ETS TOEFL

Quality Beyond Measure.

TOEFL iBT® Research Report
TOEFL iBT–26
ETS Research Report No. RR–15-08

Analyzing and Comparing Reading Stimulus Materials Across the *TOEFL*® Family of Assessments

Jing Chen

Kathleen M. Sheehan

June 2015

The *TOEFL*® test was developed in 1963 by the National Council on the Testing of English as a Foreign Language. The Council was formed through the cooperative effort of more than 30 public and private organizations concerned with testing the English proficiency of nonnative speakers of the language applying for admission to institutions in the United States. In 1965, Educational Testing Service (ETS) and the College Board assumed joint responsibility for the program. In 1973, a cooperative arrangement for the operation of the program was entered into by ETS, the College Board, and the *Graduate Record Examinations*® (*GRE*®) Board. The membership of the College Board is composed of schools, colleges, school systems, and educational associations; GRE Board members are associated with graduate education. The test is now wholly owned and operated by ETS.

ETS administers the TOEFL program under the general direction of a policy board that was established by, and is affiliated with, the sponsoring organizations. Members of the TOEFL Board (previously the Policy Council) represent the College Board, the GRE Board, and such institutions and agencies as graduate schools of business, two-year colleges, and nonprofit educational exchange agencies.

❖ ❖ ❖

Since its inception in 1963, the TOEFL has evolved from a paper-based test to a computer-based test and, in 2005, to an Internet-based test, the *TOEFL iBT*® test. One constant throughout this evolution has been a continuing program of research related to the TOEFL test. From 1977 to 2005, nearly 100 research l reports on the early versions of TOEFL were published. In 1997, a monograph series that laid the groundwork for the development of TOEFL iBT was launched. With the release of TOEFL iBT, a TOEFL iBT report series has been introduced.

Currently this research is carried out in consultation with the TOEFL Committee of Examiners (COE). Its members include representatives of the TOEFL Board and distinguished English as a second language specialists from academia. The committee advises the TOEFL program about research needs and, through the research subcommittee, solicits, reviews, and approves proposals for funding and reports for publication. Members of the TOEFL COE serve 4-year terms at the invitation of the Board; the chair of the committee serves on the Board.

Current (2014-2015) members of the TOEFL COE are:

To obtain more information about the TOEFL programs and services, use one of the following:

E-mail: toefl@ets.org
Web site: www.ets.org/toefl

ETS is an Equal Opportunity/Affirmative Action Employer.

As part of its educational and social mission and in fulfilling the organization's nonprofit Charter and Bylaws, ETS has and continues to learn from and to lead research that furthers educational and measurement research to advance quality and equity in education and assessment for all users of the organization's products and services.

TOEFL iBT Research Report Series and ETS Research Report Series ISSN 2330-8516

RESEARCH REPORT

Analyzing and Comparing Reading Stimulus Materials Across the *TOEFL*® Family of Assessments

Jing Chen & Kathleen M. Sheehan

Educational Testing Service, Princeton, NJ

The *TOEFL*® family of assessments includes the *TOEFL*® *Primary*™, *TOEFL Junior*®, and *TOEFL iBT*® tests. The linguistic complexity of stimulus passages in the reading sections of the TOEFL family of assessments is expected to differ across the test levels. This study evaluates the linguistic complexity of each passage in a corpus of TOEFL stimulus passages. The analysis was conducted using the *TextEvaluator*®, a comprehensive text analysis system developed at Educational Testing Service (ETS). For each TOEFL reading passage, TextEvaluator provides an overall complexity score and 8 component scores that measure text complexity in specific domains. The results suggest that the overall complexity scores of the reading passages selected for use at different test levels are significantly different. According to the analysis of the component scores, the TOEFL Primary, TOEFL Junior, and TOEFL iBT passages are similar with respect to some aspects of text variation and distinct with respect to others. Score ranges based on the distributions of the overall complexity scores and the distributions of the component scores of all the passages at each test level can be used as guidelines to develop or select new passages.

Keywords TextEvaluator; TOEFL family of assessments; text analysis system; text complexity

doi:10.1002/ets2.12055

The *TOEFL*® family of assessments includes the *TOEFL*® *Primary*™, *TOEFL Junior*®, and *TOEFL iBT*® tests. The TOEFL Primary test and the TOEFL Junior test are targeted at elementary school and middle school students, respectively. The TOEFL iBT test is targeted at high school and university-aged students. The linguistic complexity of stimulus passages in the reading section of the TOEFL family of assessments is expected to differ across the test levels. Reading difficulty is determined by factors such as the linguistic complexity of a text and individual differences in readers (Droop & Verhoeven, 1998; Perfetti, Wlotko, & Hart, 2005). The linguistic complexity of a text has a large impact on reading comprehension (Barrot, 2013; Fulcher, 1997). Linguistic complexity is defined as "the amount of discourse (oral or written), the types and variety of grammatical structures, the organization and cohesion of ideas and, at the higher levels of language proficiency, the use of text structures in specific genres" (Gottlieb, Cranley, & Cammilleri, 2007, p. 46). For instance, complex vocabulary and grammatical structure will make it more difficult to understand a reading passage. Similarly, less organized and coherent ideas in a reading passage will make it difficult to understand the passage. We note that reading difficulty is determined not only by textual features but also by individual differences in readers. However, in this study we focus only on evaluation of the linguistic complexity of the TOEFL passages.

This paper evaluates the linguistic complexity of each passage in a corpus of TOEFL stimulus passages using *TextEvaluator*®[1] (Sheehan, Flor, & Napolitano, 2013; Sheehan, Kostin, Futagi, & Flor, 2010). TextEvaluator is a comprehensive text analysis system developed at Educational Testing Service (ETS). TextEvaluator provides an overall linguistics complexity score for each TOEFL reading passage and eight component scores that measure the complexity in specific domains (e.g., vocabulary difficulty). As the test level increases from TOEFL Primary to TOEFL iBT, we should expect the TOEFL reading passages to be more and more complex. Therefore, in general, reading passages at TOEFL Primary, TOEFL Junior, and TOEFL iBT levels should have increasing linguistic complexity scores as evaluated by TextEvaluator.

