Measuring the Power of Learning.™

Research Report
ETS RR–15-38

Using *TextEvaluator*® to Quantify Sources of Linguistic Complexity in Textbooks Targeted at First-Grade Readers Over the Past Half Century

Kathleen M. Sheehan

Michael Flor

Diane Napolitano

Chaitanya Ramineni

December 2015

ETS Research Report Series

Since its 1947 founding, ETS has conducted and disseminated scientific research to support its products and services, and to advance the measurement and education fields. In keeping with these goals, ETS is committed to making its research freely available to the professional community and to the general public. Published accounts of ETS research, including papers in the ETS Research Report series, undergo a formal peer-review process by ETS staff to ensure that they meet established scientific and professional standards. All such ETS-conducted peer reviews are in addition to any reviews that outside organizations may provide as part of their own publication processes. Peer review notwithstanding, the positions expressed in the ETS Research Report series and other published accounts of ETS research are those of the authors and not necessarily those of the Officers and Trustees of Educational Testing Service.

The Daniel Eignor Editorship is named in honor of Dr. Daniel R. Eignor, who from 2001 until 2011 served the Research and Development division as Editor for the ETS Research Report series. The Eignor Editorship has been created to recognize the pivotal leadership role that Dr. Eignor played in the research publication process at ETS.

ETS Research Report Series ISSN 2330-8516

RESEARCH REPORT

Using *TextEvaluator*® to Quantify Sources of Linguistic Complexity in Textbooks Targeted at First-Grade Readers Over the Past Half Century

Kathleen M. Sheehan, Michael Flor, Diane Napolitano, & Chaitanya Ramineni

Educational Testing Service, Princeton, NJ

This paper considers whether the sources of linguistic complexity presented within texts targeted at 1st-grade readers have increased, decreased, or held steady over the 52-year period from 1962 to 2013. A collection of more than 450 texts is examined. All texts were selected from Grade 1 textbooks published by Scott Foresman during the targeted time period. Analyses are implemented using the *TextEvaluator*® tool, a comprehensive text complexity evaluation tool designed to help teachers, textbook publishers, and test developers identify and quantify text-based sources of comprehension difficulty within informational, literary, and mixed texts. Results suggest that 1st-grade textbooks published over the past half century have included an increasing proportion of informational passages, and this shift has been accompanied by the following specific changes: (a) an increase in the proportion of words that tend to appear less frequently in printed text, (b) an increase in the proportion of words that are more characteristic of academic text as opposed to fiction or conversation, (c) lower levels of referential cohesion, (d) lower levels of narrativity, and (e) fewer instances of an interactive/conversational style. These findings suggest that, in contrast to the claim of a "general, steady decline" in textbook complexity, text-based sources of comprehension difficulty within Grade 1 texts have either risen or held steady throughout the past half century.

Keywords Automated text complexity measurement; readability; Common Core State Standards

doi:10.1002/ets2.12085

The Common Core State Standards (CCSS) accelerate text complexity expectations for students in Grades 1 through 12 in order to address what has been described as a "general, steady decline … in the difficulty … of the texts students have been asked to read in school since 1962" (CCSS Initiative, 2010, Appendix A, p. 3). Research cited in support of this decline has included an analysis of texts targeted at students in Grades 6 and 11 (Chall, Conard, & Harris, 1977)[1] and an analysis of texts targeted at students in Grades 1 – 8 (Hayes, Wolfer, & Wolfe, 1996). In a more recent analysis, however, the claim of a general, steady decline in the complexity levels of textbooks published since 1962 was not supported. In particular, Gamson, Lu, and Eckert (2013) demonstrated that the linguistic complexity of texts targeted at students in Grades 3 and 6 have either risen or held steady over the past half century. These contradictory findings suggest that additional research focused on the complexity levels of textbooks published over the past half century is needed.

This paper presents a detailed analysis of a corpus assembled by E. Hiebert. The corpus includes first-grade textbooks published by Scott Foresman between 1962 and 2013, inclusive. Hiebert (2015) reported that the decision to limit the corpus to textbooks published by Foresman was based on the fact that the Foresman program is the only core reading program that was included in Chall's (1967) seminal analysis of first-grade texts; it has remained in publication throughout the entire targeted time period and has been realigned to address the accelerated text complexity guidelines specified in the CCSS. Thus, it provides an excellent test bed for evaluating variation in the sources of linguistic complexity that many first-grade students are likely to have encountered over the past half century.

All analyses are implemented using the *TextEvaluator*® tool, a comprehensive text complexity evaluation tool that provides text complexity classifications aligned with the CCSS (Sheehan, 2015b). The paper is organized as follows. First, the TextEvaluator tool is described, and evidence collected in a number of previously published validity studies is summarized. Next, analyses implemented with respect to the corpus of texts selected from first-grade textbooks are summarized.

Corresponding author: K. M. Sheehan, E-mail: ksheehan@ets.org
An earlier version of this paper was presented at the American Educational Research Association (AERA), April 17–20, 2015, Chicago, IL.

A final section discusses implications relative to the claim that text complexity levels have "trended downward over the past half century" (CCSS Initiative, 2010, Appendix A, p. 3)

An Overview of the Measurement Approach of the TextEvaluator Tool

Kane (2006) noted, "Measurement uses limited samples of observations to draw general and abstract conclusions about persons or other units" (p. 17). When measuring text complexity, the units that we wish to measure are *texts*, and the general conclusions about texts that we are primarily interested in making concern the levels of knowledge and skill needed to form a coherent mental representation of the information, argument, or story presented within a text; the *observations* on which conclusions are based are determined from a cognitive model of the processes engaged by readers when attempting to make sense of texts with varying combinations of features. The following section outlines the cognitive model adopted for use in selecting observable text features for consideration within the TextEvaluator tool.

A Cognitive Model for Comprehending Complex Text

Much research over the past several years has supported the view of reading as an active process in which readers attempt to build coherent mental representations of the information presented in reading materials (Alderson, 2000; Gernsbacher, 1990; Hiebert & Pearson, 2014; Just & Carpenter, 1987; Kintsch, 1998; Snow, 2002; Valencia, Wixson, & Pearson, 2014). This view suggests that the ability to form a coherent mental representation of a text requires skill at implementing four types of cognitive processes: (a) making sense of the individual words comprising a text, including retrieving definitions from long-term memory, and inferring word meanings from structural components or from surrounding context; (b) using relevant syntactic knowledge to define meaningful propositions, to assemble propositions into sentences, and to infer the meanings of individual sentences; (c) using observable textual clues (e.g., repeated content words, explicit connectives) to fill in gaps and infer connections across sentences and larger sections of text; and (d) using relevant prior knowledge and experience to develop a more complete, more integrated mental representation of a text (i.e., a situation model; Kintsch, 1998). The following section summarizes how this model was used to develop a comprehensive set of observable text features for consideration within the TextEvaluator tool.

Features and Component Scores of the TextEvaluator Tool

The feature set of the TextEvaluator tool currently includes more than 50 features. Each feature was designed to characterize the ease or difficulty of implementing one or another of the four cognitive processes listed above. Because many of the resulting features were expected to be moderately or highly correlated, analyses focused on combining evidence from multiple correlated features were implemented (Sheehan, Flor, & Napolitano, 2013; Sheehan, Kostin, Futagi, & Flor, 2010; Sheehan, Kostin, Napolitano, & Flor, 2014). Approaches for addressing feature intercorrelations are discussed in Biber et al. (2004) and Deane, Sheehan, Sabatini, Futagi, and Kostin (2006). In each case, a two-step solution is proposed. First, corpus-based multidimensional techniques are used to locate clusters of features that simultaneously exhibit high within-cluster correlation and lower between-cluster correlation. Second, linear combinations defined in terms of the identified clusters are employed for text characterization. Biber et al. (2004) justified this approach by noting that because many important aspects of text variation are not well captured by individual linguistic features, investigation of such characteristics requires a focus on "constellations of co-occurring linguistic features" (p. 45) as opposed to individual features. In other words, evidence obtained via multiple correlated features is combined so that the measures passed to the complexity estimation module are more stable, less subject to construct-irrelevant spikes, and more closely focused on targeted aspects of text variation.

