it suggested that the two logit scales were measuring the same construct. Consequently, a linking equating procedure was used to link the two logit scales (Kolen & Brennan, 2004). Finally, a linear transformation was done resulting in measures that could range from 0 to 100 on a text-complexity scale. That is, the 350 texts ordered by teachers could be assigned a measure from 0 to 100, and the texts read by students could be assigned a measure from 0 to 100.

References

Bond, T. G., & Fox, C. M. (2007). *Fundamental measurement in the human sciences* (2nd ed.). Mahwah, NJ: Erlbaum.

Kolen, M. M., & Brennan, R. L. (2004). *Test equating, scaling, and linking: Methods and practices* (2nd ed.). New York, NY: Springer-Verlag.

Shin, J., Deno, S. L., & Espin, C. (2000). Technical adequacy of the maze task for curriculum-based measurement of reading growth. *The Journal of Special Education, 34,* 164–172.

Wright, B., & Stone, M. (1999). *Measurement essentials* (2nd ed.). Wilmington, DE: Wide Range.

Supplemental Material 2

Table S1

Text Characteristics by Linguistic Level, Definitions (Sources), Operationalization Examples, and Variable Operationalizations

Linguistic Level	Text Characteristics	Definition (Source)	Possible Score Range (and Interpretation) for Final Nine Most Important and a Few Additional Operationalization Examples	Variable Operationalizations
Sounds in Words	Number of phonemes in words	The smallest unit of sound. (The MRC Psycholinguistic Database provides the phoneme values for words [Coltheart, 1981].)	1 (fewer phonemes in words, less complex) to less than 10 (more phonemes in words, more complex)	Mean number of phonemes for words in the text Mean with stop list 50 most frequent Types as test (ability 50%) Types as test with stop list 50 most frequent (ability 50%) Types as test (ability 75%) Types as test with stop list 50 most frequent (ability 75%) Types as test (ability 90%) Types as test with stop list 50 most frequent (ability 90%) Words as test (ability 50%)

				Words as test stop list 50 most frequent (ability 50%) Words as test (ability 75%) Words as test stop list 50 most frequent Words as test (ability 90%) Words as test with stop list 50 most frequent (ability 90%)
	Phonemic Levenshtein Distance	The degree to which co-occurring phonemes exist across words. (Levenshtein Distance is a standard computer metric of string edit distance that gauges the minimum number of substitution, insertion, or deletion operations required to turn one word into another. Measures phonemic similarity across words for the 20 closest words. [Levenshtein, 1965; Yarkoni, Balota, & Yap, 2008; Cf., Kruskal, 1999; Nerbonne & Heeringa, 2001; Sanders & Chinn, 2009].)		Mean Mean with stop list 50 most frequent Types as test (ability 50%) Types as test with stop list 50 most frequent (ability 50%) Types as test (ability 75%) Types as test with stop list 50 most frequent (ability 75%) Types as test (ability 90%) Types as test with stop list 50 most frequent (ability 90%) Words as test (ability 50%) Words as test stop list 50 most frequent (ability 50%) Words as test (ability 75%) Words as test stop list 50 most frequent Words as test (ability 90%) Words as test with stop list 50 most frequent (ability 90%)
	Mean Internal Phonemic Predictability	The degree to which phoneme collocations occur given the	0 (fewer phoneme collocations are repeated in the text)	Mean with chunk size 125 Mean with chunk size 125 and with stop list 50 most frequent

		totality of the phoneme collocations in the particular text. The frequencies of phoneme collocations for words in the particular text are determined. Then examining each word's phonemes, for three-phoneme collocations, what is the probability that the tri-phoneme collocation occurs in the text? (Words are converted to phonemes using the CMU [Carnegie Mellon University] Pronouncing Dictionary [Carnegie Mellon University, n.d.].)	to 1 (more phoneme collocations are repeated in the text)	Product with chunk size 125 Product with chunk size 125 and with stop list 50 most frequent
Word Structure	Decoding demand	The decoding demand of the words in the text. (Slight modification of Menon & Hiebert's [1999] decodability scale.) Sample levels are: Level 1: A, I and C-V (examples, A, I, me,	1 (less complex word structure) to 9 (most complex word structure)	Mean *Mean with stop list 50 most frequent* Percentage of sentences with 1 word over score of 4 Percentage of sentences with 1 word over score of 5 Percentage of sentences with 1 word over score of 6 Percentage of sentences with 1 word over score of 7