The linguistic complexity scores from TextEvaluator are useful in three ways: (a) helping TOEFL researchers and test developers understand similarities and differences among the linguistic features of passages developed for use on different TOEFL family assessments; (b) helping TOEFL test developers take advantage of that knowledge when selecting, evaluating, and classifying candidate passages; and (c) developing a plan for future corpus development work.

Corresponding author: J. Chen, E-mail: jchen003@ets.org

We emphasize that the complexity scores provided by TextEvaluator are only used to check or confirm whether a reading passage is placed at an appropriate level, in addition to test developers' judgment. The complexity scores should not be used as the only or main evaluation criterion to determine the test level of a reading passage, since linguistic complexity is not the only factor that determines the appropriateness of reading passages to be used at a particular test level. When the complexity score from TextEvaluator indicates that a passage is too easy or too difficult for a particular test level, test developers should use this information only with other evaluation criteria to determine whether individual passages are more or less appropriately structured for use at a particular test level.

This study was guided by the following research questions: What is the range of linguistic complexity observed across passages at each level (i.e., TOEFL Primary, TOEFL Junior, and TOEFL iBT) and what is the range of the component scores observed across passages at each level?

Reading Passages in TOEFL Assessments

Each assessment in the TOEFL family of assessments measures the ability to communicate in English in academic settings. The TOEFL Primary test assesses three skills: reading, listening, and speaking. Both the TOEFL Junior and the TOEFL iBT assess four skills: reading, listening, speaking, and writing. Reading is a part of the test across all TOEFL family assessments.

In the TOEFL Primary test, the reading section is around 30 minutes long. Students have reading tasks in two formats: reading items and reading sets. The reading materials used in the reading items format consist of separate sentences or phrases. These are not included in our analysis because TextEvaluator is designed to analyze the reading difficulty of a text passage rather than that of a single word or sentence. Only passages used in the reading sets format are included in this study. The reading sets passages range in length from 50 to 250 words. The types of the passages include academic reading, short reading, and narrative reading. Each type is explained below.

- Academic reading: A short informational text about a particular subject.
- Short reading: A brief message, such as a note or announcement.
- Narrative reading: A short narrative passage that tells a story.

In the TOEFL Junior test, the reading part takes 50 minutes, and students read three to four passages followed by 9–12 selected response questions about each passage. The passages range in length from 150 to 500 words. Passages include a number of different text types and are either academic or nonacademic (e.g., news articles or e-mails) in nature. The TOEFL Junior reading passages include the following text types:

- Expository: Material that describes events or processes objectively, categorizes information, explains situations, or presents solutions to problems.
- Biographical: Material that presents the important details of and influential moments in the life of a famous individual.
- Persuasive: Material that presents an opinion, provides evidence in support of that opinion, and may attempt to convince the reader of the correctness of a certain point of view.
- Journalism: Material that presents an account of events as they appear in a newspaper or magazine; text that includes information about the event interspersed with quotations.
- Fiction: Material that tells a story in narrative form.
- Graphic presentation of information.[2] Material that presents information in nonlinear form. Examples include schedules, advertising brochures, and bulleted announcements.
- Correspondence: Material that presents a message intended for a specific audience, either formally in the form of a business letter or e-mail, or informally in the form of a memo, friendly letter, or friendly e-mail.

In the TOEFL iBT test, the reading section takes 60–80 minutes, during which examinees read three to five passages from academic texts followed by 12–14 selected response questions about each passage. The passages range in length from 600 to 900 words. The TOEFL iBT reading passages are taken from university-level textbooks that introduce a discipline or topic. The passages cover a variety of different subjects and can be classified into three basic categories:

- Exposition: Material that provides an explanation of a topic.
- Argumentation: Material that presents a point of view about a topic and provides evidence to support it.
- Historical: Material that introduces the history of a topic.

Collecting Validity Evidence for TOEFL Reading Assessments

Enright and Tyson (2011) laid out the validity argument for the TOEFL iBT test by stating propositions that underlie the proposed test score interpretation and use and by summarizing the evidence relevant to six propositions. For instance, one proposition is that "performance on the test is related to other indicators or criteria of academic language proficiency" (p. 3). Previous studies have provided evidence for validity relevant to this proposition. Cho and Bridgeman (2012) found that TOEFL iBT scores provided information about the future academic performance of nonnative English-speaking students beyond that provided by other admissions tests. They argued that the reading and writing skills assessed in the TOEFL iBT provide unique information that could not be obtained from the other admissions tests such as the SAT® and the *GRE*® test. Another proposition is that "academic language proficiency is revealed by the linguistic knowledge, processes, and strategies test takers use to respond to test tasks" (Enright & Tyson, p. 3). For instance, Cohen and Upton (2006) collected verbal report data from 32 examinees when they responded to prototype TOEFL reading comprehension tasks closely resembling tasks now used on the TOEFL iBT test. Based on the verbal report data, they found that test takers did not rely on "test wiseness" strategies. Instead, Cohen and Upton found that test-taker strategies "reflect the fact that respondents were in actuality engaged with the reading test tasks in the manner desired by the test designers" (p. 105).

This study also aims to collect validity evidence for TOEFL reading assessments. One proposition stated in Enright and Tyson's (2011) validity argument is that the content of the TOEFL iBT test is relevant to and representative of the kinds of tasks and written and oral texts that students encounter in college and university settings. Similarly, the content of the TOEFL Primary and TOEFL Junior tests should be relevant to and representative of the tasks and texts that students encounter in K–12 schools in the United States. The complexity score provided by TextEvaluator can provide validity evidence for this proposition. It compares the linguistic complexity of TOEFL reading passages to those of the exemplar texts that were selected to be representative of the readings used at different US grade levels in the Common Core State Standards for English language arts (ELA; Common Core State Standards Initiative, 2010). Therefore, TextEvaluator evaluates whether a reading is representative of the readings that students encounter in K–12 schools in United States.