The principal components analysis (PCA) conducted with respect to the feature set of the TextEvaluator tool is documented in Sheehan et al. (2010, 2013, 2014).[2] Results suggest that more than 60% of the variation captured by the complete set of more than 50 features could be accounted for via a set of eight component scores, each estimated as a linear combination of multiple correlated features. The eight components incorporated within the current TextEvaluator tool, with illustrative text features, are listed in Table 1. Note that one or more components are provided for each process in the hypothesized cognitive model.

Table 1 Dimensions of Text Variation Addressed by the TextEvaluator Tool, by Targeted Cognitive Process

Targeted cognitive process	TextEvaluator component core	Sample features
Understanding words	Word concreteness	Average concreteness rating, average imageability rating
	Word unfamiliarity	Average ETS word frequency, average TASA word frequency, rare words—type count and token count
	Academic vocabulary	Academic word list ratio, academic words—token count, academic verbs—token count
Understanding sentences	Syntactic complexity	Average sentence length, average word count before main verb, average number of dependent clauses
Inferring connections across sentences	Lexical cohesion	Frequency of overlapping stemmed content words in adjacent sentences
	Argumentation	Causal connectives, adversative connectives, negations
Using knowledge of discourse structure	Narrativity	Pronouns: third person, singular, past tense verbs
	Interactive/conversational style	Pronouns: first person singular, words enclosed in quotes, contractions

Note. ETS = Educational Testing Service. Other features not listed in this table are also included in each component score.

Prediction Equations for the TextEvaluator Tool

In many educational contexts, inferences about students' knowledge, skills, and accomplishments are based on evidence extracted from observed item responses. Although each new item is designed to provide unbiased evidence about examinee standing relative to the targeted proficiency construct, item pretest statistics are routinely evaluated in order to ensure that the items accepted for use on operational assessments do not incorporate unintended biases. In some cases, analyses of item pretest statistics confirm that one or more items exhibit differential item functioning (DIF). DIF occurs when test takers with similar levels of a measured attribute or trait tend to score lower or higher on an item depending on the particular subpopulation to which they belong (Holland & Wainer, 1993). When items with significant levels of DIF are included on an assessment, subsequent mean scores may indicate an achievement gap among test takers in some subpopulations (e.g., male examinees vs. female examinees) even when no gap is actually present.

Just as the evidence provided by a proposed test item may be biased in favor of examinees in some subgroups (e.g., male examinees), the evidence provided by a proposed text complexity feature may be biased in favor of texts in some subgroups (e.g., informational texts). Sheehan (2015c) referred to this phenomenon as differential feature functioning (DFF) and highlighted several published instances of its occurrence. For example, the authors of the CCSS argued as follows: "The Lexile Framework, like traditional formulas, may underestimate the difficulty of texts that use simple, familiar language to convey sophisticated ideas, as is true of much high-quality fiction written for adults and appropriate for older students" (CCSS Initiative, 2010, Appendix A, p. 7). Similarly, Sheehan et al. (2013) noted that many traditional readability metrics tend to underestimate the complexity levels of literary texts while simultaneously overestimating the complexity levels of informational texts. The TextEvaluator tool addresses these effects by providing three distinct prediction models: 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). In each case, equations are designed to predict human grade level (GL) judgments conditional on the component scores obtained in the PCA. The regression coefficients estimated for informational and literary texts are listed in Table 2. Note that each component accounts for a significant amount of GL variation in one or both models.

Validity Evidence of the Text Evaluator Tool

The problem of validating claims made on the basis of scores produced by automated text analysis tools shares a number of similarities with a more familiar psychometric problem: validating claims made on the basis of test scores. The validity as argument framework is frequently employed when evaluating such claims (Cronbach, 1988; Kane, 1990, 2006, 2013). Kane (2013) described this framework as encompassing two steps. First, the specific claims to be validated are elaborated. This includes specifying the networks of inferences and supporting assumptions that underlie each claim. Second, evidence that either supports or refutes each element in the specified networks is examined. If all of the required inferences and

Table 2 Coefficients Obtained in Genre-Specific Regression Analyses Designed to Predict Human Grade Level Judgments, by Targeted Cognitive Process

Targeted cognitive process	TextEvaluator component score	Informational texts	Literary texts
Understanding words	Word unfamiliarity	0.802*	0.793*
	Word concreteness	−0.610*	−0.483*
	Academic vocabulary	1.126*	0.824*
Understanding sentences	Syntactic complexity	0.983*	1.404*
Inferring connections across sentences	Lexical cohesion	−0.266*	−0.440*
	Argumentation	0.431*	*ns*
Using knowledge of discourse structure	Degree of narrativity	*ns*	−0.361*
	Interactive/conversational style	−0.518*	*ns*

Note. ns = not significant. The informational model was estimated from the set of all informational passages in the training dataset ($n = 399$) and yielded a human/automated correlation of .86. The literary model was estimated from the set of all literary passages in the training dataset ($n = 452$) and yielded a human/automated correlation of .81.
*$p < .01$.

supporting assumptions are found to be highly plausible (either a priori or because of the evidence provided), the claim associated with the specified network would be considered plausible or valid. If any part of the argument is not plausible, however, the specified claim would be deemed invalid.

Since resulting networks of inferences and supporting assumptions can become quite large quite quickly, Kane (1990) suggested focusing on those links that appear to be most open to challenge (i.e., links that appear to be "doubtful or problematic" [p. 20]). This recommendation is based on the notion that "a serious weakness in any core inference tends to undermine the argument as a whole, even if other inferences are strongly supported" (Kane, 1990, p. 13). In other words, a proposed argument chain is only as strong as its weakest link.

Consistent with the approach outlined above, this section summarizes evidence collected to support three key claims in the validity argument for the TextEvaluator tool. For a more complete description of the available evidence, please see Sheehan (2015a).

Claim #1: The strategy of the TextEvaluator tool of estimating distinct prediction models for informational, literary, and mixed texts has succeeded in yielding text complexity scores that are less subject to genre bias. Evidence that supports this claim is summarized in Figure 1. Four plots are shown: two that present validity evidence focused on the functioning of the TextEvaluator tool and two that present validity evidence focused on the functioning of the Lexile tool (Stenner, Burdick, Sanford, & Burdick, 2007). In each plot, text complexity classifications generated via one or another of these two tools are compared to corresponding GL classifications provided by human experts. Further, within each row, results for informational texts are summarized on the left ($n = 243$), and those for literary texts are summarized on the right ($n = 305$). The plots show that while complexity scores generated via the Lexile tool tend to be too high for many informational texts and too low for many literary texts, complexity scores generated via the TextEvaluator tool exhibit minimal, if any, genre bias.

Claim #2: Overall text complexity scores generated via the TextEvaluator tool are highly correlated with text complexity judgments provided by human experts. Evidence related to this claim is summarized in Table 3. The table summarizes the degree of correlation between overall text complexity scores generated via the TextEvaluator tool and text complexity judgments provided by human experts. Three types of passages are included in the summary: passages from the TextEvaluator training corpus ($n = 941$) and passages from two additional corpora that were not considered during development and training for the TextEvaluator tool. The two independent corpora include the set of exemplar passages provided in Chall, Bissex, Conrad, & Harris-Sharples (1996; $n = 52$) and the set of exemplar passages provided in Appendix B of the CCSS ($n = 168$). These two corpora are alike in some ways and different in others. A key similarity is that each was intentionally assembled to illustrate the increases in text complexity that students can expect to experience in their journey from beginning reader to proficient college-ready reader. A key difference is that while the human-assigned ratings in the Chall corpus were collected more than 20 years ago, those in the Common Core corpus were collected just 5 years ago.