	we, my, so)	Percentage of sentences with 1 word over score of 4
	Level 4: (C)-(C)-(C)-V-C-e (examples, bake, ride, plate)	Percentage of sentences with 1 word over score of 5
		Percentage of sentences with 1 word over score of 6
	Level 7: Diphthongs (examples, boy, draw)	Percentage of sentences with 1 word over score of 7
		Types as test (ability 50%)
	Level 8: Multisyllabic words	Types as test with stop list 50 most frequent (ability 50%)
		Types as test (ability 75%)
	Level 9: Other more difficult	Types as test with stop list 50 most frequent (ability 75%)
		Types as test (ability 90%)
		Types as test with stop list 50 most frequent (ability 90%)
		Words as test (ability 50%)
		Words as test stop list 50 most frequent (ability 50%)
		Words as test (ability 75%)
		Words as test stop list 50 most frequent
		Words as test (ability 90%)
		Words as test with stop list 50 most frequent (ability 90%)
Orthographic Levenshtein Distance	Levenshtein Distance is a standard computer metric of string edit distance that gauges the minimum number of substitution, insertion, or deletion operations required to turn one word into the	Mean
		Mean with stop list 50 most frequent
		Types as test (ability 50%)
		Types as test with stop list 50 most frequent (ability 50%)
		Types as test (ability 75%)
		Types as test with stop list 50 most frequent (ability 75%)
		Types as test (ability 90%)

	other. Measures orthographic similarity across words for the 20 closest words. (Levenshtein, 1965; cf. Kruskal, 1999; Yarkoni, et al., 2008.)		Types as test with stop list 50 most frequent (ability 90%)
			Words as test (ability 50%)
			Words as test stop list 50 most frequent (ability 50%)
			Words as test (ability 75%)
			Words as test stop list 50 most frequent
			Words as test (ability 90%)
			Words as test with stop list 50 most frequent (ability 90%)
Number of Syllables in Words	Number of syllables in words. (The MRC Psycholinguistic Database provides syllable values for words [Coltheart, 1981].)		Mean
			Mean with stop list 50 most frequent
			Percent of sentences with one word of more than 1 syllable
			Percent of sentences with one word of more than 2 syllables
			Percent of sentences with two words of more than 1 syllable
			Percent of sentences with two words of more than 2 syllables
			Types as test (ability 50%)
			Types as test with stop list 50 most frequent (ability 50%)
			Types as test (ability 75%)
		1 (few words with many syllables) to 8 (more words with more syllables) (0 if all the words in the text are on the stop list)	*Types as test with stop list 50 most frequent (ability 75%)*
			Types as test (ability 90%)
			Types as test with stop list 50 most frequent (ability 90%)
			Words as test (ability 50%)
			Words as test stop list 50 most frequent (ability 50%)
			Words as test (ability 75%)
			Words as test stop list 50 most frequent

			Words as test (ability 90%) Words as test with stop list 50 most frequent (ability 90%)
Mean Internal Orthographic Predictability	The degree to which letter collocations occur in a text given the totality of the letter collocations in the particular text. The frequencies of letter collocations for words in the particular text are determined. Then examining each word, what is the probability that the tri-gram occurs in the text? (Researcher computer coded; Cf. Solso, Barbuto, & Juel, 1979).	0 (fewer orthographic trigrams are repeated in the text) to 1 (more orthographic trigrams are repeated in the text)	Mean with chunk size 125 Mean with chunk size 125 and with stop list 50 most frequent Product with chunk size 125 Product with chunk size 125 and with stop list 50 most frequent
Sight Words	The most commonly occurring words in primary grades texts. Children are expected to be able to look at and pronounce them within one-quarter second, generally all of them on the lists by end of third grade. (Dolch word list, n.d.; Fry Word List, n.d.)		*Dolch List:* Percent of words in text on Preprimer list Percent on Primer list Percent on Dolch list 1 Percent on Dolch list 2 Percent on Dolch list 3 Percent on all lists *Fry List:*

