Measuring the Power of Learning.®

Research Report
ETS RR–16-12

A Review of Evidence Presented in Support of Three Key Claims in the Validity Argument for the *TextEvaluator*® Text Analysis Tool

Kathleen M. Sheehan

June 2016

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

A Review of Evidence Presented in Support of Three Key Claims in the Validity Argument for the *TextEvaluator*® Text Analysis Tool

Kathleen M. Sheehan

Educational Testing Service, Princeton, NJ

The *TextEvaluator*® text analysis tool is a fully automated text complexity evaluation tool designed to help teachers and other educators select texts that are consistent with the text complexity guidelines specified in the Common Core State Standards (CCSS). This paper provides an overview of the TextEvaluator measurement approach and summarizes evidence related to three key claims in the TextEvaluator validity argument: (a) TextEvaluator has succeeded in expanding construct coverage beyond the two dimensions of text variation that are traditionally assessed by readability metrics; (b) the TextEvaluator strategy of estimating distinct prediction models for informational, literary, and mixed texts has succeeded in generating text complexity predictions that exhibit little, if any, genre bias; and (c) TextEvaluator scores are highly correlated with text complexity judgments provided by human experts, including judgments generated via the inheritance method and judgments generated via the exemplar method. Implications with respect to the goal of helping teachers and other educators select texts that are closely aligned with the accelerated text complexity exposure trajectory outlined in the CCSS are discussed.

Keywords readability; text complexity; genre bias; validity; TextEvaluator; Common Core State Standards

doi:10.1002/ets2.12100

The *TextEvaluator*® text analysis tool is a fully automated text complexity measurement tool designed to help teachers and other educators select texts that are consistent with the text complexity guidelines specified in the Common Core State Standards (CCSS). This paper provides an overview of the TextEvaluator measurement approach and summarizes evidence related to three key claims in the TextEvaluator validity argument.

The TextEvaluator Measurement Approach

Kane (2006) noted that "[m]easurement 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*; 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, and the observations on which conclusions are based are determined from a cognitive model of the processes engaged in by readers when attempting to make sense of texts with varying combinations of observable features. The following paragraphs illustrate how this measurement process is implemented within the TextEvaluator tool.

An Overview of the Approach

The TextEvaluator scoring engine has been under development at Educational Testing Service (ETS) for more than 8 years.[1] Each new version is implemented in four steps, as follows:

- First, a corpus of texts selected to represent the aspects of text variation addressed by students at successive points in the progression from beginning reader to proficient, college-ready reader is assembled.
- Second, a cognitive model of the processes engaged by readers during comprehension is proposed, observable text features that may facilitate or hinder the successful completion of each process are identified, and a vector of corresponding feature scores is extracted from each text.

Corresponding author: K. Sheehan, E-mail: ksheehan@ets.org

Table 1 Numbers of Passages in the TextEvaluator Corpus

Source	PCA	TC	CVC	No. of passages	Total words
Passages from high-stakes state or national reading assessments[a]	√	√		863	562,713
Passages from high-stakes college admissions assessments[b]	√	√		78	42,799
Reading passages from the Stanford Achievement Test	√		√	59	19,068
Passages from Appendix B, CCSS (CCSS Initiative, 2010)	√		√	168	80,078
Passages from Chall et al. (1996)[c]			√	52	8,193
Total				1,220	712,851

Note. PCA = included in the principal components analysis; TC = included in the TextEvaluator training corpus; CVC = included in the TextEvaluator cross-validation corpus; CCSS = Common Core State Standards.
[a]Includes passages from 24 different states and from the NAEP Reading Assessment. Fewer than half of these passages were also included in the collection of state passages analyzed in Nelson et al. (2012). [b]Includes passages from the SAT and the ACT. [c]These passages were not included in the PCA because they were added to the corpus after the PCA analysis was conducted.

- Third, a principal components analysis (PCA) is used to translate each text's vector of observed feature scores into a profile of component scores defined such that each component is focused on a single, construct-relevant dimension of text variation.
- Fourth, individual component scores are combined to form a single, overall measure of text complexity. Because many important complexity features are known to function differently within texts from different genres (Hiebert, 2012; Hiebert & Mesmer, 2013a, 2013b; Sheehan, 2013; Sheehan, Flor, & Napolitano, 2013; Sheehan, Kostin, Futagi, & Flor, 2010), three distinct prediction models are estimated: one optimized for application to informational texts, one optimized for application to literary texts, and one optimized for application to mixed texts (i.e., texts that incorporate a mixture of informational and literary elements).[2]

Additional information about each step is summarized below.

Step 1: Assemble a Representative Corpus of Texts

Biber, Conrad, and Reppen (1998) presented key design criteria to consider when conducting corpus-based analyses. They stated:

> A corpus is not simply a collection of texts. Rather, a corpus seeks to represent a language or some part of a language. The appropriate design for a corpus therefore depends upon what it is meant to represent. The representativeness of the corpus, in turn, determines the kinds of research questions that can be addressed and the generalizability of the results of the research .... [Thus] issues of representativeness in corpus design are crucial. (p. 246)

The TextEvaluator corpus is designed to permit valid inferences about the observable features of text that may contribute to lower or higher levels of comprehension difficulty. It currently includes a total of 1,220 passages comprising more than 700,000 words of text. Each passage was originally selected by a human expert for use when addressing a related measurement problem: inferring students' mastery of critical reading comprehension skills by observing their performances on specific types of reading tasks. Table 1 shows the numbers of passages selected from each of five subcorpora: (a) a collection of passages selected from high-stakes, standards-based state or national reading assessments designed to provide evidence of students' proficiencies relative to published reading standards; (b) a collection of passages selected from the reading/verbal reasoning sections of two different college admissions assessments (i.e., the *SAT*® and the ACT); (c) the set of all passages administered on Form S of the Stanford Achievement Test, Version 9; (d) the set of 52 text complexity exemplars presented in Chall, Bissex, Conrad, and Harris-Sharples (1996); and (e) the set of 168 text complexity exemplars presented in Appendix B of the CCSS (CCSS Initiative, 2010).