The estimates in Table 3 confirm that scores from the TextEvaluator tool are highly correlated with judgments provided by human experts, both when compared to the earlier human ratings provided by Chall and her colleagues and when compared to the more recent ratings provided by the experts who classified the Common Core exemplar texts. The fact that

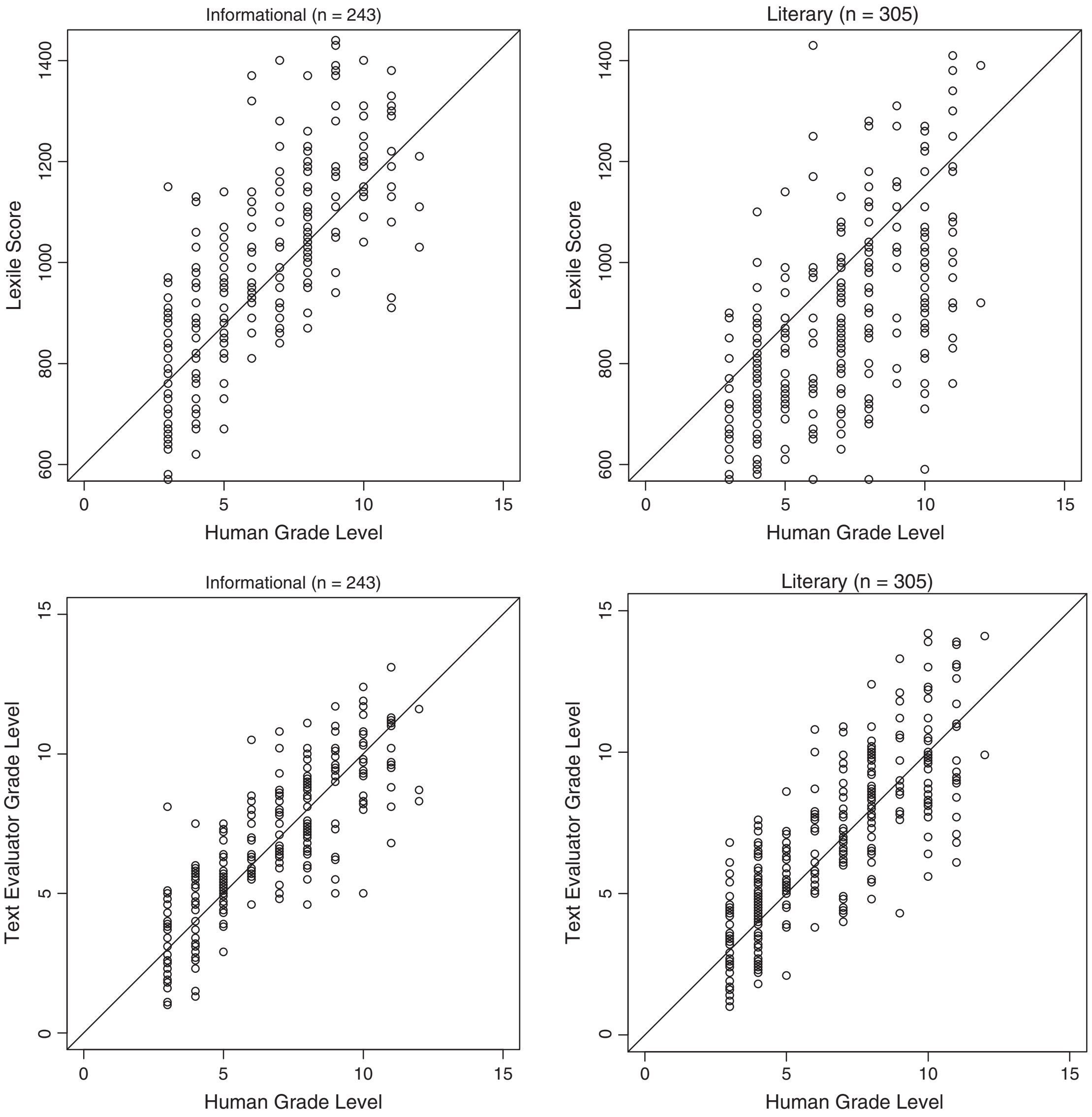


Figure 1 Overall text complexity scores generated via the Lexile Reading Analyzer (top) and the TextEvaluator tool (bottom) compared to grade level classifications provided by experienced educators, for informational texts (left) and literary texts (right). All texts were selected from high-stakes, standards-based reading assessments.

high correlations were obtained when evaluating ratings provided at both earlier and later time points is relevant because a key goal of the current study involves comparing the complexity levels of texts published at earlier and later time points.

Table 3 also shows lower correlations for the Common Core passages compared to the other two sets of passages. When considering this result, it is important to note that the reading experts who rated the Common Core passages employed a 5-point grade band scale, whereas the experts who rated both the training passages for the TextEvaluator tool and the Chall passages employed GL scales that each included at least 10 points. The lower precision afforded by the use of a 5-point grade band scale, as opposed to a 10-point grade level scale, may account for a portion of the decreased correlation observed for the Common Core passages. Across all three corpora, however, the results in Table 3 support the claim that overall text complexity scores generated via the TextEvaluator tool are highly correlated with complexity judgments provided by human experts.

Claim #3: Each of the eight component scores of the TextEvaluator tool provides valid feedback about the levels of linguistic complexity detected within informational and literary texts. Evidence that supports this claim is presented in Tables 2–4. Table 2 confirms that each of the eight component scores of the TextEvaluator tool contributes significantly to the overall text complexity score of the TextEvaluator tool, whereas Table 3 confirms that the resulting overall complexity

Table 3 Correlation Between Overall Text Complexity Scores from the TextEvaluator Tool and Complexity Judgments Provided by Human Experts, by Type of Corpus and Genre

Type of corpus/genre	Number of texts	Total words	Correlation
TextEvaluator training corpus[a]			
Informational	399	226,823	.86
Literary	459	323,194	.80
Mixed	83	57,520	.82
Total	941	607,537	.83
Validation corpus #1 (Common Core exemplars)			
Informational	54	19,872	.79
Literary	56	25,255	.58
Mixed	58	34,951	.77
Total	168	80,078	.73
Validation corpus #2 (Chall exemplars)			
Informational	28	4,438	.93
Literary	24	3,755	.90
Total	52	8,193	.91

Note. All correlations are significant at $p < .001$.
[a] The TextEvaluator training corpus includes passages from standardized state reading assessments targeted at students in Grades 2–12, passages from the National Assessment of Educational Progress (NAEP) reading assessments targeted at students in Grades 4, 8 and 12, and passages from two different college admissions assessments: the SAT and the ACT.

Table 4 Correlations Between the Component Scores from the TextEvaluator Tool and Grade Level Judgments Provided by Human Experts, by Targeted Cognitive Process and Genre

		Correlation with human grade level judgments	
Targeted cognitive process	TextEvaluator component score	Informational	Literary
Understand words	Unfamiliar vocabulary	.69*	.62*
	Academic vocabulary	.78*	.64*
	Word concreteness	−.74*	−.60*
Understand sentences	Syntactic complexity	.75*	.69*
Infer connections	Referential cohesion	−.36*	−.31*
	Argumentation	.32*	.19
Use discourse knowledge	Interactive/conversational style	−.29*	−.36*
	Degree of narrativity	.19	−.33*

Note. Estimated from 399 informational passages and 459 literary passages selected from the TextEvaluator training corpus. Detailed descriptions of each score are provided in Sheehan et al. (2014).
*$p < .001$.

scores are closely aligned with text complexity judgments provided by human experts, including both the experts who rated the passages in the Chall corpus, and the experts who rated the Common Core exemplars. Table 4 further bolsters this claim by demonstrating that several of the component scores of the TextEvaluator tool are also significantly correlated with human text complexity judgments when considered individually. These three sets of results support the claim that each of the eight component scores of the TextEvaluator tool provides valid feedback about the levels of text complexity that students can expect to encounter in their school and home-based reading.

The validity analyses summarized above suggest that scores generated via the TextEvaluator tool may help users understand similarities and differences in the levels of linguistic challenge presented within first-grade textbooks published over the past half century.

Method

Valencia et al. (2014) noted that characteristics of the text represent only one of several factors that may influence comprehension. Other factors include characteristics of the reader, the task, and the particular context in which comprehension

Table 5 Numbers of Passages in the Scott Foresman Grade 1 Corpus, Before and After Removing Outlying Passages, by Year

Year	No. of passages	Median length (in words)	No. with length ≥300 words	No. of outliers	Passages remaining
1962	114	163	16	1	113
1983	107	271	45	2	105
1993	64	86	7	4	60
2000	66	170	15	5	61
2007	59	126	13	3	56
2013	54	121	13	2	52
Total	464	179	109	17	447

Note. Passages from 1974 are not included in this summary because these passages were not available at the time that TextEvaluator scores were generated.

is to take place. They propose the construct of a text-task scenario as a way to illuminate the interplay between texts and tasks when determining comprehension success or failure. In this study, however, we focus only on the text and consider whether the complexity levels of texts targeted at first-grade readers have increased, decreased, or remained the same over the 52-year period from 1962 to 2013.