				Percent of words in text on Fry list 100 Percent on Fry list 200 Percent on Fry list 300 Percent on Fry list 400 Percent on Fry list 500 Percent on Fry list 600 Percent on all lists
Word Meaning	Age of Acquisition	Age at which a word's meaning is first known. (Kuperman, Stadthagen-Gonzalez, & Brysbaert, 2012.)		Mean Mean with stop list 50 most frequent Percent of sentences with 1 word over 4 years old Percent of sentences with 1 word over 5 years old Percent of sentences with 1 word over 6 years old Percent of sentences with 1 word over 7 years old Percent of sentences with 1 word over 8 years old Percent of sentences with 2 words over 4 years old Percent of sentences with 2 words over 5 years old Percent of sentences with 2 words over 6 years old Percent of sentences with 2 words over 7 years old Percent of sentences with 2 words over 8 years old Types as test (ability 50%)
			1 to 25 in our study (lower score means more of the words are known by	*Types as test with stop list 50 most frequent (ability 50%)*
				Types as test (ability 75%) Types as test with stop list 50 most frequent

			younger readers and a higher score means fewer of the words are known by younger readers)	(ability 75%)
				Types as test (ability 90%)
				Types as test with stop list 50 most frequent (ability 90%)
				Words as test (ability 50%)
				Words as test stop list 50 most frequent (ability 50%)
				Words as test (ability 75%)
				Words as test stop list 50 most frequent
				Words as test (ability 90%)
				Words as test with stop list 50 most frequent (ability 90%)
	Abstractness	Degree to which the text contains words that reference general or complex concepts such as “honesty” and cannot be seen or imaged. (Index of abstractness, Paivio, Yuille, & Madigan, 1968, updated in the MRC Psycholinguistic Database [Coltheart, 1981]).		Mean
				Mean with stop list 50 most frequent
				Percent of sentences with 1 word with score over 200
				Percent of sentences with 1 word with score over 400
				Percent of sentences with 1 word with score over 600
				Percent of sentences with 2 words with score over 200
				Percent of sentences with 2 words with score over 400
				Percent of sentences with 2 words with score over 600
				Types as test (ability 50%)
			0 (less abstract, less complex) to 700 (more abstract, more complex)	*Types as test with stop list 50 most frequent (ability 50%)*
				Types as test (ability 75%)
				Types as test with stop list 50 most frequent (ability 75%)
				Types as test (ability 90%)
				Types as test with stop list 50 most frequent

				(ability 90%) Words as test (ability 50%) Words as test stop list 50 most frequent (ability 50%) Words as test (ability 75%) Words as test stop list 50 most frequent Words as test (ability 90%) Words as test with stop list 50 most frequent (ability 90%)
	Word Rareness	The inverse of the frequency with which a word appears in running text in a corpus of 1.39billion words from 93,000 kindergarten through university texts normalized to link to the frequencies in the Carroll, Davies, & Richman frequency 5million word list. (MetaMetrics, n.d.; Carroll, Davies, & Richman, 1971.)	0 (less rare) to 6 (more rare) (Reverse scored from frequency)	Mean Mean with stop list 50 most frequent Types as test (ability 50%) Types as test with stop list 50 most frequent (ability 50%) Types as test (ability 75%) Types as test with stop list 50 most frequent (ability 75%) *Types as test (ability 90%)* Types as test with stop list 50 most frequent (ability 90%) Words as test (ability 50%) Words as test stop list 50 most frequent (ability 50%) Words as test (ability 75%) Words as test stop list 50 most frequent Words as test (ability 90%) Words as test with stop list 50 most frequent (ability 90%)
Within Sentence/Syntax	Sentence Length	Number of characters, words, unique words, or phrases in a sentence. (Researcher computer coded).	1 (fewer characters, words, unique words, or phrases) and above 1 (more)	*Characters:* Mean number of letters and spaces in sentences Mean number of letters in sentences *Tokens and types:*