Table 1 also shows that, although all passages were included in one or another of the analyses reported below, certain subsets of passages were excluded during model training so that differences in grade-level (GL) expectations could be accommodated and so that a completely independent set of passages would be available for consideration in subsequent cross-validation analyses. Check marks indicate the subsets of passages included in each analysis.

An important characteristic of the passages in the TextEvaluator corpus is that a human-generated text complexity classification is available for each text. Each individual classification was obtained via one or another of two approaches: (a) the inheritance method or (b) the exemplar method. These approaches are described below.

The Inheritance Method

In this approach, passages are selected from high-stakes assessments designed to provide evidence of students' proficiencies relative to the reading skills specified in state standards documents or in alternative frameworks such as the NAEP reading framework. Passage complexity classifications are then "inherited" from the GLs targeted by parent test forms. For example, all passages selected from assessments targeted at fourth grade students are assigned a complexity classification of Grade 4. One limitation of this approach is that the passages classified at successive GLs represent a spread of complexities, from passages appropriate for the least proficient students at each targeted GL to passages appropriate for the most proficient students at each targeted GL. Because the majority of passages at each GL are designed to be appropriate for students reading near the average for their GL, however, text complexity ratings assigned via the inheritance method can help us distinguish text characteristics that tend to be more challenging for students at lower GLs and less challenging for students at higher GLs.

The Exemplar Method

Information about the aspects of text variation that distinguish texts scaling at lower or higher levels on a text complexity scale can also be developed by analyzing collections of exemplar texts. The TextEvaluator corpus currently includes two collections of exemplar texts: the set of 52 exemplars presented in Chall et al. (1996) and the set of 168 exemplars presented as Appendix B of the CCSS (CCSS Initiative, 2010). In each case, passage complexity classifications were obtained by first specifying a numeric text complexity scale and then searching for texts believed to be optimally indicative of the aspects of text variation that distinguish texts located at lower and higher levels on that scale. Note that, unlike the passage classifications obtained via the inheritance method described above, all of the passages classified via the exemplar method are selected to represent the targeted text complexity level. This difference means, for example, that all of the passages classified at Level 4 are expected to be more difficult than all of the passages classified at Level 3 and less difficult than all of the passages classified at Level 5.

Table 2 summarizes the set of 52 exemplar texts presented in the Chall et al. (1996) collection. Each passage is classified as scaling at a particular point on a quantitative text complexity scale that ranges from Level 1 (suitable for students who have successfully completed Grade 1 to Level 16 (suitable for students who have successfully completed 4 or more years of college). Chall et al. reported that the following aspects of text variation were considered by the human experts who classified each passage:

- *Language*. This aspect was evaluated by considering the proportion of words viewed as being "unfamiliar, abstract, polysyllabic, and/or technical" (p. 16).
- *Sentence complexity*. This aspect was evaluated by considering the proportion of sentences that was "longer, more complex, less direct, with greater embedding of ideas" (p. 5).
- *Conceptual difficulty*. This aspect was evaluated by considering "the conceptual understanding required to comprehend the text (e.g., the degree of abstractness, the amount of prior knowledge needed to understand the text" (p. 16).
- *Cognitive difficulty*. This aspect was evaluated by considering the amount of "thought, reasoning, analysis, and critical abilities [needed] to fully understand [the text]" (p. 6).

Resulting text complexity classifications were validated via comparisons to four types of reference scores: (a) difficulty rankings provided by groups of teachers and school administrators, (b) difficulty rankings provided by students, (c) cloze comprehension scores obtained by administering the passages to groups of students, and (d) text complexity scores obtained via readability formulas. Results reported in Chall et al. (1996) suggested that the proposed text complexity ratings were both reliable and valid and that the resulting set of 52 scaled passages can help educators understand the aspects of text variation that distinguish texts likely to be more or less difficult for students with varying levels of reading ability.

Table 2 Numbers of Exemplar Passages Comprising the Chall et al. (1996) Text Complexity Scale by Reading Level, Genre, and Content Area

		Genre = Literary			Genre = Informational			
Reading level	Score[a]	Literature	Popular fiction	Narr. soc. st.	Life sciences	Physical sciences	Expos. soc. st.	Total
1	1	1	1	1	1	0	0	4
2	2	1	1	1	1	1	1	6
3	3	1	1	1	1	1	1	6
4	4	1	1	1	1	1	1	6
5–6	5.5	1	1	1	1	1	1	6
7–8	7.5	1	1	1	1	1	1	6
9–10	9.5	1	1	1	1	1	1	6
11–12	11.5	1	0	1	1	1	1	5
13–15	14	1	0	0	1	1	1	4
16+	16	0	0	0	1	1	1	3
Total	—	9	7	8	10	9	9	52

Note. Narr. = narrative; expos. = expository; soc. st. = social studies.
[a]Score = Scaled score employed in quantitative analyses.

Table 3 Numbers of Exemplar Passages Comprising the Common Core Text Complexity Scale by Grade Band, Genre, and Content Area

		Genre = Literary		Genre = Informational		
Grade band	Score[a]	Fiction	Other literary[b]	STEM[c]	Other expository[d]	Total
2–3	2.5	10	0	6	4	20
4–5	4.5	10	0	10	10	30
6–8	7.0	15	6	9	11	45
9–10	9.5	9	2	9	19	36
11–CCR	11.5	10	0	8	20	37
Total	—	54	8	42	64	168

Note. STEM = science, technology, engineering, and mathematics.
[a]Score = Scaled score employed in quantitative analyses. [b]Includes historical documents and letters written in a literary style. [c]Includes texts from the content areas of STEM. [d]Includes historical documents, biographies, and speeches.

The TextEvaluator corpus also includes the set of 168 exemplar texts listed in Appendix B of the CCSS (CCSS Initiative, 2010). This collection was assembled by a working group that, according to the information provided in Appendix A of the CCSS, recommended texts that they or their colleagues had used successfully with students in a given grade band (GB). The three-part text complexity model described in Appendix A of the CCSS was then used to generate a final GB classification for each text. Table 3 shows the number of exemplars classified into each of the five GBs defined in the CCSS. Note that, in contrast to the 10 scale levels defined in Chall et al. (1996), this scale includes just five levels.