A database of 464 texts was assembled for consideration in the analyses. All texts were selected from Grade 1 textbooks published by Scott Foresman during the targeted time period. Table 5 shows the numbers of passages available from textbooks published in each of the following years: 1962, 1983, 1993, 2000, 2007, and 2013.[3] In addition to providing the numbers of passages available at each of six different publication years, Table 5 also lists the median length (in words) of the passages included in each sample. Note that, except for the 1983 publication year, most samples yielded median text lengths that were quite a bit shorter than 300 words.

Analyses were implemented as follows. First, the TextEvaluator tool was used to generate a feature vector and a corresponding set of component scores for each text. Next, individual outputs were examined, and 17 texts with outlying feature values were identified. Although the TextEvaluator tool was originally designed for use with texts containing at least 300 words, the outlier analysis was implemented because many of the texts in the current corpus fall quite a bit short of that threshold. The outlier analysis is documented in Appendix.

At the completion of the outlier analysis, a total of 17 texts were classified as exhibiting outlying values on one or more features and were excluded from all subsequent analyses. Findings based on the remaining set of 547 texts (564 minus 17) are summarized in the next section.

Results

Each of the eight component scores of the TextEvaluator tool is expressed on a standardized scale that ranges from 1 to 100, where 1 is slightly lower than the lowest score observed among the 941 passages in the tool's training corpus and 100 is slightly larger than the largest score observed in the training corpus. When interpreting these scores, it is important to remember that higher scores may indicate more or less difficulty, depending on the case. For example, because syntactic complexity tends to increase with GL, the syntactic complexity component is structured such that higher scores indicate more difficulty (i.e., more complex sentences) and lower scores indicate less difficulty (i.e., less complex sentences). By contrast, because the level of concreteness in a text tends to decrease with GL, the concreteness component is structured such that higher scores indicate less difficulty (i.e., more concrete words) and lower scores indicate more difficulty (i.e., fewer concrete words). The particular scaling option selected for each component is indicated in Table 6. The table presents the median scores observed, by year, for each of nine components. This set includes the eight components discussed in previous research about the TextEvaluator tool plus an additional component that has just been added to the research version of the tool. To facilitate interpretation, results for individual components are grouped according to the type of cognitive process targeted (i.e., understanding words, understanding sentences, inferring connections across sentences, or using knowledge of discourse structure). The median scores observed across the 52-year period from 1962 to 2013 are evaluated in Table 7. To better highlight key trends, Table 7 shows how each median compares to the medians estimated from the 1962 sample. Increases (or decreases) that represent at least one tenth of the total scale (about 10 points) are meaningful because a large proportion of the passages in the training sample for the TextEvaluator tool were targeted at students in Grades 3 through 12, a span of 10 GLs.

Table 6 Median Scores for Each of Nine Components of the TextEvaluator Tool, by Targeted Cognitive Process and Year

	Understanding words			Und. sent.	Inferring connections across sentences			Using discourse knowledge	
Year	WU	AV	Con.	SC	LC	LT	Arg	Nar	IC
1962	13	4	76	16	66	82	38	82	76
1983	21	9	74	20	60	81	39	62	74
1993	40	16	78	27	48	81	37	58	52
2000	34	11	78	13	51	81	34	66	70
2007	41	14	77	19	55	79	20	62	58
2013	43	14	79	20	51	79	26	65	57

Note. Und. Sent. = understanding sentences; WU = word unfamiliarity (+); LC = lexical cohesion (−); AV = academic vocabulary (+); LT = lexical tightness (−); Con = concreteness (−); Arg = argumentation (+); SC = syntactic complexity (+); Nar = narrativity (−); IC = interactive/conversational style (−).In the abbreviation explanations, a (+) after the name signals that higher values indicate higher complexity; (−) signals that higher values indicate lower complexity.

Table 7 Changes in Component Scores of the TextEvaluator Tool Relative to the 1962 Median, by Targeted Cognitive Process and Year

	Understanding words			Und. sent.	Inferring connections across sentences			Using discourse knowledge	
Year	WU	AV	Con.	SC	LC	LT	Arg	Nar	IC
1962	—	—	—	—	—	—	—	—	—
1983	+8	+5	+2	+4	+6	+1	+1	+20	+2
1993	+27	+12	−2	+11	+18	+1	−1	+24	+24
2000	+21	+7	−2	−3	+15	+1	−4	+16	+6
2007	+28	+10	−1	+3	+11	+3	−18	+20	+18
2013	+30	+10	−3	+4	+15	+3	−12	+17	+19

Note. Und. Sent. = understanding sentences; WU = word unfamiliarity (+); LC = lexical cohesion (−); AV = academic vocabulary (+); LT = lexical tightness (−); Con = concreteness (−); Arg = argumentation (+); SC = syntactic complexity (+); Nar = narrativity (−); IC = interactive/conversational style (−). Directional signs for some changes have been reversed so that a (+) always indicates an increase in complexity, and a (−) always indicates a decrease in complexity.

Process #1: Understanding Words

The TextEvaluator tool includes three components that characterize the level and type of word knowledge needed to successfully comprehend a text: the word unfamiliarity score, the academic vocabulary score, and the concreteness score. These three components characterize three different aspects of required word knowledge: the knowledge and skill needed to understand words that tend to appear more or less frequently in printed text, the knowledge and skill needed to understand words that tend to appear more or less frequently in academic text, and the knowledge and skill needed to understand words that are more or less concrete. The first two components (word unfamiliarity and academic vocabulary) are scaled such that higher scores indicate more difficulty; the concreteness score is scaled such that higher scores indicate less difficulty.

Note in Table 7 that the changes for the concreteness score are reported with an opposite sign. This strategy was adopted so that a positive change is always indicative of an increase in comprehension difficulty and a negative change is always indicative of a decrease in comprehension difficulty. Results suggest that the level of word knowledge needed to comprehend Grade 1 texts published between 1962 and 2013 has trended upward, not downward as is suggested in Appendix of the CCSS. For example, the word unfamiliarity score increased from a median of 13 in 1962 to a median of 43 in 2013, an increase of 30 points. The median scores listed for the academic vocabulary component also suggest an increasing trend rather than a decreasing trend. In particular, the academic vocabulary score increased from a median of 4 in 1962 to a median of 14 in 2013, an increase of 10 points. By contrast, changes in the median concreteness scores estimated across the 52-year span from 1962 to 2013 are all less than 10 points, indicating that the concreteness levels of the texts published during this time span were all quite similar.

To better illustrate these trends, Figure 2 displays the individual values obtained for two of the eight features that are included in the word unfamiliarity component: (a) the average frequency of words determined via the Educational Testing Service (ETS) word frequency index (Sheehan et al., 2013) and (b) the average frequency of words determined via the TASA word frequency index (Touchstone Applied Science Associates; Zeno, Ivens, Millard, & Duvvuri, 1995). Both indices suggest that passages published in later years tend to include more words with lower frequency values, compared to passages published in earlier years. Note that, while this trend is visible in each display, it is more pronounced in the ETS word frequency display. The larger size of the ETS corpus compared to the TASA corpus may account for this difference. That is, the ETS word frequency index may provide more precise evidence about word familiarity because it was estimated from a corpus that included more than 400 million word tokens, whereas the TASA index was estimated from a corpus that included just slightly more than 17 million word tokens (Zeno et al., 1995).

When considering the trends suggested in Figure 2, it is important to remember that although word familiarity may vary across years, many word frequency indices, including the two indices incorporated within the TextEvaluator tool, assume a single, invariant frequency value for each word. Thus, anomalies such as words that are more frequent in later years and less frequent in earlier years are not addressed.

When interpreting the changes indicated in Table 6 and Figure 2, it is also important to remember that informational passages tend to include a higher proportion of low frequency words, and a higher proportion of academic words, compared to literary passages at the same GL (Sheehan et al., 2014). Human genre classifications are not available for the texts analyzed in this study; however, classifications obtained via the automated genre classifier (Sheehan et al., 2013) of the TextEvaluator tool suggest that the 1962 sample is primarily composed of literary passages, whereas each subsequent sample includes an increasing proportion of informational passages. Thus, the increasing proportion of informational texts included in each successive sample may account for the increasing proportions of unfamiliar words and academic words shown in Tables 6 and 7.