			Mean number of words in sentences Log of mean number of words in sentences with slice 125 Mean number of unique words in sentences. *Phrases:* Mean number of phrases in sentences
Grammar	Link Type, which is a linguistic convention that ties a word in a sentence to another word in the sentence. e.g., one link type connects adjectives to verbs in cases where the adjective is fronted, such as in questions and indirect questions like "How BIG IS it?" (Sleator & Temperley, 1991; Temperley, Sleator, & Lafferty, 2012; Definitions of all link types can be found at http://www.link.cs.cmu.edu/link/dict/summarize-links.html.)		*Unique Link Types:* Mean number of unique link types in sentences
	Distance to Verb, which is the distance from the beginning of a sentence to the first verb. (Bird, Loper, Klein, 2009; cf.		*Distance to Verb:* Mean distance to first verb in a sentence with slice 125

		Maximum Entropy POS- [Part of Speech] Tagging Model, n.d.; Collins, 2002.)		
		Link Distance: The average number of words between linked words within a sentence. (Sleator & Temperley, 1991; Temperley, Sleator, & Lafferty, 2012.)		Mean of distances between links averaged across all sentences
Discourse (Across Sentences)				
Intersentential Complexity	Linear Edit Distance	The degree of word, phrase, and letter pattern repetition across *adjacent* sentences. The number of single character replacements required to turn one sentence into the next one. (Levenshtein, 1965.) Ex., "This is my pretty coat. This my pretty coat." (score 0) "This is my pretty coat. This is my pretty	0 (if all sentences are identical or there is only one sentence; lots of redundancy, less complex) to approximately 110 in our study (not much redundancy, more complex)	*Lexical Emphasis/Linear* *Mean linear edit distance* Mean linear edit percentage *Syntactic Emphasis/Linear* Mean linear edit distance for part of speech Mean linear edit percentage for part of speech

	hat." (score 2)	
Linear Word Overlap	Degree to which unique words in a first sentence are repeated in a following sentence, comparing sentence pairs sequentially. (Researcher computer coded.)	<u>*Lexical Emphasis/Linear*</u> Mean linear word overlap with slice 125 Mean linear percentage word overlap with slice 125 Mean of upper quartile Cartesian word overlap with slice 125 <u>*Syntactic Emphasis/Linear*</u> Mean linear word overlap with slice 125 for part of speech Mean linear percentage word overlap with slice 125 for part of speech Mean of upper quartile Cartesian word overlap with slice 125 for part of speech
Cohesion Triggers	Words that indicate occurrence of cohesion in text. Five categories of cohesive devices between words in text work to hold a text together. e.g., In the following sentences, "She" is an anaphoric cohesive tie with "Susie." "Susie away. She was unhappy." Cohesion trigger words are words that typically link with other words in the text. "She" in the preceding two example sentences is a	<u>*Lexical Emphasis/Context*</u> Percent of words in text that are on the cohesion trigger word list

		cohesion trigger word. (Cf. Halliday & Hasan, 1976; Researcher devised beginning with words listed at: Cohesion [linguistics], n.d.)		
Lexical/Syntactic Diversity	Type-Token Ratio	An indicator of word diversity, or the number of unique words in a text divided by the total number of words in a text. (Cf. Malvern, Richards, Chipere, & Durán, 2009.)		*Lexical Emphasis Context* Type-token ratio with chunk 125 Type-token ratio with chunk 125 and stop list 50 most frequent
Phrase Diversity	Longest Common String	Degree of word, phrase, and letter pattern repetition across *multiple* sentences. Captures couplets and triplets. (Gusfield, 1997)	0 (a lot of overlap, a lot of redundancy across multiple sentences, less complex) to 1 (not much overlap, more complex)	*Lexical Emphasis/Context* Cartesian LCSequence percentage with slice 125 Cartesian LCSubsequence with slice 125 Cartesian LCSubstring with slice 125 Mean of upper quartile Cartesian LCSubstring with slice 125 Mean linear LCS percentage with slice 125 *Mean Cartesian LCS percentage with slice 125* Mean LCSubsequence percentage with slice 125 Mean of upper quartile Cartesian LCSubsequence percentage with slice 125 Mean of upper quartile Cartesian LCSubsequence percentage with slice 125 for part of speech Mean linear LCSubsequence with slice 125