An Introduction to TextEvaluator

Evaluating TOEFL Text Complexity Using TextEvaluator

The TextEvaluator scoring engine was described by Sheehan et al. (2013). In the current version of TextEvluator, eight component scores that measure text complexity in different domains are used as predictors in linear regression models to predict human judgments of text complexity. Each component score is estimated via clusters of correlated text features extracted using natural language processing (NLP) techniques. Previous research has demonstrated that TextEvaluator provides valid information about the linguistic features of texts. For instance, Sheehan and Kostin (2012) showed that TextEvaluator provided ratings of source material comparable to ratings from human raters, which suggested that TextEvaluator was successful in capturing useful information about the characteristics of texts. Nelson, Perfetti, Liben, and Liben's study (2011) confirmed that the complexity scores obtained via TextEvaluator were highly correlated with classifications of text complexity provided by expert human raters, as well as with scores determined from student performance data.

The TextEvaluator system is uniquely suited to measure the linguistic complexity of TOEFL reading passages for two main reasons. First, it includes a wide variety of features that have been shown to be of use in assessing the linguistic complexity of passages selected for use on different types of reading assessments, including assessments targeted at second-language readers (Brown, 1998; Crossley, Greenfield, & McNamara, 2008; Greenfield, 2004). Second, it also provides a set of reference distributions designed to help test developers compare the linguistic characteristics of each new text to the range of linguistic complexity observed for texts from certain known populations, that is, texts included on the TOEFL iBT assessments.

Previous research has suggested that several of TextEvaluator's features are valid for evaluating text complexity for second-language readers. For example, features such as the average number of syllables per sentence, the percentage of function words, the number of letters per word, and the number of words per sentence were found to be effective at predicting reading difficulty for second-language readers (Brown, 1998; Greenfield, 2004). All these features are available in TextEvaluator. In the following section, we introduce the feature set of TextEvaluator along with its modeling approach.

TextEvaluator's Feature Set

The extreme complexity of the reading comprehension process suggests that large numbers of text features may be needed to adequately represent variation in text complexity (National Reading Panel, 2000; Rand Reading Study Group, 2002). For example, the Rand Reading Study Group (2002) mentioned that "processing the text involves decoding the text, higher-level linguistic and semantic processing, and self-monitoring for comprehension—all of which depend on reader capabilities as well as on the various text features" (p. xv). However, complexity is evaluated via a mere handful of text features in many popular readability formulas. For example, both the Flesch-Kincaid grade level score (Kincaid, Fishburne, Rogers, & Chissom, 1975) and the Lexile score (Stenner, Burdick, Sanford, & Burdick, 2006) rely on just two features, and the Coh-Metrix Second Language Readability Index (Crossley, Allen, & McNamara, 2011) relies on just three features.[3]

The TextEvaluator scoring engine employs eight components to assess text complexity, each of which is structured to provide evidence about a distinct dimension of text variation. The score for each component is estimated via clusters of correlated text features extracted using NLP techniques, as opposed to individual features, in order to broaden construct representation. For example, the syntactic complexity component measures the complexity of sentence structure. It is measured by a set of text features such as the average length of sentences, the average number of dependent clauses, the average number of words before the main verb, the average word depth (Yngve, 1960), and so forth. The eight components of TextEvaluator are described in Table 1.

A previous study (Sheehan & Kostin, 2012) demonstrated that, consistent with theoretical expectations, four components (syntactic complexity, vocabulary difficulty, degree of academic orientation, and argumentation) were positively associated with text complexity. The other four components (concreteness, cohesion, interactive/conversational style, and degree of narrativity) were negatively associated with text complexity. As test level increases, we should expect that the TOEFL reading passages become more and more complex. Therefore, in general, reading passages at TOEFL Primary, TOEFL Junior, and TOEFL iBT levels should have increased complexity in some or all of these components. For example, TOEFL iBT passages probably have more difficult words and include more abstract ideas than the TOEFL Primary and the TOEFL Junior passages do.

Based on previous studies (Sheehan & Kostin, 2012; Sheehan et al., 2013), the feature set used to predict text complexity is tailored for different genres of texts. TextEvaluator classifies reading materials into three genres: informational texts, literary texts, and mixed texts that incorporate a mixture of informational and literary texts. The text complexity scores of informational texts are generated based on a linear regression model with seven feature scores as the predictors (excluding degree of narrativity). The text complexity scores of literary texts are generated based on a linear regression of six feature scores (excluding argumentation and interactive/conversational style). For any mixed text, the text complexity score is an average of the text complexity scores estimated using the informational model and the literary model. Each feature has a different weight in the linear regression model that generates the overall complexity score. The model used to generate the complexity score of informational text and the model used to generate that of literary text are different (see Sheehan & Kostin, 2012).

Complexity Scores From the Common Core State Standards

The overall complexity score can be classified into a grade-level scale (see Table 2). The information in Table 2 is designed to help teachers, textbook publishers, and test developers determine the appropriate grade level at which a reading passage should be used for any passage that has been evaluated by TextEvaluator. For example, if a text evaluated by TextEvaluator has a complexity score of 10, it provides some evidence that the text is appropriate to be used for students at Grades 6–8.

This table was developed from 168 exemplar texts presented as Appendix B of the Common Core State Standards for ELA (Common Core State Standards Initiative, 2010). These texts were selected to be representative of the readings used at different US grade levels. They exemplify the level of complexity, quality, and range that the Standards require of all students in each of five possible grade bands to engage with: Grades 2–3, Grades 4–5, Grades 6–8, Grades 9–10, and Grades 11–College and Career Ready. TextEvaluator was used to analyze the linguistic complexity of all the exemplar texts. Table 2 shows the range of complexity scores for the texts in each grade band. Some relatively easier and harder readings are needed in each grade band because of the variety of reading ability among students in the same grade band. Thus the range of the complexity scores at each grade band is wide, and the ranges of complexity scores of adjacent grade bands overlap.