Alternative approaches for increasing the size of the TextEvaluator corpus were examined. For example, the strategy of including texts selected from a corpus developed by Touchstone Applied Science Associates (TASA) was evaluated. As demonstrated in Sheehan, Kostin, Napolitano, & Flor (2014), however, analyses confirmed that the computer-generated text slices in the TASA corpus differ systematically from the types of texts typically encountered by students in school and at home, so the strategy of adding TASA texts to the corpus was rejected, and the high degree of corpus representativeness recommended by Biber et al. (1998) was preserved.

Step 2: Define Construct-Relevant Text Features

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 stimulus materials (Alderson, 2000; Gernsbacher, 1990; Grabe, 2009; Just & Carpenter, 1987; Kintsch, 1998; Snow, 2002). 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 TextEvaluator feature set was developed by considering the observable features of text that might serve to facilitate or impede one or another of these processes. Natural language processing (NLP) tools designed to automatically detect the identified features were then built, and relationships between the resulting feature scores and text complexity judgments provided by human experts were quantified. Features that exhibited significant correlations with human complexity judgments were retained; those that showed no evidence of a significant correlation with human complexity judgments were rejected.

The detailed analyses conducted for each feature have been illustrated in several previous publications. For example, Sheehan et al. (2014) presented evidence focused on two particular features: one that was retained and one that was rejected. The ETS word frequency (WF) index is presented as an illustration of a feature that was retained. This highly successful feature was constructed from two large corpora: a collection of more than 1,000 articles extracted from online journals and magazines classified as appropriate for readers in Grade 3 through graduate school and a set of more than 17,000 complete books. The resulting combined corpus included more than 400 million words of running text. As is recommended in Carroll, Davies, and Richman (1971), the index provides WF estimates expressed on a standardized logarithmic scale. Analyses summarized in Sheehan et al. (2014) confirm that text complexity classifications generated via the ETS WF index are highly correlated with classifications provided by human experts.

Sheehan et al. (2014) also summarized analyses focused on the stem overlap adjacent (SOA) score, a feature that was not retained. The SOA score is calculated as the proportion of adjacent sentences that share one or more stemmed content words (i.e., treating inflected forms of a word such as *argues*, *argued*, and *arguing* as equivalent). Although researchers have frequently argued that SOA scores provide valid information about the ease or difficulty of inferring connections across sentences, studies focused on the validity of this feature have yielded mixed results. For example, McNamara, Louwerse, McCarthy, and Graesser (2010) demonstrated that SOA scores predict differences in human coherence judgments for original and modified versions of the same text. When measuring text complexity, however, we are concerned with a much more complex task (i.e., characterizing differences in the cohesion levels detected within different texts). This task is more complex because valid text-to-text comparisons are only possible when the scores generated for different texts are expressed on a common scale, a psychometric hurdle that does not exist when comparisons are limited to original and modified versions of the same text.

In several other cases (e.g., Pitler & Nenkova, 2008; Sheehan, 2013), SOA scores failed to predict variation in the human complexity judgments provided for different texts. Thus, the SOA score was not included in the subset of features considered at subsequent steps of the model development process.

Step 3: Define a Set of Construct-Relevant Component Scores

Since many of the retained features were expected to be moderately or highly correlated, analyses focused on combining evidence from multiple correlated features were implemented. Methods for addressing this problem are discussed in a number of recent papers (see e.g., Deane, Sheehan, Sabatini, Futagi, & Kostin, 2006; Graesser, McNamara, & Kulikowich, 2011; Sheehan, Kostin, & Futagi, 2007; Sheehan et al., 2010). 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.

Consistent with the above research, intercorrelations among the TextEvaluator features were investigated by implementing a PCA followed by a Promax rotation. Results suggested that more than 60% of the variation captured by the

Table 4 Dimensions of Text Variation Addressed by the TextEvaluator Tool by Targeted Cognitive Process

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

Note. Other features not listed in this table are also included in each component score. ETS = Educational Testing Service; TASA = Touchstone Applied Science Associates.

selected feature set could be accounted for via a set of eight component scores, each estimated as a linear combination of multiple correlated features. These eight components, along with illustrative text features, are listed in Table 4. A more detailed description of each component is provided in the Appendix.

The results summarized in the appendix suggest that the PCA was successful at identifying eight dimensions of text variation that are conceptually distinct yet still correlated. Correlations ranged from 0.00 to +0.60. The following section details the statistical techniques used to transform this evidence into valid, unbiased estimates of text complexity.

Step 4: Model Variation in Human Complexity Judgments

The authors of the CCSS highlight an important modeling concern: literary texts frequently employ common, everyday language to express complex, sophisticated ideas. A key implication of this fact is illustrated in Figure 1. The plot shows that the same word familiarity score is indicative of a higher GL classification if the text in question is a literary text and a lower GL classification if the text in question is an informational text. This finding suggests that any model that incorporates evidence about word familiarity without also accounting for genre effects may tend to yield predictions that fall between the two curves thereby yielding GL predictions that are too high for many informational texts and too low for many literary texts. Evidence that this prediction is true for two popular readability metrics — the Flesch–Kincaid GL Score (Kincaid, Fishburne, Rogers, & Chissom, 1975) and the Lexile tool (Stenner, Burdick, Sanford, & Burdick, 2006) — is provided in Figure 2. Each plot compares complexity scores obtained via one or the other of these two approaches to corresponding human GL judgments. Results for informational texts are plotted on the left; those for literary texts are plotted on the right. Note that in each case the predicted pattern of over- and underestimation is present.

Alternative approaches for addressing these biases are possible. For example, distinct concordance tables could be estimated for informational and literary texts. A problem with this approach, however, is that the same reported score would then have different GL interpretations depending on the genre of the text under evaluation.