		Mean LCSubstring with slice 125 Mean upper quartile Cartesian LCSubstring percentage with slice 125 *Syntactic Emphasis/Context* Cartesian LCSequence percentage with slice 125 for part of speech Cartesian LCSubsequence percentage with slice 125 for part of speech Cartesian LCSubstring with slice 125 for part of speech Mean of upper quartile Cartesian LCSubstring with slice 125 for part of speech Mean linear LCS percentage with slice 125 for part of speech Mean linear LCSubsequence percentage with slice 125 for part of speech Mean linear LCSubsequence with slice 125 for part of speech Mean LCSubstring with slice 125 for part of speech Mean upper quartile Cartesian LCSubstring percentage for part of speech
Edit Distance	Number of single character additions, deletions, or replacements required to turn one string (or sentence) into another. (Levenshtein, 1965; Kruskal, 1999.)	*Lexical Emphasis/Context* Mean Cartesian edit distance with slice 125 Mean of lower quartile Cartesian edit distance with slice 125 Mean Cartesian edit percentage with slice 125 Mean of lower quartile Cartesian edit percentage with slice 125 *Syntactic Emphasis/Context*

				Mean Cartesian edit distance with slice 125 for part of speech Mean of lower quartile Cartesian edit distance with slice 125 for part of speech Mean Cartesian edit percentage with slice 125 for part of speech Mean of lower quartile Cartesian edit percentage with slice 125 for part of speech
	Cartesian Word Overlap	Degree to which unique words in a first sentence are repeated in a following sentence comparing all possible pairs in a 125 slice. (Researcher computer coded.)		*Lexical Emphasis/Context* Mean Cartesian word overlap with slice 125 Percentage Cartesian word overlap with slice 125 for part of speech *Syntactic Emphasis/Context* Mean of Cartesian word overlap with slice 125 for part of speech Percentage of Cartesian word overlap with slice 125 for part of speech
Text Density	Information Load	Total information load in text. Denser texts have more information load, less redundancy, and are more complex. Also taps overlap of *groups* of co-occurring word repetition. (Researcher devised incorporating Latent Semantic Analysis [Deerwester, Dumais, Furnas,Landauer, Harshman, 1990;	0 (low density, low information load, lots of novel co-occurring word-group repetition) to 1 (denser text, higher information load, not as much novel co-occurring word-group repetition)	*Lexical Emphasis/Context* *Normalized percent reduction of information load across sentences* for 10 dimensions with slice 125 . . . *for 10 dimensions with slice 500* . . . for 5 dimension with slice 125 . . . for 5 dimensions with slice 500 . . . for 3 dimensions with slice 125 . . . for 3 dimensions with slice 500 Number of dimensions to capture 5% of content word space across sentences with slice 125 . . . with slice 500 Number of dimensions to capture 7% . . . with slice 125

		Landauer & Dumais, 1997].) Ex. “Mat. Mat sat. Sam. Sam Sat. Mat sat. Mat sat on Sam. Sam sat on Mat. Mat sat. Sam sat.” (score .28) “Button. I did it. Pull. I did it. Tie. I did it. Zip. I did it. Snap. I did it. Open. I did it. Wait! Hug. I did it!” (score .58)		. . . with slice 500 Number of dimensions to capture 9% . . . with slice 125 . . . with slice 500
Non-Compressibility	Compression Ratio	The degree to which information in the text can be compressed. Novel text is less compressible. (Burrows & Wheeler, 1994.)	0 (more compressible, more redundancy, less complex) to 1 (less compressible, less redundancy, more complex)	$\frac{\textit{Compression ratio with slice 125}}{\textit{Compression ratio with chunk 125}}$

References

Bird, S., Loper, E., & Klein, E. (2009). *National language processing with Python.* Sebastopol, CA: O'Reilly Media.

Burrows, M., & Wheeler, D. J. (1994). *A block sorting lossless data compression algorithm* (Technical Rep. no. 124). Maynard, MA: Digital Equipment Corporation.