Table 1 Description of Eight TextEvaluator Components

Component	Sample features
1. Syntactic complexity — the complexity of sentence structure that is measured by sentence length, number of clauses, and so forth.	Average sentence length; average number of dependent clauses; average number of words before the main verb; average word depth (Yngve, 1960); average length of text span between punctuation marks (measured in log words)
2. Vocabulary difficulty — the difficulty of vocabulary that is measured by the length of the vocabulary, the rareness of the vocabulary, and so forth.	Average word frequency determined via the ETS Word Frequency Index (Sheehan et al., 2010); average word frequency determined via the TASA Word Frequency Index (Zeno, Ivens, Millard, & Duvvuri, 1995); average frequency of rare words determined via the ETS Word Frequency Index; average frequency of unique rare words determined via the ETS Word Frequency Index
3. Academic orientation — the extent to which a given text exhibits features that are more characteristic of academic texts than of nonacademic texts such as fiction or transcripts of conversations.	Average frequency of words from the Academic Word List (Coxhead, 2000); average frequency of abstract nouns; average frequency of academic verbs (apply, develop, indicate); average frequency of nominalizations; average frequency of unique nominalizations
4. Argumentation — a measure of the amount of argumentation detected in the text.	Average frequency of concessive subordinators (e.g., although, though); average frequency of adversative/concessive conjuncts (e.g., alternatively, on the other hand); average frequency of negations (e.g., no, neither, nor); average frequency of causal conjuncts (consequently, therefore)
5. Concreteness — a measure of the specificity of the text that is measured by frequency of concrete words and so forth.	Average concreteness score determined from the MRC database (Coltheart, 1981); average imageability score determined from the MRC database; average frequency of words with high concreteness ratings; average frequency of words with high imageability ratings
6. Cohesion — the property of a text that enables it to be interpreted as a coherent message rather than a collection of unrelated sentences and clauses.	Average frequency of overlapping content words (nouns and verbs) across adjacent sentences within paragraphs, adjusted to account for differences in genre and sentence length
7. Degree of narrativity — the extent to which a given text exhibits features that are more characteristic of narrative text than of nonnarrative text.	Average frequency of third person singular pronouns (he, she); average frequency of past tense verbs; average frequency of past perfect aspect verbs
8. Interactive/conversational style — a measure of the degree to which a text exhibits an interactive/conversational style as opposed to a noninteractive, nonconversational style.	Average frequency of conversation verbs (e.g., go, get, put); average frequency of contractions; average frequency of first person singular pronouns (I, me, my)

Note. Reprinted from "SourceRater: Helping teachers and test developers determine the difficulty of text for instruction and assessment," by K. M. Sheehan and I. Kostin, 2012, paper presented at the annual meeting of the National Council on Measurement in Education (NCME), Vancouver, British Columbia.

Table 2 TextEvaluator Complexity Score Ranges for Common Core Standard Readings (After Collapsing by Grade Band)

Common Core grade band	TextEvaluator complexity score range
2–3	0.4–5.6
4–5	4.0–8.4
6–8	5.9–10.9
9–10	8.4–12.3
11–College and Career Ready	9.6–13.5

Note. This table is from an ETS internal document that documents the range of the TextEvaluator complexity scores of Common Core Standard Readings.

The grade level standards in Table 2 were developed using texts that were selected to be representative of the readings used at different US grade levels. These texts may be slightly more difficult than the TOEFL reading passages, which are designed for English language learners. However, since TOEFL tests measure examinees' English language proficiency for selecting students who will study in an English-speaking country, it is reasonable to expect that the reading difficulty of TOEFL passages should be close to the difficulty of the reading passages that are representative of the reading standards specified for students at different US grade levels. Thus, the grade level standards in Table 2 provide a useful reference to compare the complexity scores of the TOEFL passages to those passages that were selected as representative of the reading standards specified for US students at different grade levels.

Currently, TextEvaluator does not provide reference scores to evaluate the complexity of each of the eight components expressed on a grade level scale. So we don't have specific reference scores to evaluate the component scores of TOEFL reading passages at each test level. But for each component, we can compare the component score of a reading passage with the component scores from the other passages at the same test level to see whether the passage is too difficult or too easy in a particular domain (e.g., vocabulary difficulty) compared to the other passages at the same test level.

Method

A total of 256 TOEFL passages was analyzed, including 17 TOEFL Primary passages, 159 TOEFL Junior passages, and 89 TOEFL iBT passages. The TOEFL Primary test was still at a very early stage of development. Therefore, only 17 TOEFL Primary passages were available for analysis. The results for TOEFL Primary passages are therefore preliminary and should be supported with future studies. All of the passages were reformatted into text files and then analyzed using TextEvaluator.

TextEvaluator provides two types of feedback about the linguistic properties of a text: an overall complexity score expressed on a grade-level scale and a profile of eight component scores. We analyzed descriptive statistics for the overall complexity scores and the component scores of the TOEFL readings at different test levels. In addition, we performed independent samples *t* tests to see whether the complexity scores and component scores of the TOEFL Junior passages are significantly different from those of the TOEFL iBT passages. Because of the small sample size of the TOEFL Primary passages, we were not able include the TOEFL Primary passages in this analysis.

Results

Complexity Scores of the TOEFL Stimulus Passages

The complexity scores estimated for each passage are summarized in Table 3. Minimum and maximum scores indicate that passage difficulty overlaps between adjacent assessments and that, on average, the TOEFL iBT passages tend to be more complex than the TOEFL Junior passages, and the TOEFL Junior passages tend to be more complex than the TOEFL Primary passages. The mean complexity scores of the TOEFL Primary, TOEFL Junior, and TOEFL iBT passages are 2.4, 6.7, and 11.9, respectively. Normality tests suggest that the complexity scores of the TOEFL Junior and TOEFL iBT passages follow normal distributions but those of the TOEFL Primary passages do not. This difference is probably because of the small sample size of the TOEFL Primary passages. Results from the independent samples *t* tests (see Table 4) suggest that the mean complexity scores estimated for the TOEFL Junior and the TOEFL iBT passages are significantly different. According to the ranges in Table 2, these results indicate that, for the most part, the reading passages selected for use on the three different TOEFL assessments tend to incorporate an appropriate level of linguistic complexity.