Alternatively, the mapping from feature scores to overall text complexity scores could be defined such that genre effects are addressed. This second approach is the one implemented within TextEvaluator; that is, three distinct prediction models are provided: one optimized for application to informational texts, one optimized for application to literary texts, and one optimized for application to mixed texts. In each case, equations are designed to predict human GL judgments conditional on the component scores obtained via the PCA.

Results are summarized in Table 5. Two sets of regression coefficients are listed: one that is entirely estimated from informational passages ($n = 399$) and one that is entirely estimated from literary passages ($n = 452$). Common Core passages,

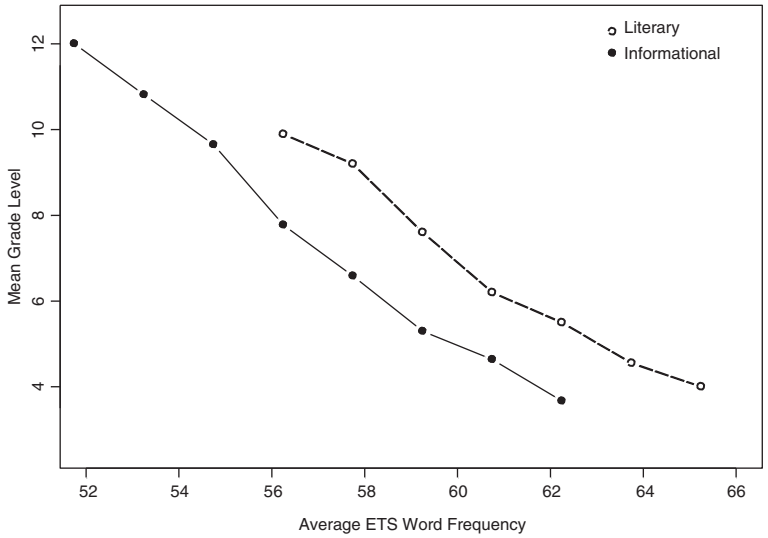

Figure 1 Mean human grade level by average ETS word frequency score for literary texts ($n = 452$) and informational texts ($n = 399$) from the training portion of the TextEvaluator corpus.

passages from the Stanford Achievement Test, and passages from Chall et al. (1996) were not included in these analyses so they would then be available for use in subsequent, cross-validation analyses.

Evidence Presented in Support of Key Claims in the TextEvaluator Validity Argument

The problem of validating claims made on the basis of scores produced by automated text complexity measurement 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 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 examines three claims that represent critical links in the TextEvaluator validity argument. Each claim is elaborated and relevant validity evidence is summarized.

Claim 1: TextEvaluator Has Succeeded in Expanding Construct Coverage Beyond the Two Dimensions of Text Variation Assessed by Traditional Readability Metrics

Two types of text complexity measurement tools are frequently discussed in the literature: (a) metrics that only measure the traditional readability dimensions of word familiarity and syntactic complexity and (b) metrics that also measure additional dimensions of text variation such as the ease or difficulty of inferring connections across sentences and the ease or difficulty of reconciling new information with prior knowledge about discourse structures. Both the Flesch–Kincaid tool and the Lexile tool belong to the first category because, in each case, all of the evidence extracted from each text is entirely focused on two particular dimensions of text variation: the ease or difficulty of determining the meaning of individual words (word familiarity) and the ease or difficulty of assembling words into sentences (syntactic complexity).