Carnegie Mellon University. (n.d.) *CMU pronouncing dictionary.* Retrieved from http://www.speech.cs.cmu.edu/cgi-bin/cmudict

Carroll, J. B., Davies, P., & Richman, B. (1971). *The American heritage word frequency book.* New York, NY: American Heritage.

Cohesion (linguistics). (n.d.) In *Wikipedia.* Retrieved from http://en.wikipedia.org/wiki/Cohesion_%28linguistics%29

Collins, M. (2002, July). Discriminative training methods for hidden Markov models: Theory and experiments with perceptron algorithms. In J. Hajič & Y., Matsumoto (Eds.), *Proceedings of the conference on empirical methods in natural language processing* (pp. 1–8). Philadelphia, PA: Special Interest Group on Linguistic Data and Corpus-Based Approaches to NLP (SIGDAT).

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

Deerwester, S., Dumais, S. T., Furnas, G. W., Landauer, T. K., & Harshman, R. (1990). Indexing by latent semantic analysis. *Journal of the American Society for Information Science, 41,* 391–407.

Dolch word list. (n.d.). In *Wikipedia.* Retrieved from http://en.wikipedia.org/wiki/Dolch_word_list

Fry Word List—1,000 High Frequency Words. (2012). In *K12Reader: Reading instruction resources for teachers & parents.* Retrieved from http://www.k12reader.com/fry-word-list-1000-high-frequency-words/

Gusfield, D. (1997). *Algorithms on strings, trees and sequences: Computer science and computational biology*. Cambridge, England: University of Cambridge.

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

Kruskal, J. B. (1999). An overview of sequence comparison. In D. Sankoff & J. B. Kruskal (Eds.), *Time warps, string edits, and macromolecules: The theory and practice of sequence comparison* (pp. 1–44). Stanford, CA: Center for the Study of Language and Information.

Kuperman, V., Stadthagen-Gonzalez, H., & Brysbaert, M. (2012). Age-of-acquisition ratings for 30,000 English words. *Behavior Research Methods, 44,* 978–990.

Landauer, T. K., & Dumais, S. T. (1997). A solution to Plato's problem: The latent semantic analysis theory of acquisition, induction and representation of knowledge. *Psychological Review, 104,* 211–240.

Levenshtein, V. I. (1965). Binary codes capable of correcting deletions, insertions, and reversals. *Doklady Akademii Nauk SSSR, 163,* 845–848.

Malvern, D. D., Richards, B. J., Chipere, N., & Durn, P. (2009). *Lexical diversity and language development: Quantification and assessment.* New York, NY: Palgrave Macmillan.

Maximum Entropy POS-Tagging Model. (n.d.). http://acl.ldc.upenn.edu/W/W96/W96-0213.pdf

Menon, S., & Hiebert, E. H. (1999). *Literature anthologies: The task for first-graders.* Ann Arbor, MI: Center for the Improvement of Early Reading Achievement.

MetaMetrics. (n.d.). *Word corpus.* Durham, NC: Author.

Nerbonne, J., & Heeringa, W. J. (2001). Computational comparison and classification of dialects. *Dialectologia et Geolinguistica, 9,* 69–83.

Paivio, A., Yuille, J. C., & Madigan, S. A. (1968). Concreteness, imagery, and meaningfulness values for 925 nouns [Monograph]. *Journal of Experimental Psychology, 76*(1, Pt. 2), 1–25.

Sanders, N. C., & Chinn, S. B. (2009). Phonological distance measures. *Journal of Quantitative Linguistics*, *16*, 96–114.

Sleator, D., & Temperley, D. (1991, October). *Parsing English with a link grammar* (Carnegie Mellon University Computer Science Technical Rep. no. CMU-CS-91-196). Pittsburgh, PA: Carnegie Mellon University.

Solso, R. L., Barbuto, P. F., Jr., & Juel, C. L. (1979). Methods & designs: Bigram and trigram frequencies and versatilities in the English language. *Behavior Research Methods & Instrumentation*, *11*, 475–484.

Temperley, D., Sleator, D., & Lafferty, J. (2012). *Link grammar*. Retrieved from http://www.link.cs.cmu.edu/link/

Yarkoni, T., Balota, D., & Yap, M. (2008). Moving beyond Coltheart's *N*: A new measure of orthographic similarity. *Psychonomic Bulletin & Review*, *15*, 971–979.