According to Table 2, passages with complexity scores in the range from 0.36 to 8.40 are structured in accordance with the reading standards specified for students in the primary grades, (i.e., Grades 2–5). All 17 TOEFL Primary passages included in this study fall within this range. Table 2 also shows that passages with complexity scores in the range from 5.9 to 10.9 are structured according to the reading standards specified for middle school students (i.e., Grades 6–8). While many TOEFL Junior passages fall within this range, some do not. Table 2 also shows that passages with complexity scores in the range from 9.6 to 13.5 are structured in accordance with the reading standards specified for students in Grades 11 through college. Since the TOEFL iBT assessment is targeted at the university level, TOEFL iBT passages should have complexity scores above 9.6. While most passages fall above this cutoff, some do not. The TOEFL iBT passages with complexity scores below 9.6 may exhibit complexity levels that are below what is required for successful college-level reading.

Table 3 Descriptive Statistics of the Complexity Scores and the Component Scores of the TOEFL Primary, TOEFL Junior, and TOEFL iBT Passages

	Passage level	*N*	Mean	Median	*SD*	Min	Max
TextEvaluator complexity score	TOEFL Primary	17	2.4	1.7	1.526	1.0	5.8
	TOEFL Junior	159	6.7	6.8	1.650	1.8	12.2
	TOEFL iBT	89	11.9	12.0	1.225	8.9	14.2
	Total	265	8.2	7.8	3.217	1.0	14.2
1. Syntactic complexity	TOEFL Primary	17	33.3	28.0	12.539	17	56
	TOEFL Junior	159	50.7	52.0	8.451	30	66
	TOEFL iBT	89	67.7	68.0	5.237	51	79
	Total	265	55.3	56.0	12.531	17	79
2. Vocabulary difficulty	TOEFL Primary	17	29.9	28.0	10.662	15	54
	TOEFL Junior	159	50.2	51.0	11.525	19	79
	TOEFL iBT	89	79.7	79.0	6.124	62	93
	Total	265	58.8	58.0	18.552	15	93
3. Degree of academic orientation	TOEFL Primary	17	22.8	20.0	10.791	6	50
	TOEFL Junior	159	50.2	51.0	12.184	14	83
	TOEFL iBT	89	76.7	77.0	8.943	52	95
	Total	265	57.3	57.0	18.893	6	95
4. Argumentation[a] (informational texts only)	TOEFL Primary	8	26.0	26.0	18.385	7	53
	TOEFL Junior	115	51.6	51.0	19.643	7	90
	TOEFL iBT	89	52.2	52.0	13.825	20	82
	Total	212	50.9	51.0	17.998	7	90
5. Cohesion	TOEFL Primary	17	58.0	56.0	12.505	42	81
	TOEFL Junior	159	53.6	53.0	10.485	29	81
	TOEFL iBT	89	55.0	54.0	8.990	38	76
	Total	265	54.4	54.0	10.175	29	81
6. Concreteness	TOEFL Primary	17	60.5	62.0	12.304	39	86
	TOEFL Junior	159	43.7	44.0	10.333	19	72
	TOEFL iBT	89	29.0	28.0	9.272	13	59
	Total	265	39.9	40.0	13.336	13	86
7. Degree of narrativity[a] (literary texts only)	TOEFL Primary	6	71.0	68.5	13.446	56	90
	TOEFL Junior	38	72.2	76.0	12.429	40	90
	Total	44	72.0	75.5	12.414	40	90
8. Interactive/conversational style[a] (informational texts only)	TOEFL Primary	8	24.0	25.0	12.995	5	45
	TOEFL Junior	115	39.9	40.0	14.030	5	74
	TOEFL iBT	89	23.3	24.0	9.903	5	47
	Total	212	32.3	31.5	14.855	5	74

[a] The degree of narrativity component is only considered when evaluating literary passages, a genre type that rarely occurs on TOEFL iBT assessments, but occasional occurs on TOEFL Junior or TOEFL Primary assessments. Similarly, the argumentation and the interactive/conversational style components are only considered when evaluating informational passages.

Figure 1 illustrates the distributions of the complexity scores of the TOEFL Primary, TOEFL Junior, and TOEFL iBT passages estimated via TextEvaluator. From this figure, one can see that the complexity scores of the TOEFL iBT passages are in the high range, the complexity scores of the TOEFL Junior passages are in the medium range, and those of the TOEFL Primary passages are in the low range. The complexity scores of the reading passages are clearly different across these three test levels, which reflect significant differences in the complexity of the passages used in the three different TOEFL family assessments.

From Figure 1, one can see that the complexity scores of some TOEFL Junior and TOEFL iBT passages are much higher or lower than the majority of the other passages used at the same test level. Figure 2, box plots of the complexity scores of the TOEFL Primary, TOEFL Junior, and TOEFL iBT passages, presents the outliers of the complexity scores at each test level. For example, the complexity score of a TOEFL Junior passage (Passage 176) is 12.2, which is very high compared to that of the other TOEFL Junior passages. Meanwhile, the complexity scores of two TOEFL Junior passages (Passage 18 and Passage 19) are 1.8 and 2.1, respectively, which are much lower than those of most of the TOEFL Junior passages. These passages are considered as outliers. These passages may need to be revised to make the difficulty of these passages more consistent with the difficulty of the other passages at the same test level.