Process #2: Understanding Sentences

The TextEvaluator tool includes a single component focused on the ease or difficulty of understanding the individual sentences presented within a text: the syntactic complexity score. Median syntactic complexity scores are summarized in Tables 6 and 7. Results suggest that the process of assembling words into sentences and then inferring the meaning of those sentences has remained focused at approximately the same difficulty level in Grade 1 texts published between 1962 and 2013. For example, Table 6 shows that median syntactic complexity increased from 16 in 1962 to 20 in 2013, an increase of just four points. To better illustrate this finding, Figure 3 displays the individual values obtained for two of the eight features included in the syntactic complexity component: (a) the average frequency of dependent clauses and (b) the average number of words between punctuation marks. Although both measures show slight increases, the magnitude of the increase was not sufficient to exceed the specified threshold of at least 10 scale points.

Process #3: Inferring Connections Across Sentences

In addition to inferring the meaning of individual sentences, competent readers must also be proficient at inferring connections across sentences. The TextEvaluator tool currently includes three measures focused on this particular aspect of text variation: the lexical cohesion score, the lexical tightness score, and the level of argumentation score. The lexical cohesion score measures the degree of lexical overlap across successive sentences. The lexical tightness score measures global cohesion (i.e., the degree to which a text employs words that tend to co-occur in printed text; see Flor, Beigman Klebanov, & Sheehan, 2013). The level of argumentation score measures whether comprehension of successive sentences requires processing of argumentative conjuncts (e.g., although, on the other hand) and/or negations. Previous research has confirmed that a high score on either the lexical cohesion component or the lexical tightness component contributes to processing ease, whereas a high score on the argumentation component contributes to processing difficulty (Flor et al., 2013; Sheehan et al., 2013).

The median scores in Table 6 show that levels of lexical cohesion have tended to decrease over the years. In particular, texts in later years exhibit less sentence-to sentence overlap compared to texts in earlier years. This finding suggests that the process of inferring connections across sentences may be slightly more difficult in later years compared to earlier years.

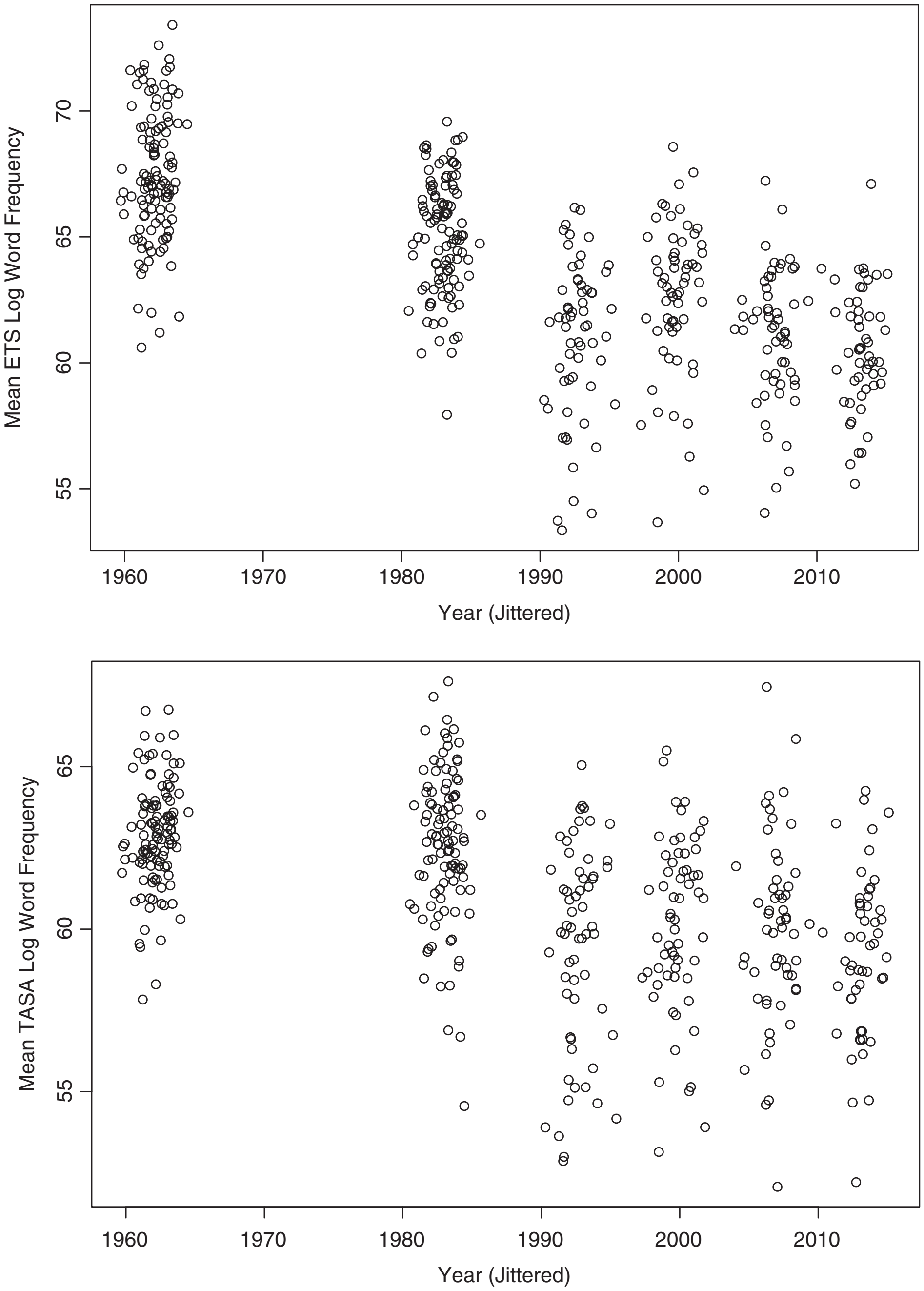


Figure 2 Scatter plots of Educational Testing Service word frequency versus year of publication (top) and TASA word frequency versus year of publication (bottom). Each plot includes 447 points, one for each of 447 texts. For both measures, texts with lower scores are expected to more difficult.

Table 6 also presents results for the lexical tightness score, a measure of the degree to which passages are comprised of words that are more or less highly interassociated. Results suggest that interassociations among words have also trended downward but only by a small amount.

A somewhat different pattern is present in the median scores estimated for the level of argumentation component. These scores suggest that levels of argumentation were fairly similar in the period from 1962 to 2000 but then decreased in 2007 and 2013. The lower argumentation scores observed in 2007 and 2013 suggest that passages in these samples tended to include fewer argumentative conjuncts, indicating that less argumentative reasoning may have been needed to infer connections across sentences. Additional research is needed to more fully understand these results.