Supplemental Material 3

Table S2

Discourse Variable Families and Lexical Versus Syntactic Emphases in Operationalizations

		Emphases in Variable Operationalizations			
		Lexical Emphasis		Syntactic Emphasis	
Five Discourse Variable Families	Text Characteristic	Linear/Adjacent Sentences	Cartesian/Context Larger than Adjacent Sentences	Linear/Adjacent Sentences	Cartesian/Context Larger than Adjacent Sentences
Intersentential	Linear Edit Distance	Y		Y	
Complexity	Linear Word Overlap	Y		Y	
	Cohesion Triggers		Y		

Lexical/Syntactic Diversity	Type-Token Ratio	Y	
Phrase Diversity	Longest Common String	Y	Y
	Edit Distance	Y	Y
	Cartesian Word Overlap	Y	Y
Text Density	Information Load	Y	
Non-Compressibility	Compression Ratio	Y	

Note. Y = Yes, operationalizations were employed for the specific emphasis.

Supplemental Material 4: Random Forest Regression: Comparison to Linear Regression

To better understand random forest regression, and partly to better understand why it is potentially beneficial for analyzing text complexity, comparison to linear regression can be informative. First, as a parametric technique, data are fit to a linear regression model with the assumption that the data come from probability distributions, and parameters of the variable distributions and relationships can be inferred using the probability distributions. While linear regression may be robust to violations of some assumptions, minimally, homoscedasticity is required (Cohen & Cohen, 1983). As a nonparametric technique, random forest regression makes no assumptions about the underlying distribution of the data or the population. The absence of such assumptions is an advantage when text characteristics are operationalized because such distributions in early-grades texts are not known.

Second, the number of variables and the number of interactions among variables that can be accommodated in linear regression is somewhat limited, whereas random forest regression can handle an extremely large number of variables as well as many interactions including higher order ones. When examining text complexity, a very large number of text characteristics can be imagined, and it seems entirely possible that some text characteristics might interact with others to impact complexity.

Third, linear regression enforces a specific functional form on the relationship between independent variables and dependent variables. The relationship between independent variables is additive and interactions must be explicitly modeled. Random forest makes no assumptions about the functional form of relationships between independent and dependent variables. Arbitrary nonlinear relationships can be implicitly modeled, including nonlinear interactions. The implicit incorporation of interaction effects is one of the most important differentiators between linear regression and random forest regression, one that is significant in a study of text complexity where multiple variable interactions might be possible.

Fourth, linear regression yields statistics and associated probability values revealing the statistical significance of the variable relationships. Random forest regression yields "Importance" values for each variable. Importance for a variable is the amount of increased error in the model when that variable is prevented from influencing the outcome measure (Liaw &

Wiener, 2002; Strobl, Malley, & Tutz, 2009). Variables with higher Importance values often are involved in interactions with other variables. Determining most-important text characteristics while acknowledging potential interactions is a main goal of the present study.

Fifth, a linear regression model involves one statistical run or a small set of runs. Random forest regression involves hundreds or even thousands of iterations of individually trained decision tree models of the relations between predictors and the outcome. Averaging over many models reduces the risk of model overfit, that is training a model to a specific set of data—an advantage in the present study where generalization to other similar early-grades texts is desirable.

Sixth, in linear regression the final model is tested through a statistical "fit" of the model to the data. Model fit in random forest regression is tested through a "validation phase" involving examination of the predictive power of the model using a previously "unseen" data set. Predicting model performance on hold-out data is recommended by some statisticians for all modeling procedures but is rarely employed in linear regression.

References

Cohen, J., & Cohen, P. (1983). *Applied multiple regression/correlation analysis for the behavioral sciences* (2nd ed.). Hillsdale, NJ: Erlbaum.

Liaw, A., & Wiener, M. (2002, December). Classification and regression by random Forest. *R News*, *2*(3), 18–22.

Strobl, C., Malley, J., & Tutz, G. (2009). An introduction to recursive partitioning: Rationale, application, and characteristics of classification and regression trees, bagging, and random forests. *Psychological Methods, 14,* 323–348.