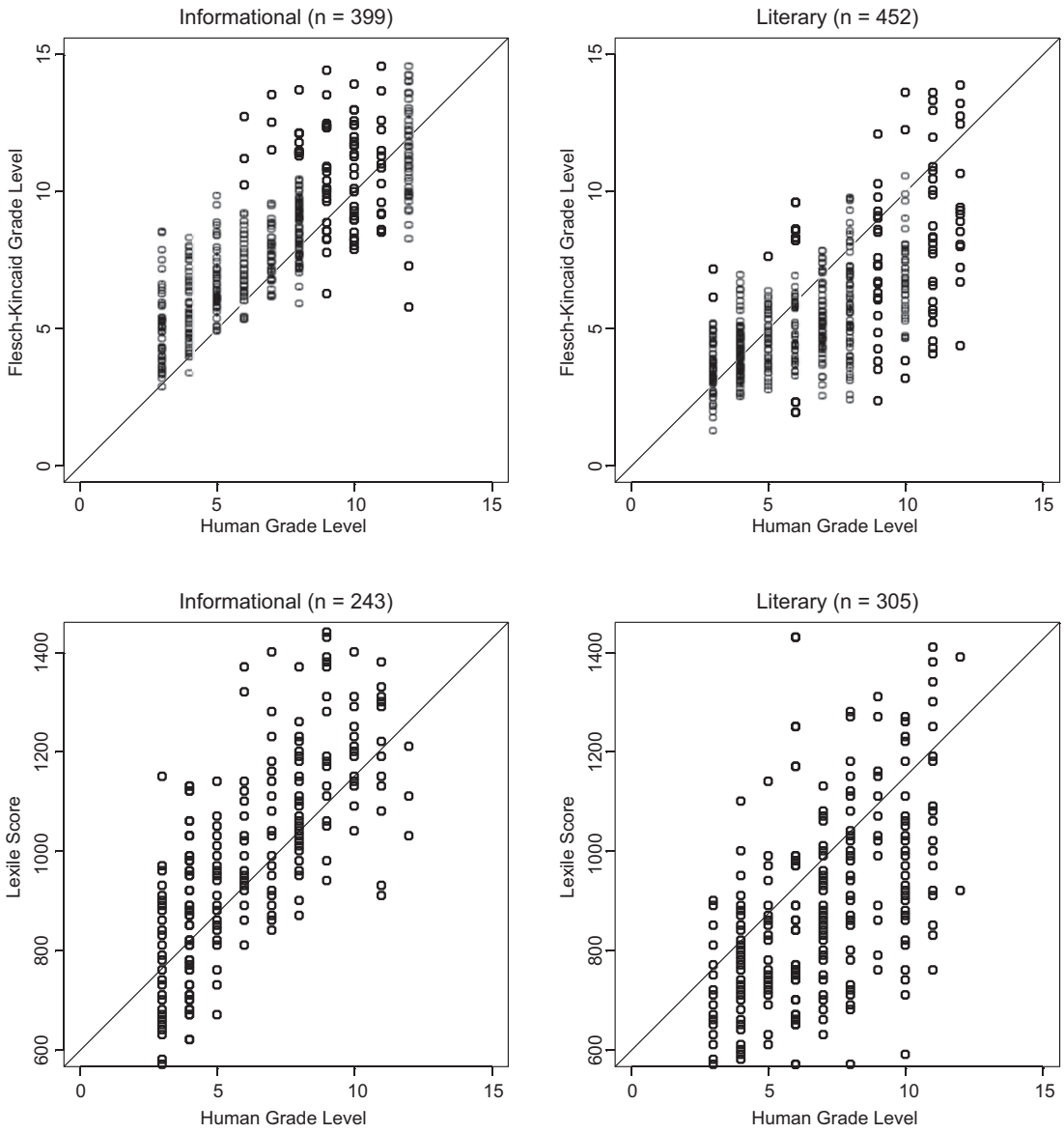

Figure 2 Overall text complexity scores generated via the Flesch – Kincaid tool (top) and the Lexile tool (bottom) compared to grade-level classifications provided by human experts, for informational texts (left) and literary texts (right).

By contrast, a key claim in the TextEvaluator validity argument is that, in addition to measuring word familiarity and syntactic complexity, TextEvaluator also measures six additional dimensions of text variation. The set of eight dimensions of text variation assessed by TextEvaluator are listed in Table 4. Additional, more in-depth descriptions are provided in the Appendix.

Table 5 presents evidence to support the claim that these additional dimensions reflect the theoretical relationships specified in many cognitively based theories of reading comprehension. The table shows that, in both the informational and literary models, one or more components associated with each of four targeted cognitive processes contribute independently to predictive accuracy. For example, three components are associated with the understanding words process: word unfamiliarity, word concreteness, and academic vocabulary. Consistent with the hypothesized cognitive model, word unfamiliarity and academic vocabulary contribute to significant increases in complexity, whereas word concreteness contributes a significant decrease in complexity.

The coefficients in Table 5 also show that two components are associated with the process of inferring connections across sentences (i.e., lexical cohesion and argumentation); two additional components (i.e., degree of narrativity and

Table 5 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	.802*	.793*
	Word concreteness	−.610*	−.483*
	Academic vocabulary	1.126*	.824*
Understanding sentences	Syntactic complexity	.983*	1.404*
Inferring connections across sentences	Lexical cohesion	−.266*	−.440*
	Argumentation	.431*	*ns*
Using knowledge of discourse structure	Degree of narrativity	*ns*	−.361*
	Interactive style	−.518*	*ns*

Note. The informational model was estimated from the set of all informational passages in the training corpus ($n = 399$) and yielded a human/automated correlation of 0.86. The literary model was estimated from the set of all literary passages in the Training corpus ($n = 452$) and yielded a human/automated correlation of 0.81.
*$p < .01$. *ns* = not significant.

interactive/conversational style) are associated with the process of using knowledge of discourse structure to develop a more complete, more integrated model of the situation presented in a text. These latter two components measure the extent to which a given text requires knowledge of more familiar discourse structures; by contrast, the argumentation component measures the extent to which a given text requires knowledge of less familiar discourse structures. Consistent with the hypothesized cognitive model, both degree of narrativity and interactive/conversational style have significant negative coefficients, whereas argumentation has a significant positive coefficient. Note that the coefficients listed for each of the other components are also consistent with the hypothesized cognitive model (i.e., the syntactic complexity component has a significant positive coefficient, and the lexical cohesion component has a significant negative coefficient). These results suggest that TextEvaluator has succeeded in expanding construct coverage to encompass additional dimensions of text variation including dimensions associated with inferring connections across sentences and dimensions associated with using knowledge of discourse structure to build a mental representation of the text that is more complete and more integrated.

Claim 2: Text Complexity Scores Generated by TextEvaluator Exhibit Little, If Any, Genre Bias

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 (2015) referred to this phenomenon as differential feature functioning (DFF). That such biases are possible has been noted in a number of recent publications. For example, the authors of the CCSS (CCSS Initiative, 2010) 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" (Appendix A, p. 7).

The degree of success achieved by TextEvaluator relative to this particular validity threat is illustrated in Figure 3. Note that the pattern of over- and underestimation that was easily distinguishable in similar displays generated from Flesch–Kincaid scores and Lexile scores is not present.

The claim that TextEvaluator scores are not subject to the types of genre biases detected when text complexity scores are generated by traditional readability metrics is also supported by the regression coefficients in Table 5. These findings suggest that TextEvaluator has succeeded in capturing important differences in the aspects of text variation that contribute to complexity variation among informational and literary texts. For example, the estimated coefficients suggest that the interactive/conversational style score is a significant component in the informational model but not in the literary model. This finding reflects the fact that, although literary texts at *all* GLs tend to exhibit moderate to high interactive/conversational style scores, moderate to high interactive/conversational style scores among informational texts tend to only be found at the lowest GLs. Thus, a moderate to high interactive/conversational style score is an indication of low complexity if the text in question is an informational text but provides no statistically significant evidence about complexity if the text in question is a literary text.