Table 4 Independent Samples *t* Test Comparing Mean TextEvaluator Complexity Score and Eight Component Scores Across TOEFL Junior and TOEFL iBT Passages

	Mean difference	*SE* difference	*p* (2-tailed)
TextEvaluator complexity score	−5.171	.1843	.000
1. Syntactic complexity	−16.981	.870	.000
2. Vocabulary difficulty	−29.481	1.121	.000
3. Academic orientation	−26.538	1.354	.000
4. Argumentation	−1.010	2.098	.631
5. Cohesion	−1.394	1.321	.292
6. Concreteness	14.731	1.319	.000
7. Degree of narrativity	6.745	2.294	.004
8. Interactive/conversational style	24.982	1.871	.000

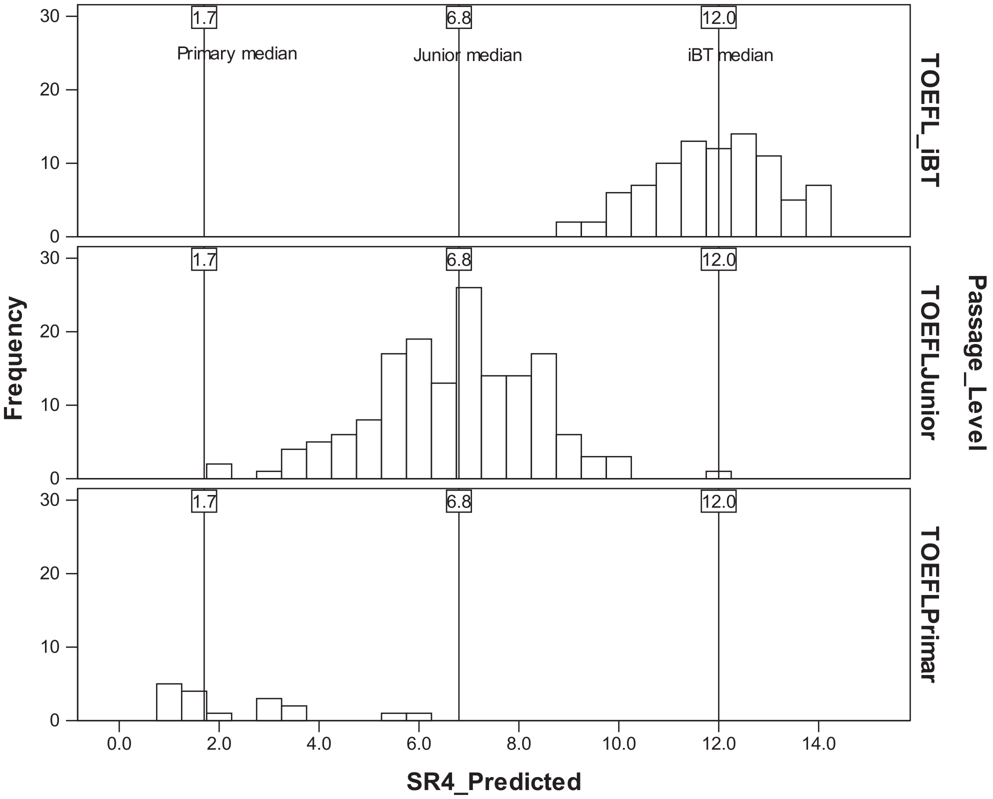

Figure 1 Distribution of the complexity scores of the TOEFL Primary, TOEFL Junior, and TOEFL iBT passages.

Component Scores of the TOEFL Stimulus Passages

This report also provides an analysis of the component scores estimated for each passage. The eight component scores of the reading passages can provide more detailed information about each reading passage. If the overall complexity score of a reading passage suggests that the passage is too easy or too difficult for a target test level, we can identify which aspect of the passage makes it too easy or too difficult using the component scores, and corresponding changes can be made to make the reading passage appropriate for the target level.

In general, when the test level increases, the reading passages should be more complex and the scores of the first four components (syntactic complexity, vocabulary difficulty, degree of academic orientation, and argumentation) should increase, while the scores of the other four components should decrease. Table 3 provides the mean, the standard deviation, and the maximum and minimum values of each of the eight TextEvaluator component scores for all of the TOEFL passages. Table 3 shows that as the test level increases, the mean scores of Components 1–4 increase and the mean score

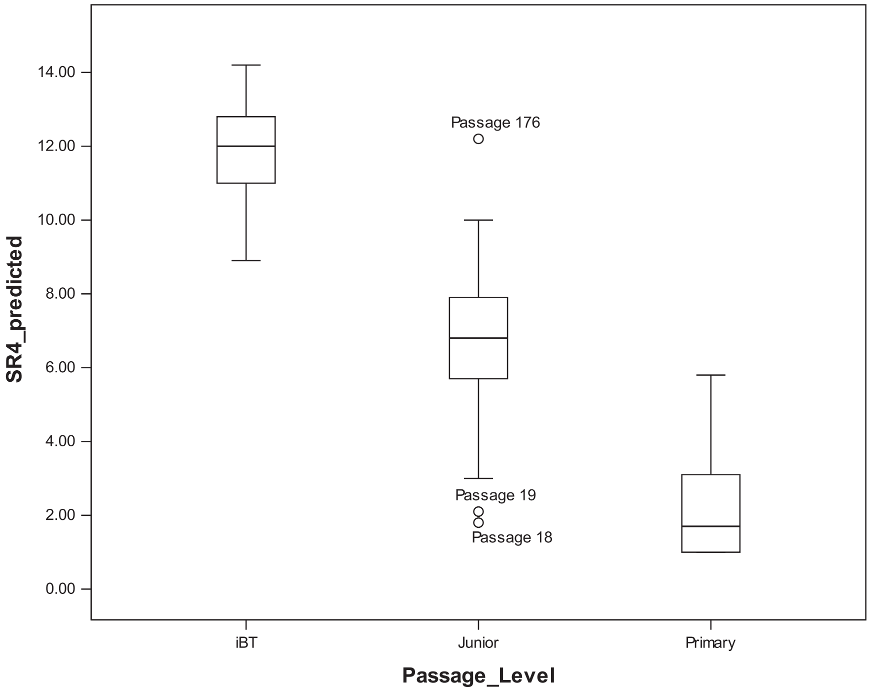

Figure 2 Box plot of the complexity scores of the TOEFL Primary, TOEFL Junior, and TOEFL iBT passages.

of Component 6 decreases, which is consistent with what is expected. However, the mean scores of the other three components do not decrease as expected. We conducted independent samples *t* tests to see whether the component scores are significantly different between TOEFL Junior and TOEFL iBT test levels. Because of the small sample size of the primary passages, we did not include TOEFL Primary passages in our test. Independent samples *t* test results presented in Table 4 suggest the reading passages across these two test levels are clearly distinct in their complexity in all the components except argumentation and cohesion.