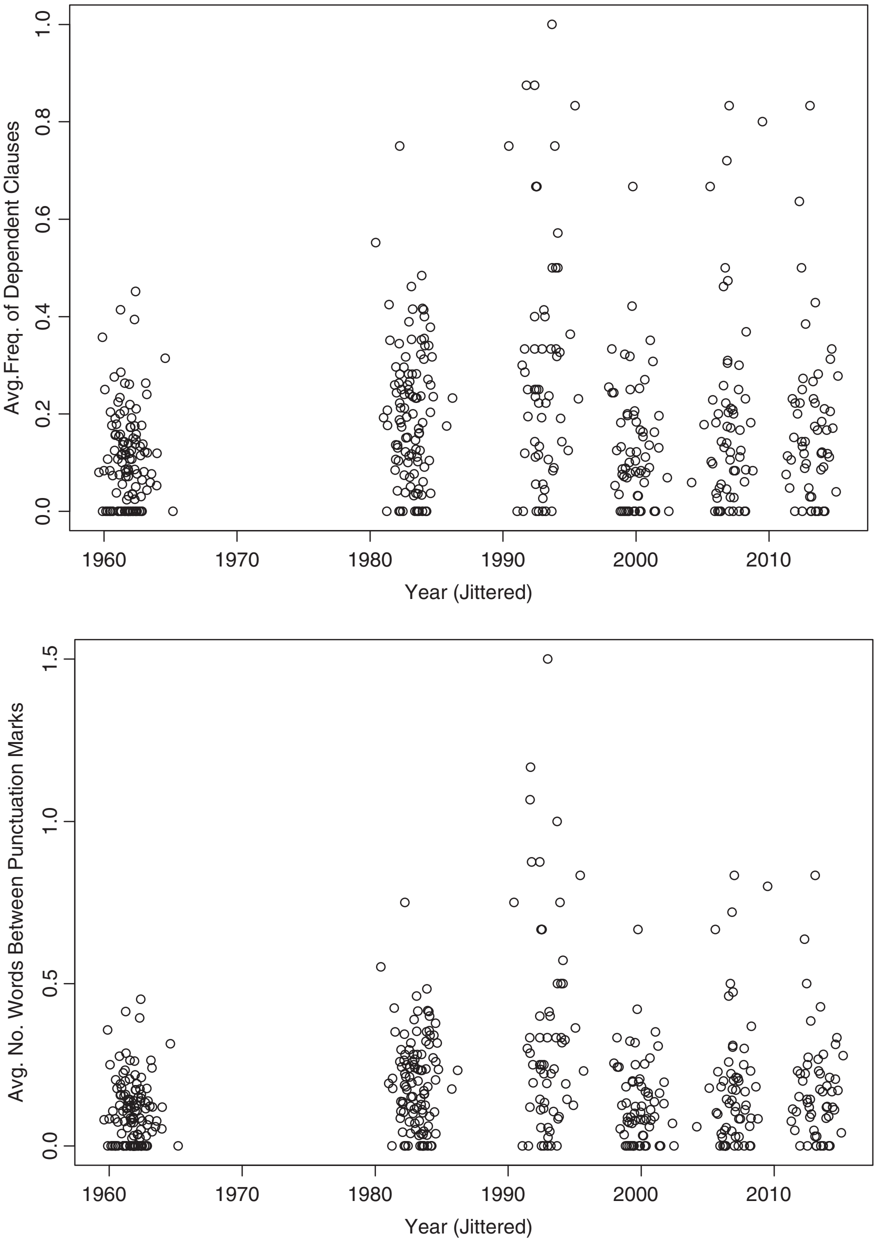


Figure 3 Scatter plots of average frequency of dependent clauses versus year of publication (top) and average number of words between punctuation marks versus year of publication (bottom). Each plot includes 447 points, one for each of 447 texts. For both measures, texts with higher scores are expected to more difficult.

Process #4: Using Prior Knowledge About Rhetorical Patterns

Texts that employ more familiar rhetorical patterns may be less difficult to comprehend compared to texts that employ less familiar rhetorical patterns. The TextEvaluator tool includes two components focused on this aspect of text variation: the degree of narrativity score and the interactive/conversational style score. In each case, higher scores indicate more familiar discourse patterns and thus are associated with slight decreases in comprehension difficulty. Median scores estimated for these components are summarized in Table 6. The table shows that both the narrativity score and the interactive/conversational style score showed meaningful decreases, although the narrativity decrease was confined to the 21-year period from 1962 to 1983. Because informational texts tend to be less narrative and less conversational, these decreases provide additional evidence to support the claim that first-grade textbooks published between 1962 and 2013 have included an increasing proportion of informational text.

Conclusions, Limitations, and Directions for Additional Research

This study considered differences in the processing difficulties of texts selected from first-grade textbooks published between 1962 and 2013. Analyses were implemented using TextEvaluator, a comprehensive text complexity evaluation tool developed at ETS. Nine dimensions of text variation were examined, including dimensions focused on each of the following cognitive processes: (a) inferring the meanings of the individual words comprising a text, (b) assembling words into sentences and then inferring the meanings of resulting sentences, (c) inferring connections across sentences and larger sections of text, and (d) using knowledge of discourse structure to generate a more complete, more integrated mental representation of the information, argument or story presented within a text.

Although the authors of the CCSS have argued that texts published over this time period have exhibited a “general, steady decline” (CCSS Initiative, 2010, Appendix A, p. 3) in complexity, no evidence of that decline was observed. Instead, results suggested that first-grade textbooks published over the studied time period have included an increasing proportion of informational texts, and this shift has been accompanied by the following specific changes: (a) an increasing proportion of words that tend to appear less frequently in printed text, (b) an increasing proportion of words that are more characteristic of academic text as opposed to fiction or conversation, (c) lower levels of referential cohesion, (d) lower levels of narrativity, and (e) fewer instances of an interactive/conversational style. The analyses also indicated that levels of syntactic complexity increased slightly and levels of global cohesion decreased slightly. Note that, in each case, previous research has confirmed that the direction of these changes is toward more difficulty not less difficulty (Sheehan et al., 2010, 2013, 2014). Thus, the findings are not consistent with the claim of a general, steady decline in textbook complexity. Rather, the findings suggest that the comprehension challenges presented within Grade 1 texts have either risen or held steady over the past half century.

Certain limitations of the above analyses should be noted. First, many of the current features of the TextEvaluator tool were originally designed for application to passages containing at least 300 words. Because many of the passages in the current corpus are shorter than 300 words, additional analyses of those passages could help to further clarify the observed changes. Second, because all of the analyses are focused on texts published by Scott Foresman, additional analyses are needed to determine whether similar effects are observed when texts from other publishers are examined. Third, the analyses did not include any passages from the 1974 publication year because they were not received in time to be included in the study. A fourth limitation concerns the use of word frequency indices that are not structured to capture changes in word familiarity over time. Despite these limitations, however, the current results are informative in that they provide no support for the claim that complexity levels of texts targeted at first-grade readers have trended downward over the past half century. Rather, the data are more consistent with the conclusion reported in Gamson et al. (2013) that textbook complexity has either held steady or risen over the past half century.

Notes

[1] Although analyses of first-grade texts are also discussed in Chall et al. (1977), the texts considered in those analyses were published at an earlier time period (i.e., between 1930 and 1962).

[2] An earlier description of the PCA for the TextEvaluator tool appears in the US patent full text database located at http://patft.uspto.gov. As is noted in that document, the analysis was originally described in a patent application submitted on January 31, 2008. A patent covering the analysis was subsequently awarded on September 27, 2013.

[3] This summary does not include any passages from 1974 because these were received too late to be included in the analyses.

References

Alderson, J. C. (2000). *Assessing reading*. Cambridge, UK: Cambridge University Press.

Biber, D., Conrad, S., Reppen, R., Byrd, P., Helt, M., Clark, V., . . . Urzua, A. (2004). *Representing language use in the university: Analysis of the TOEFL 2000 spoken and written academic language corpus* (TOEFL Monograph Series, MS-25). Princeton, NJ: Educational Testing Service.

Chall, J. S. (1967). *Learning to read: The great debate*. New York, NY: McGraw-Hill.

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, J. S., Conard, S., & Harris, S. (1977). *An analysis of textbooks in relation to declining SAT scores*. New York, NY: Advisory Panel on the Scholastic Aptitude Test Score Decline.

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 and technical subjects*. Washington, DC: CCSSO & National Governors Association.

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

Cronbach, L. J. (1988). Five perspectives on validity argument. In H. Wainer & H. Braun (Eds.), *Test validity* (pp. 3 – 17). Hillsdale, NJ: Lawrence Erlbaum.

Deane, P., Sheehan, K. M., Sabatini, J. P., Futagi, Y., & Kostin, I. (2006). Differences in text structure and its implications for the assessment of struggling readers. *Scientific Studies of Reading*, *10*(3), 257 – 275.

Flor, M., Beigman Klebanov, B., & Sheehan, K. M. (2013). Lexical tightness and text complexity. In L. Rello, H. Saggion, & R. Baeza-Yates (Eds.), *Proceedings of the 2nd Workshop on Natural Language Processing for Improving Textual Accessibility* (pp. 29 – 38). Stroudsburg, PA: Association for Computational Linguistics.

Gamson, D. A., Lu, X., & Eckert, S. A. (2013). Challenging the research base of the Common Core State Standards: A historical reanalysis of text complexity. *Educational Researcher*, *42*(7), 381 – 391.

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

Hayes, D. P., Wolfer, L. T., & Wolfe, M. F. (1996). Schoolbook simplification and its relation to the decline in SAT-Verbal scores. *American Educational Research Journal*, *33*, 489 – 509.