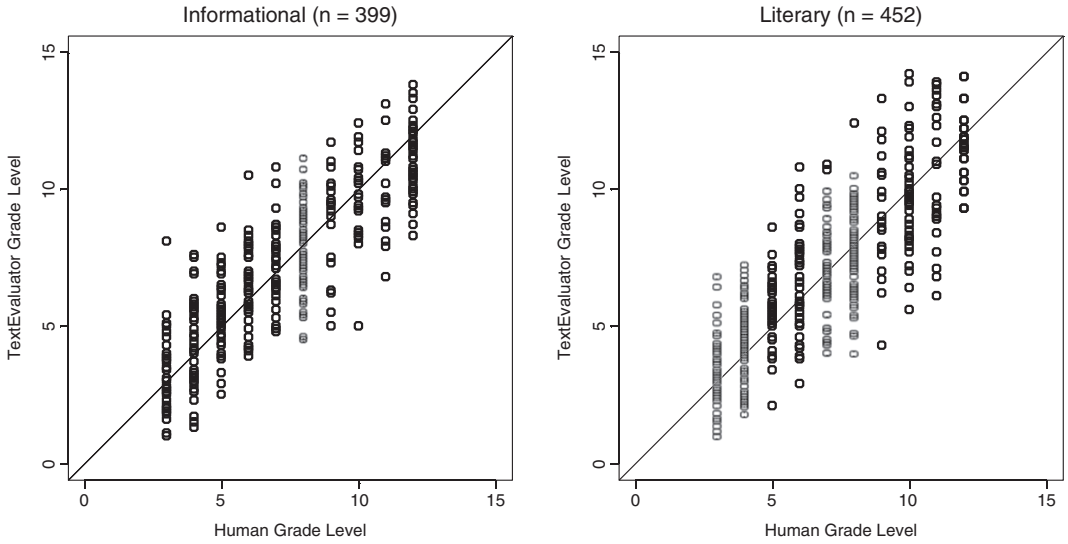

Figure 3 Overall text complexity scores obtained via TextEvaluator compared to grade-level classifications provided by human experts for informational texts (left) and literary texts (right). All informational and literary texts in the TextEvaluator training corpus are included.

Table 6 Estimated Coefficients for a Model That Predicts Variation in the Reading Level Classifications Provided for Each of the 52 Passages in the Chall et al. (1996) Corpus

Source	Coefficient	Standard error	*t* statistic	$p(>t)$
Genre	0.7475	0.8963	0.8340	.4084
TextEvaluator score	0.8152	0.0874	9.3266	.0000
Genre X TextEvaluator score	0.1072	0.1109	0.9663	.3387

Note. $N = 52$, $R^2 = 0.86$.

The claim that TextEvaluator scores are not subject to the types of genre biases detected when text complexity scores are generated by traditional readability metrics was also evaluated with respect to the set of 52 exemplar passages described in Chall et al. (1996). This additional analysis was implemented as follows. First, a model that permits estimation of genre bias, if present, was proposed, as follows:

$$Y_i = \beta_0 + \beta_1 X_i + \beta_2 G_i + \beta_3 \left(G_i * X_i\right) + \varepsilon_i \quad (1)$$

where Y_i is a human estimate of the complexity level of Text i, X_i is the text complexity score generated by TextEvaluator, and G_i is a genre indicator coded as $G_i = 1$ if the ith text is informational and $G_i = 0$ otherwise.

Next, the model was evaluated on the set of 52 exemplar passages presented in Chall et al. (1996). This set is appropriately structured for use in assessing genre effects because both informational and literary texts are included at each of 10 well-spaced text complexity levels. Results are summarized in Table 6. Note that both the main effect of genre and the interaction effect yielded coefficients that are not significantly different from zero. The table also shows that the coefficient of the TextEvaluator effect is quite large ($\widehat{\beta}_2 = 0.81$, $p < .0001$), yielding an R^2 value of 0.86. This evidence supports the claim that TextEvaluator scores do not exhibit the pattern of overestimation of informational texts and underestimation of literary texts that is typical of many traditional readability metrics.

Claim 3: TextEvaluator Scores Are Highly Correlated With Text Complexity Judgments Provided by Human Experts, Including Judgments Generated Via the Inheritance Approach and Judgments Generated Via the Exemplar Approach

Evidence related to this claim is summarized in Table 7. The table shows that relatively high correlations were obtained in each of three cross-validation datasets. The lower correlations observed for the Common Core texts are due to the fact

Table 7 Correlation Between TextEvaluator Scores and Grade Level Classifications Provided by Human Experts

Source	No. of passages	Type of corpus	Spearman correlation	Pearson correlation
Passages from high-stakes state or national reading assessments and from college-admissions assessments[a]	941	TC	0.83***	0.83***
Reading passages from the Stanford Achievement Test, Version 9, Form S	59	CVC	0.89***	0.89***
Passages from Appendix B, CCSS	168	CVC	0.72***	0.73***
Passages from Chall et al. (1996)	52	CVC	0.93***	0.91***

Note. TC = training corpus; CVC = cross-validation corpus; CCSS = Common Core State Standards. Separate correlations are reported for each of three different cross-validation corpora because the human-generated text complexity classifications obtained for these texts are not necessarily expressed on a common scale.

[a] Includes passages from 24 different state assessments, from the NAEP Reading Assessment, and from the SAT and the ACT. *** $p < .0001$

that the human-generated complexity classifications provided for these texts are reported on a 5-point scale, as opposed to the 10- or 12-point scales employed in the other two validation datasets. These high correlations suggest that the TextEvaluator measurement approach has succeeded in generating overall text complexity scores that are closely aligned with complexity classifications generated by human experts, including classifications generated via the inheritance method and classifications generated via the exemplar method.

Discussion

This paper illustrated three critical aspects of the TextEvaluator measurement approach: (a) defining observable text features that are consistent with a cognitive model of the processes engaged in by readers when attempting to make sense of complex texts, (b) enhancing scoring reliability by combining information extracted via multiple measures of the targeted dimensions of text variation, and (c) addressing genre effects by estimating distinct prediction models for informational, literary, and mixed texts.

Evidence related to three key claims in the TextEvaluator validity argument was also presented. Analyses suggested that this evidence provides support for each of the following claims:

- Claim 1: The TextEvaluator measurement approach has succeeded in expanding construct coverage beyond the two dimensions of text variation assessed by traditional readability metrics.
- Claim 2: The TextEvaluator measurement approach has succeeded in generating text complexity classifications that are not subject to the types of genre biases detected when text complexity scores are instead generated by traditional readability metrics.
- Claim 3: Text complexity scores generated via TextEvaluator are highly correlated with complexity classifications provided by human experts, including classifications generated via the inheritance approach and classifications generated via the exemplar approach.

These findings suggest that TextEvaluator scores may help teachers and other educators understand the aspects of text variation that may make texts more or less difficult for students and may facilitate the goal of ensuring that all students are exposed to increasingly complex texts at successive stages of their education.