The descriptive statistics presented in Table 3 and the independent samples *t* test results presented in Table 4 suggest that the TOEFL Primary, TOEFL Junior, and TOEFL iBT passages are more similar with respect to some aspects of text variation, and more distinct with respect to others. Because of small sample size of TOEFL Primary passages ($n = 17$) and that of literary passages ($n = 6$ for TOEFL Primary passages and $n = 38$ for TOEFL Junior passages), outliers may have a big impact on the descriptive statistics presented in Table 3. Therefore, the findings need to be verified with a larger sample size in future studies.

Table 3 provides the range of each component score of all the passages at each test level. These ranges can be used as a guideline for developing or selecting new passages so that they are more likely to achieve the targeted overall text complexity at each test level. For example, the TOEFL iBT passage (Passage 206) with the highest syntactic complexity score has a score of 79, and the TOEFL iBT passage (Passage 181) with lowest syntactic complexity score has a score of 51 (texts of these two passages are available to TOEFL assessment developers upon request). Since the TOEFL iBT passages included in this study are representative of all the TOEFL iBT passages, 51 – 79 is a reasonable range into which the syntax complexity score of an iBT passage should fall. When developing or selecting new iBT passages, if the syntax complexity score of the passage is far out of the range from 51 to 79, it is likely that the syntax used in this passage may be too difficult or too easy, and the overall text complexity might be inappropriate for the TOEFL iBT test. Similarly, for each of the other components, score ranges based on the distribution of the component scores listed in Table 3 can be used as a guideline for passage development and selection.

Discussion and Conclusions

Using appropriate stimulus reading materials for each TOEFL family assessment is important because this helps to assess examinees' English language proficiency more precisely. TextEvaluator may be a useful tool to aid the selection, evaluation, and classification of appropriate stimulus materials for the TOEFL family of assessments. However, reading difficulty is determined not only by textual features, but also by individual differences in readers. Additional aspects of text variation beyond those measured by TextEvaluator also play a role in determining the suitability of a particular text for use in TOEFL.

The analyses confirmed that the complexity scores of the reading passages at different TOEFL test levels are significantly different and tend to incorporate an appropriate level of linguistic complexity; that is, they tend to fall within the range of variation expected for readers at the specified grade levels according to the Common Core Concordance Table (Table 2). However, in some cases, passages were classified as having overall complexity scores that were either too low or too high for readers in the targeted population, according to the Common Core Concordance Table. These passages may need further review and/or modification. The complexity scores of the passages at each test level can be used as a reference to develop and select new passages. For instance, a targeted complexity score range based on the distribution of the complexity scores of all the passages at one test level (e.g., mean ± 2 SD) can be used as a guideline to develop future passages at each test level.

It is worth noting that the grade level standards specified in the Common Core Concordance Table were developed based on some representative readings used at different US grade levels. The standards specified in the Common Core Concordance Table may be too high for evaluating the reading materials of a language test, especially at the TOEFL Primary and TOEFL Junior test levels, where the reading ability of native and nonnative speakers may differ substantially. So the grade level standards can only be taken as a frame of reference to compare the text complexity of TOEFL reading materials to that of the typical readings used at different US grade levels and should not be taken as fixed standards. Another use of the complexity score is to compare the text complexity among TOEFL readings to see whether some readings are much more difficult or easier compared to the others at the same test level.

Some students' reading ability may be far above or below the level of the test that they are taking. To measure the reading ability of these students, it is desirable to have some passages that are somewhat outside of the targeted range. Our results suggest that there are some TOEFL Junior level passages that are much more difficult or easier than the other TOEFL Junior level passages. The overall text complexity scores of the TOEFL Junior level passages fall in a wide range from 1.8 to 12.2 compared to the range of the text complexity scores of TOEFL Junior level passages specified in the Common Core Concordance Table, which is between 5.9 and 10.9. This suggests that the TOEFL Junior level reading materials can also measure the reading ability of students who are somewhat outside of the targeted range. When developing and selecting reading passages at each test level, test developers need to consider the extent to which it is desirable to have passages with a wide range of text complexity levels to more precisely assess students who are far below or above the test level.

According to the analysis of the component scores, the TOEFL Primary, TOEFL Junior, and TOEFL iBT passages are more distinct with respect to some aspects of text variation (e.g., syntactic complexity and vocabulary difficulty) and more similar with respect to others (e.g., argumentation and cohesion). This result needs to be verified with a larger sample size of TOEFL Primary texts and literary texts. To further distinguish the levels of linguistic complexity presented among TOEFL passages targeted at different levels, test developers may consider paying more attention to the components of text variation when selecting and evaluating passages.

This study provides the range of each component score for all the passages at each test level. These ranges can be used as a guideline to develop or select new passages to achieve the targeted overall text complexity at each test level. When developing or selecting a new TOEFL passage, it is useful to check whether the component scores of the passage fall within the identified ranges. If a reading passage is too easy or too difficult for a particular test level, it is likely that one or more of the component scores of the passage is out of the identified range. Test developers might consider modifying passages in one or more particular domain(s) to achieve the targeted complexity. For example, if the syntactic complexity score of a passage is too high, test developers can reduce text complexity by using simplified sentences that convey the same meaning.

Limitations of the Study

Only 17 TOEFL Primary passages are included in this study because the TOEFL Primary test is still in the development stage. The TOEFL Primary passages included in this study came from a pilot test, which might be changed slightly in real administration. Owing to the small sample size of the TOEFL Primary reading passages, the overall text complexity scores and the component scores of the TOEFL Primary passages may not represent the pool of the TOEFL Primary passages after more reading passages are developed in the near future. So the conclusions about the TOEFL Primary reading passages can only be generalized with caution.