Hiebert, E. H. (2015, April). *Multiple perspectives on the complexity of first grade text over a 50-year period*. Paper presented at the 96th annual meeting of the American Educational Research Association, Chicago, IL.

Hiebert, E. H., & Pearson, P. D. (2014). Understanding text complexity: Introduction to the special issue. *Elementary School Journal*, *115*(2), 153 – 160.

Holland, P. W., & Wainer, H. (1993). *Differential item functioning*. Hillsdale, NJ: Lawrence Erlbaum.

Just, M. A., & Carpenter, P. A. (1987). *The psychology of reading and language comprehension*. Boston, MA: Allyn & Bacon.

Kane, M. T. (1990). *An argument-based approach to validation* (Research Report No. 90-13). Iowa City, IA: ACT.

Kane, M. T. (2006). Validation. In R. L. Brennan (Ed.), *Educational Measurement* (4th ed., pp. 17 – 64). Westport, CT: Praeger.

Kane, M. T. (2013). Validating the interpretations and uses of test scores. *Journal of Educational Measurement*, *50*, 1 – 73.

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

Sheehan, K. M. (2015a). *A review of evidence presented in support of three key claims in the TextEvaluator validity argument*. Manuscript submitted for publication.

Sheehan, K. M. (2015b). *Aligning TextEvaluator scores with the accelerated text complexity guidelines specified in the Common Core State Standards* (Research Report No. RR-15-21). Princeton, NJ: Educational Testing Service. http://dx.doi.org/10.1002/ets2.12068

Sheehan, K. M. (2015c, April). *What proportion of the high school/college text complexity gap is due to genre-based Differential Feature Functioning (DFF)?* Invited distinguished paper presentation at the 96th annual meeting of the American Educational Research Association Chicago, IL.

Sheehan, K. M., Flor, M., & Napolitano, D. (2013). A two-stage approach for generating unbiased estimates of text complexity. In L. Rello, H. Saggion, & R. Baeza-Yates (Eds.), *Proceedings of the 2nd Workshop on Natural Language Processing for Improving Textual Accessibility*, 49 – 58. Stroudsburg, PA: Association for Computational Linguistics.

Sheehan, K. M., Kostin, I., Futagi, Y., & Flor, M. (2010). *Generating automated text complexity classifications that are aligned with targeted text complexity standards* (Research Report No. RR-10-28). Princeton, NJ: Educational Testing Service. http://dx.doi.org/10.1002/j.2333-8504.2010.tb02235.x

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.

Snow, C. (2002). *Reading for understanding: Toward an R&D program in reading comprehension*. Santa Monica, CA: RAND.

Stenner, A. J., Burdick, H., Sanford, E. E., & Burdick, D. S. (2007). *The Lexile Framework for Reading technical report*. Durham, NC: MetaMetrics.

Vajjala, S., & Meurers, D. (2012). On improving the accuracy of readability classifications using insights from second language acquisition. In *Proceedings of the 7th Workshop on the Innovative Use of NLP for Building Educational Applications*, 163 – 173. Stroudsburg, PA: Association for Computational Linguistics.

Valencia, S. W., Wixson, K. K., & Pearson, P. D. (2014). Putting text complexity in context: Refocusing on comprehension of complex text. *Elementary School Journal*, *115*(2), 270 – 289.

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

Appendix

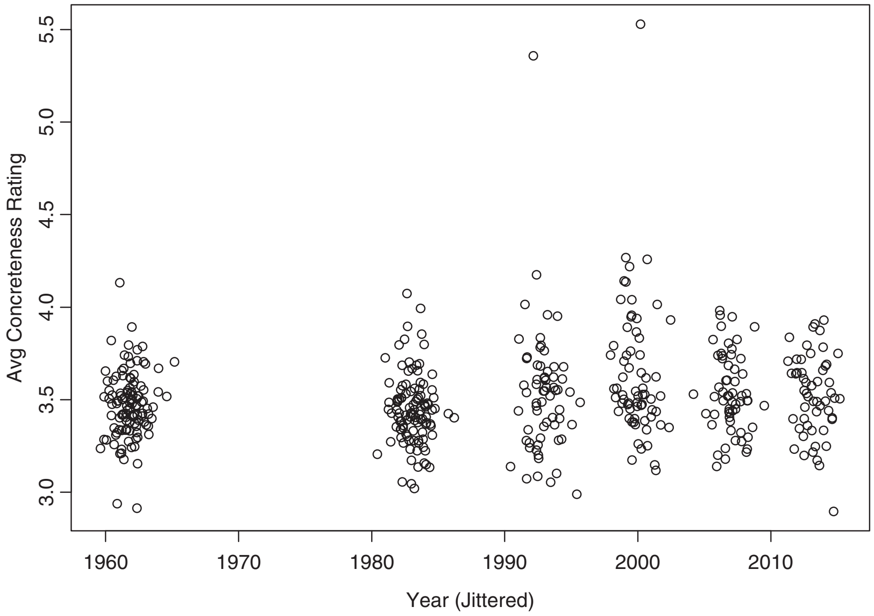
Outlier Analysis

This appendix illustrates the analyses conducted to identify texts with outlying feature values. Analyses focused on four illustrative features are included: average concreteness rating, ratio of academic words to total words, average sentence length, and mean lexical tightness.

The Average Concreteness Rating

The average concreteness rating is one of five features included in the concreteness component of the TextEvaluator tool. It is estimated from the database of word concreteness ratings provided in Coltheart (1981). The database provides human-generated concreteness ratings expressed on a 7-point scale that ranges from 1 (*least concrete*) to 7 (*most concrete*). An average score is generated for each text by averaging over the individual concreteness ratings obtained for each individual word.

Figure A1 displays the average scores generated for each text. A total of 464 points are plotted, one for each of the 464 texts included in the original corpus. Note that unusually high averages were obtained for two texts: one from the 1993 sample and one from the 2000 sample. The individual texts that yielded these unusually high averages are also shown. As can be seen, the individual words comprising each text are all very concrete. In order to ensure that these outlying scores would not exert an undue influence on the inferences of interest, these two texts were classified as outliers and excluded from all subsequent analyses.

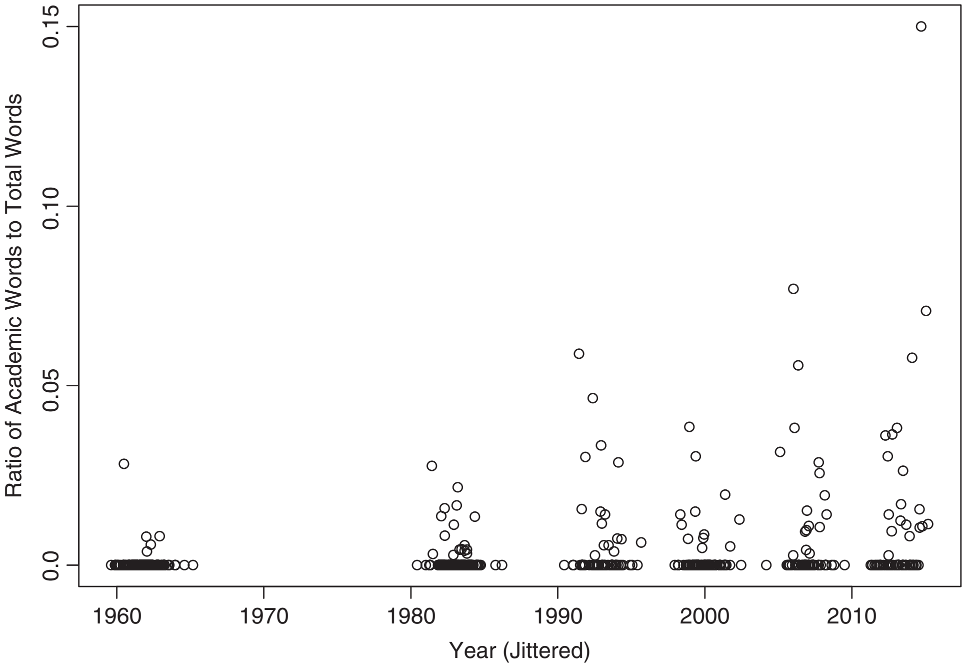


Year =1993	Year = 2000
Frogs jump. Caterpillars hump. Worms wiggle. Bugs jiggle. Rabbits hop. Horses clop. Snakes slide. Sea gulls glide. Mice creep. Deer leap. Puppies bounce. Kittens pounce. Lions stalk. But I walk!	Trucks carry logs. Trucks haul garbage. Trucks hold milk. Trucks mix cement. Trucks dump rocks. Trucks plow snow. Trucks tow cars. Trucks bring mail.