Note, however, while many current uses of the tool are supported, there is still room for improvement. For example, alternative approaches for measuring the eight dimensions of text variation addressed by the current TextEvaluator tool should be investigated as they could lead to additional improvements in predictive accuracy. Furthermore, since additional dimensions of text variation might also play a role in determining comprehension ease or difficulty, analyses focused on the goal of measuring additional dimensions of text variation might also be of use. Differences in the numbers of features loading on each component are also a concern, as component scores that are estimated from larger numbers of features may be more reliable than component scores that are estimated from fewer numbers of features (Biber et al., 2004). Despite these limitations, however, the results summarized above suggest that text complexity scores generated via the current

TextEvaluator scoring engine may help teachers and other educators make more informed judgments when selecting texts for use in instruction and assessment.

Acknowledgments

I am grateful to Diane Napolitano, Michael Flor, and Yoko Futagi for building several of the feature extraction algorithms discussed in this paper.

Notes

1 Three patents focused on TextEvaluator's innovative measurement approach have been awarded. The first patent was filed on January 31, 2008, and was awarded on September 27, 2013. Two subsequent patents were awarded in 2014. See Sheehan, Kostin, & Futagi (2013, 2014a, 2014b).

2 The particular scoring model employed in each new scoring episode can either be selected by the user or determined via an automated genre classifier. See Sheehan, Flor, and Napolitano (2013) for a description of the TextEvaluator genre classifier.

References

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

Biber, D. (1986). Spoken and written textual dimension in English: Resolving the contradictory findings. *Language*, *62*, 394 – 414.

Biber, D. (1988). *Variation across speech and writing*. Cambridge, England: Cambridge University Press.

Biber, D., Conrad, S., Reppen, R. (1998). *Corpus linguistics: Investigating language structure and use*. Cambridge, England: 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 No. MS-25). Princeton, NJ: Educational Testing Service.

Biber, D., Johansson, S., Leech, G., Conrad, S., & Finegan, E. (1999). *Longman grammar of spoken and written English*. Essex, England: Pearson Education.

Bormuth, J. R. (1964). Mean word depth as a predictor of comprehension difficulty. *California Journal of Educational Research*, *15*, 226 – 231.

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

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.

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

Common Core State Standards Initiative. (2010). *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: 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.

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

Grabe, W. (2009). *Reading in a second language: Moving from theory to practice*. New York, NY: Cambridge University Press.

Graesser, A. C., McNamara, D. S., & Kulikowich, J. M. (2011). Coh-Metrix: Providing multilevel analyses of text characteristics. *Educational Researcher*, *40*(5), 223 – 234.

Graesser, A. C., McNamara, D. S., Louwerse, M. W., & Cai, Z. (2004). Coh-Metrix: Analysis of text on cohesion and language. *Behavior Research Methods, Instruments, & Computers*, *36*(2), 193 – 202.

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

Hiebert, E. H. (2012). The Common Core State Standards and text complexity. *Teacher Librarian*, *39*(5), 13 – 19.

Hiebert, E. H., & Mesmer, H. A. (2013a). Meeting standard 10: Reading complex text. *Principal Leadership*, *13*(5), 30 – 33.

Hiebert, E. H., & Mesmer, H. A. (2013b). Upping the ante of text complexity in the Common Core State Standards: Examining its potential impact on young readers. *Educational Researcher*, *42*(1), 44 – 51.

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 Series 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.

Kincaid, J. P., Fishburne, R. P., Rogers, R. L., & Chissom, B. S. (1975). *Derivation of new readability formulas (automated readability index, Fog count and Flesch reading ease formula) for Navy enlisted personnel* (Research Branch Report 8–75). Memphis, TN: Naval Air Station.

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

McNamara, D. S., Louwerse, M. M., McCarthy, P. M., & Graesser, A. C. (2010). Coh-Metrix: Capturing linguistic features of cohesion. *Discourse Processes*, *47*, 292–330.

Nelson, J., Perfetti, C., Liben, D., & Liben, M. (2012). *Measures of text difficulty: Testing their predictive value for grade levels and student performance*. New York: Student Achievement Partners.

Pitler, E., & Nenkova, A. (2008). Revisiting readability: A unified framework for predicting text quality. In M. Lapata & H. T. Ng (Eds.), *Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing* (pp. 186–195). Stroudsburg, PA: Association for Computational Linguistics.

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

Sheehan, K. M. (2015, 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 American Educational Research Association (AERA), Chicago, IL.

Sheehan, K. M., Flor, M., & Napolitano, D. (2013). A two-stage approach for generating unbiased estimates of text complexity. In L. Rello (Ed.), *Proceedings of the Second Workshop on Natural Language Processing for Improving Textual Accessibility (NLP4ITA)* (pp. 49–58). Stroudsburg, PA: Association for Computational Linguistics.

Sheehan, K. M., Kostin, I., & Futagi, Y. (2007, August). *Reading level assessment for literary and expository texts.* Paper presented at the 29th Annual Meeting of the Cognitive Science Society, Nashville, TN.

Sheehan, K. M., Kostin, I., & Futagi, Y. (2013). *U.S. Patent No. 8,517,738*. Washington, DC: U.S. Patent and Trademark Office.

Sheehan, K. M., Kostin, I., & Futagi, Y. (2014a). *U.S. Patent No. 8,892,421*. Washington, DC: U.S. Patent and Trademark Office.

Sheehan, K. M., Kostin, I., & Futagi, Y. (2014b). *U.S. Patent No. 8,888,493*. Washington, DC: U.S. Patent and Trademark Office.

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. 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. *The 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., & Burdick, D. (2006). How accurate are Lexile text measures? *Journal of Applied Measurement*, *7*(3), 307–322.

Tabachnick, B. G., & Fidell, L. S. (2001). *Using multivariate statistics*. Boston, MA: Allyn & Bacon.