Readability of texts is determined by many factors besides linguistic complexity. However, TextEvaluator mainly evaluates linguistic complexity based on some features that have been proven to be related to linguistic complexity. It is also likely that the evaluation of linguistic complexity is not complete. In addition, similar to many automated scoring engines, TextEvaluator cannot evaluate propositions or content of the readings as human raters do. For example, TextEvaluator cannot identify the topic of the reading and therefore cannot judge the difficulty of the passage based on the background knowledge required to understand the passage. In particular, literary texts have a lot of implied meanings, which depend not only on the text itself, but also on the context, the inferred intent of the author, any preexisting knowledge related to the text, and so forth. TextEvaluator does not provide a measure of pragmatics for literary texts. Therefore, the features of TextEvaluator cannot provide a comprehensive evaluation of the difficulty of TOEFL reading stimulus materials. The results from this study need to be integrated with other information to evaluate TOEFL readings to determine whether individual passages are more or less appropriately structured for use at a certain test level.

Acknowledgments

The study was funded by the ETS research allocation under the initiative of English Language Learning and Assessment. Several members of the ETS staff in addition to the authors contributed to this study. Mitchell Ginsburgh and Paul Rybinski prepared the TOEFL reading passages. Michael Flor reformatted the TOEFL reading passages, and Diane Napolitano analyzed the reading passages using TextEvaluator. We would like to thank these people for their great contributions. We also would like to thank Gary Ockey for his helpful feedback on an earlier version of the report.

Notes

[1] TextEvaluator was formerly called SourceRater.

[2] In this study, the TOEFL Junior passages in the graphic presentation information category were not included, because TextEvaluator is not designed to analyze graphic presentations.

[3] The Coh-Metrix system calculates a large number of features, but the Coh-Metrix Second Language Readability Index only uses three of them.

References

Barrot, J. S. (2013). Revisiting the role of linguistic complexity in ESL reading comprehension. *Southeast Asian Journal of English Language Studies*, *19*(1), 5–18.

Brown, J. (1998). An EFL readability index. *JALT Journal*, *20*(2), 7–36.

Cho, Y., & Bridgeman, B. (2012). Relationship of TOEFL iBT scores to academic performance: Some evidence from American universities. *Language Testing*, *29*(3), 421–442.

Cohen, A. D., & Upton, T. A. (2006). *Strategies in responding to the New TOEFL reading tasks* (TOEFL Monograph Series Report No. TOEFL-MS-33). Princeton, NJ: Educational Testing Service. Retrieved from http://www.ets.org/Media/Research/pdf/RR-06-06.pdf

Coltheart, M. (1981). The MRC psycholinguistic database. *Quarterly Journal of Experimental Psychology*, *33*(4), 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.

Crossley, S. A., Allen, D., & McNamara, D. S. (2011). Text readability and intuitive simplification: A comparison of readability formulas. *Reading in a Foreign Language*, *23*(1), 84–101.

Crossley, S. A., Greenfield, J., & McNamara, D. S. (2008). Assessing text readability using cognitively-based indices. *TESOL Quarterly*, *42*(3), 475–493.

Droop, M., & Verhoeven, L. (1998). Background knowledge, linguistic complexity, and second-language reading comprehension. *Journal of Literacy Research*, *30*, 253–271.

Enright, M., & Tyson, E. (2011). *Validity evidence supporting the interpretation and use of TOEFL iBT scores* (TOEFL iBT Research Insight, Series I, Volume 4). Princeton, NJ: Educational Testing Service.

Fulcher, G. (1997). Text difficulty and accessibility: Reading formulae and expert judgement. *System*, *25*(4), 497–513.

Gottlieb, M., Cranley, M. E., & Cammilleri, A. (2007). *Understanding the WIDA English language proficiency standards: A resource guide.* Madison: University of Wisconsin Board of Regents System. Retrieved from http://www.wida.us/standards/Resource_Guide_web.pdf

Greenfield, J. (2004). Readability formulas for EFL. *JALT Journal*, *26*(1), 5 – 24.

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 No. 8-75). Millington, TN: Naval Air Station Memphis.

National Reading Panel. (2000). *Teaching children to read: An evidence-based assessment of the scientific research literature on reading and its implications for reading instruction*. Washington, DC: National Institute of Child Health and Human Development.

Nelson, J., Perfetti, C., Liben, D., & Liben, M. (2011). *Measures of text difficulty*: *Testing their predictive value for grade levels and student performance.* Technical Report to the Gates Foundation (also to be submitted for publication). Retrieved from http://www.ccsso.org/Documents/2012/MeasuresofText Difficulty_final.2012.pdf

Perfetti, C. A., Wlotko, E. W., & Hart, L. A. (2005). Word learning and individual differences in word learning reflected in event-related potentials. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *31*(6), 1281 – 1292.

Rand Reading Study Group. (2002). *Reading for understanding: Toward an R & D program in reading comprehension*. Santa Monica, CA: Rand.

Sheehan, K. M., Flor, M., & Napolitano, D. (2013, June). 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 (NLP4ITA), Annual Conference of the Association for Computational Lingusitics* (pp. 49 – 58). Stroudsburg, PA: Association for Computational Linguistics.

Sheehan, K. M., & Kostin, I. (2012, April). *SourceRater: Helping teachers and test developers determine the difficulty of text for instruction and assessment.* Paper presented at the annual meeting of the National Council on Measurement in Education (NCME), Vancouver, British Columbia.

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

Stenner, A. J., Burdick, H., Sanford, E. E., & Burdick, D. (2006). How accurate are Lexile text measures? *Journal of Applied Measurement*, *7*(3), 307 – 322.

Yngve, V. H. (1960). A model and an 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.

Suggested citation:

Chen, J., & Sheehan, K. W. (2015). *Analyzing and comparing reading stimulus materials across the TOEFL® family of assessments* (TOEFL iBT Research Report No. 26, ETS Research Report No. RR-15-08). Princeton, NJ: Educational Testing Service. 10.1002/ets2.12055

Action Editor: Gary Ockey

Reviewers: This report was reviewed by the Research Subcommittee of the TOEFLCommittee of Examiners