Figure A1 Average concreteness ratings plotted versus year (top), and two passages with outlying concreteness scores (bottom).

The Ratio of Academic Words to Total Words

The ratio of academic words to total words is one of 10 features included in the academic vocabulary component of the TextEvaluator tool. It is estimated from the database of academic words provided in Coxhead (2000). Vajjala and Meurers (2012) demonstrated that this feature was effective at distinguishing texts at lower and higher levels in the weekly reader corpus. Figure A2 displays the ratios obtained for the set of 464 texts in the current corpus. Note that a particularly high ratio was obtained for one of the texts in the 2013 sample. As is indicated in the box below the plot, the text is quite short and also includes a large number of words from the academic word list. In order to ensure that this outlying score would not exert an undue influence on the inferences of interest, the text was classified as an outlier and was excluded from all subsequent analyses.



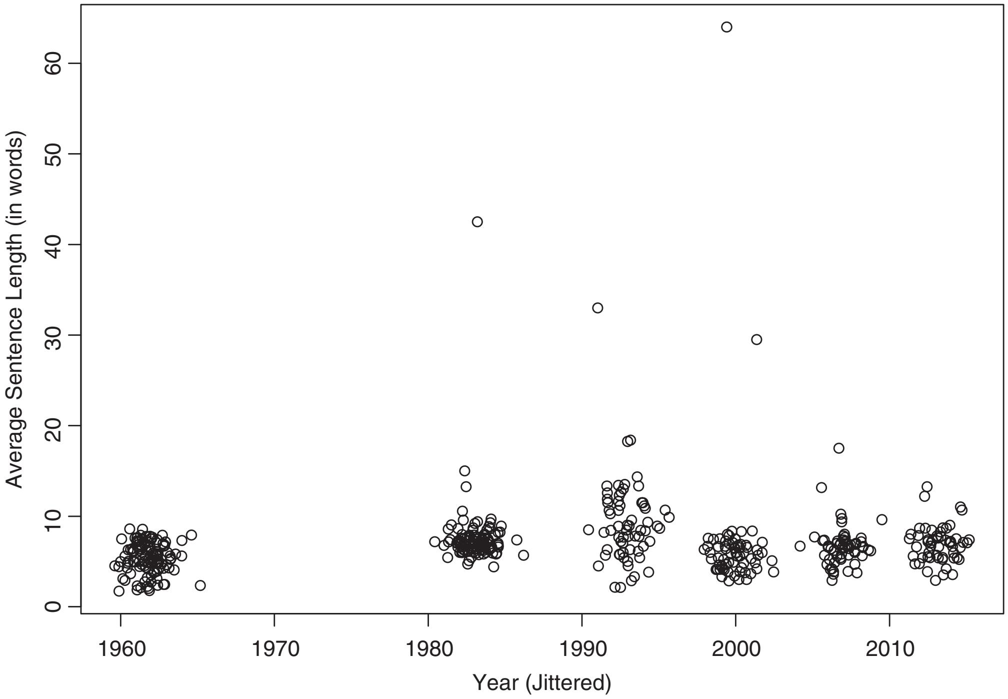
Year = 2013

```
Computers are made up of many parts. The more you get to
know each part, the easier it will be for you to use the
computer.
Cursor, keyboard, mouse, CD-rom, monitor, printer.
```

Figure A2 The ratio of academic words to total words plotted versus year (top), and a passage with an outlying value on this feature (bottom).

The Average Sentence Length Feature

The average sentence length feature is one of seven features included in the syntactic complexity component of the TextEvaluator tool. Figure A3 displays the average sentence length scores generated for each of the 464 texts in the current corpus. Note that unusually high scores were obtained for four texts. An examination of each text confirmed that, in each case, nonstandard punctuation had been employed, either because the text was intended to be read as a poem (see the left-hand text in the box below the plot) or because the text was designed to look like a student composition (see the right-hand text in the box below the plot). In order to ensure that the use of nonstandard punctuation would not exert an undue influence on subsequent inferences, these four texts were classified as outliers and were excluded from all subsequent analyses.



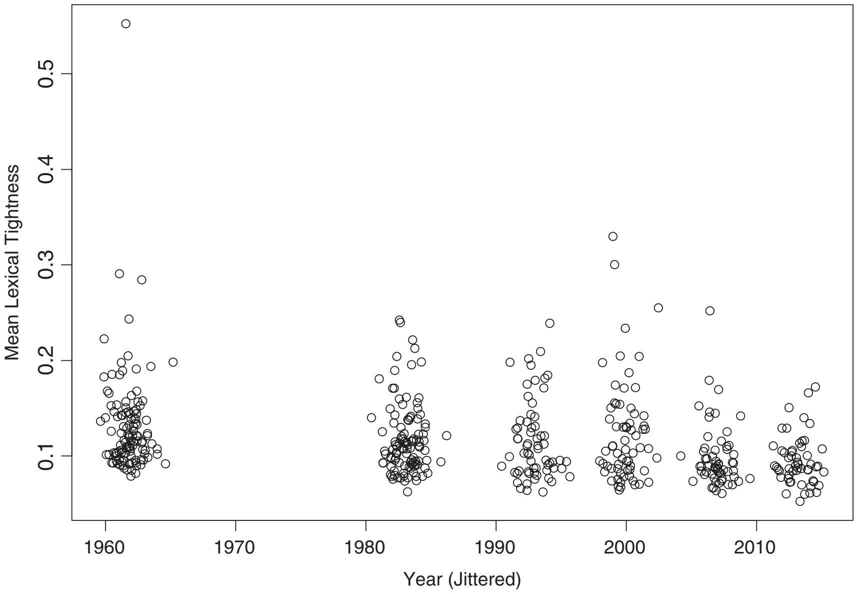
Year=1983	Year=2000
My Box. Nobody knows what's there but me, Knows where I keep my silver key And my baseball cards And my water gun And my wind up card that doesn't run, And a stone I found with a hole clear through And a blue jay feather that's mostly blue, And a note that I wrote to a guy next door And never gave him and lots, lots more Of important things that I'll never show To anyone, anyone else I know.	Wen I was in kindergarten This new girl came in our class one day And the teacher told her to sit beside me And I didn't know what to say So I wiggled my nose and made my bunny face And she laughed Then she puffed out her cheeks And she made a funny face And I laughed So then We were friends

Figure A3 Average sentence length plotted versus year (top), and two of four texts with outlying values on this feature (bottom).

The Lexical Tightness Score

The lexical tightness score measures the extent to which the individual words in a text tend to co-occur in printed text. When words frequently co-occur in text, the inferences needed to connect those words in any new text may be easier to generate. Flor et al. (2013) demonstrated that scores generated via the lexical tightness feature are highly correlated with passage GL classification provided by human experts.

Figure A4 displays the mean lexical tightness scores generated for the set of 464 texts included in the current corpus. Note that a particularly high score was obtained for one of the texts in the 1962 sample. As indicated in the box below the plot, the text includes just three unique words: *go*, *Dick*, and *help*. Because the lexical tightness feature measures both repetition and association, the text yielded an unusually high score. In order to ensure that this outlying score would not



Year = 1962

```
Go, Go, Go.

Go, Dick, go.
Go, go, go.
Help, help!
```

Figure A4 Mean lexical tightness scores plotted versus year (top), and a text with an outlying score on this feature (bottom).

exert an undue influence on the inferences of interest, the text was classified as an outlier and excluded from all subsequent analyses.

Suggested citation:

Sheehan, K. M., Flor, M., Napolitano, D., & Ramineni, C. (2015). *Using TextEvaluator® to quantify sources of linguistic complexity in textbooks targeted at first-grade readers over the past half century* (Research Report No. RR-15-38). Princeton, NJ: Educational Testing Service. http://dx.doi.org/10.1002/ets2.12085

Action Editor: Beata Beigman Klebanov

Reviewers: Paul Deane and Tenaha O'Reilly