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

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

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

Appendix

A Detailed Description of the TextEvaluator Component Scores

Hypotheses about the specific aspects of text variation measured by each of the eight components identified in the TextEvaluator PCA were evaluated by implementing a marker variable approach. Marker variables are variables that can be reasonably expected to provide relatively pure measurement of specific targeted dimensions of variation (Tabachnick & Fidell, 2001, p. 587). Examination of the available marker variables suggested that the identified components can be characterized as indicated below.

Component 1: Academic Vocabulary

Ten features loaded heavily on this dimension. Two are based on the academic word list described in Coxhead (2000). These include the frequency per thousand words of all words on the academic word list and the ratio of listed words to total words. In a previous study, Vajjala and Meurers (2012) demonstrated that the ratio of listed words to total words was effective at distinguishing texts at lower and higher levels in the weekly reader corpus. Two additional features focus on the frequency of nominalizations: one estimated from token counts and one estimated from type counts. Four additional features are based on word lists developed by Biber and his colleagues. These include the frequency per thousand words of academic verbs, abstract nouns, topical adjectives, and cognitive process nouns (see Biber, 1986, 1988; Biber, Johansson, Leech, Conrad, & Finegan, 1999; Biber et al., 2004). Two measures of word length are also included: average word length measured in syllables and the frequency per thousand words of words containing more than eight characters. Based on these results, the component is classified as measuring the extent to which the language of a text is more characteristic of academic texts than of nonacademic texts such as fiction or memoirs.

Component 2: Syntactic Complexity

Seven features loaded heavily on this component. These include features determined from the output of a syntactic parser, as well as more easily computed measures such as average sentence length and the average frequency of long sentences. Parse-based features include average number of dependent clauses, average number of words before the main verb (a measure of the extent to which the sentence is front-loaded, see Graesser et al., 2011), and an automated version of the word "depth" measure introduced by Yngve (1960). This last feature, called average maximum Yngve depth, is designed to capture variation in the memory load imposed by sentences with varying syntactic structures. It is estimated by first using a syntactic parser to assign a depth classification to each word in a text, then determining the maximum depth represented within each sentence, and then averaging resulting sentence-level estimates to obtain a passage-level estimate. Several studies of this word depth measure have been reported. For example, Bormuth (1964) reported a correlation of −0.78 between mean word depth scores and cloze fill-in rates provided by Japanese English-as-foreign-language learners. These results suggest that the second component extracted in the PCA is a measure of syntactic complexity.

Component 3: Concreteness

Words that are more concrete are more likely to evoke meaningful mental images, a response that has been shown to facilitate comprehension (Coltheart, 1981). Alderson (2000) argued that the level of concreteness present in a text is a useful feature to consider when evaluating passages for use on reading assessments targeted at readers whose first language is not English. A total of five concreteness and imageability measures loaded heavily on this dimension. All five measures are based on concreteness and imageability ratings downloaded from the Medical Research Council (MRC) Psycholinguistic Database (Coltheart, 1981). Ratings are expressed on a 7-point scale from 1 (*least concrete or least imageable*) to 7 (*most concrete or most imageable*).

Component 4: Word Unfamiliarity

This component summarizes variation detected via six different features. Two of the features are measures of word familiarity: (a) the average log WF determined via the ETS WF index and (b) the average log WF determined via the TASA index (see Zeno, Ivens, Millard, & Duvvuri, 1995). Both features have negative loadings, suggesting that the component is measuring vocabulary difficulty as opposed to vocabulary easiness. The other features with high loadings on this component are all measures of rare WF. These all have positive loadings since texts with large numbers of rare words are expected to be more difficult. Two types of rare word indices are included: indices based on token counts and indices based on type counts. Vocabulary measures based on token counts view each new word as an independent comprehension challenge, even when the same word occurs repeatedly throughout the text. By contrast, vocabulary measures based on type counts assume that a passage containing five *different* unfamiliar words may be more challenging than a passage containing the same unfamiliar word repeated five times. This difference is consistent with the notion that each repetition of an unknown word provides an additional opportunity to connect to prior knowledge.

Component 5: Interactive/Conversational Style

Many of the features with high loadings on this component also had high loadings on the first dimensions reported in one or more of the following studies: Biber (1986, 1988) and Biber et al. (2004). In each of these previous studies, the authors demonstrated that transcripts of spoken texts yielded high scores on the dimension whereas written texts yielded much lower scores. Based on this evidence, Component 5 is characterized as a measure of the degree to which a given text exhibits an interactive/conversational style as opposed to a noninteractive, nonconversational style.

Component 6: Degree of Narrativity

Three features had high positive loadings on this dimension: namely, the frequency of past perfect aspect verbs, the frequency of past tense verbs, and the frequency of third person singular pronouns. All three features have previously been classified as providing positive evidence of the degree of narrativity exhibited by a text (Biber, 1986, 1988; Graesser, McNamara, Louwerse, & Cai, 2004).

Component 7: Cohesion

Cohesion is that property of a text that enables it to be interpreted as a coherent message rather than a collection of unrelated clauses and sentences (Sheehan et al., 2014, p. 195). Halliday and Hasan (1976) argued that readers are more likely to develop a coherent mental representation of a text when certain observable features are present. These include repeated content words and explicit connectives (e.g., *consequently*, *as a result*, etc.). The seventh component extracted in the PCA includes three different types of cohesion features. The first two features measure the frequency of content word repetition across adjacent sentences within paragraphs. These measures are reported on a normalized scale designed to allow for valid text-to-text comparisons (see Sheehan, 2013).

Component 8: Argumentation

Two features have high loadings on this dimension: the frequency of concessive and adversative conjuncts and the frequency of negations (Just & Carpenter, 1987). These results suggest that texts that score high on this component may be more likely to invite the reader to consider alternative explanations or arguments. Consequently, this component is viewed as a measure of the degree of argumentation present in a text.

Suggested citation:

Sheehan, K. (2016). *A review of evidence presented in support of three key claims in the validity argument for the TextEvaluator® text analysis tool* (Research Report No. RR-16-12). Princeton, NJ: Educational Testing Service. http://dx.doi.org/10.1002/ets2.12100

Action Editor: Rebecca Zwick

Reviewers: Michael Kane and Chaitanya Ramineni