Journal of Textbook Research
Vol. 15, No. 1 (April, 2022), 43-80
doi: 10.6481/JTR.202204_15(1).02

Do Senior High School English Textbooks Sufficiently Prepare Students for the High-Stakes College Entrance Examinations? A Corpus-Based Analysis of Text Difficulty

Yuh-Show Cheng Sheng-Chieh Chang

The aim of this study was to uncover how sufficiently senior high school English textbooks in Taiwan prepare students for reading the passages in the high-stakes college entrance examinations in terms of text difficulty. A corpus-based approach was adopted to compare the vocabulary load and readability of passages extracted from senior high school textbooks and from the English test papers of college entrance examinations. Two corpora were compiled: a textbook corpus comprising texts extracted from all five editions of Ministry of Education (MOE)-authorized senior high school textbooks and a test corpus containing all of the reading passages in the college entrance English tests from the 2002 to 2017 school years. The results indicate that the passages in the English textbooks do not match those in the tests in terms of the vocabulary load and several Coh-Metrix readability metrics. The reading passages in the English tests generally have lower overall readability, lower narrativity, and higher syntactic complexity than those in the textbooks. The test passages also require a much larger vocabulary size than the textbooks do. Implications of the findings for students, textbook writers, English teachers, and the MOE are provided.

Keywords: corpus-based, textbook analysis, vocabulary load, readability

Received: August 13, 2021; Revised: November 24, 2021; Accepted: March 18, 2022

Yuh-Show Cheng, Professor, Department of English, National Taiwan Normal University, E-mail: t22035@ntnu.edu.tw
Sheng-Chieh Chang, Research Assistant, Department of English, National Taiwan Normal University.

教科書研究
第十五卷 第一期（2022 年 4 月），43-80
doi: 10.6481/JTR.202204_15(1).02

高中英文教科書是否準備好學生面對高風險的大學入學考試？一個語料庫為本的文本困難度分析研究

程玉秀　張盛傑

本研究旨在發掘就文本困難度而言，臺灣高中英文教科書是否準備好學生閱讀大學入學考試英文試題之文本。以語料庫爲本的方法，本研究建置了兩個語料庫：教科書語料庫和大學入學考試試題語料庫，進而比較高中教科書和大學入學考試試題英文文本之詞彙量和可讀性。教科書語料庫涵蓋研究期間市面流通的所有審定本高中英文教科書，共五套；試題語料庫則包含了 2002 年至 2017 年大學入學考試英文學測和指考的所有篇章。研究結果顯示高中英文教科書和大學入學考試的英文文本，在詞彙量和數個 Coh-Metrix 可讀性指標上有落差。學測和指考文本的整體可讀性和敘事性通常顯著低於教科書文本，但其語法複雜性則顯著高於教科書文本。閱讀學測和指考試題文本所需的詞彙量也遠高於閱讀教科書文本所需的詞彙量。文末，作者依據研究結果對學生、教科書編輯者、英語教師和教育部提出一些建議。

關鍵詞：語料庫為本、教科書分析、詞彙量、可讀性

收件：2021年8月13日；修改：2021年11月24日；接受：2022年3月18日

程玉秀，國立臺灣師範大學英語學系教授，E-mail: t22035@ntnu.edu.tw
張盛傑，國立臺灣師範大學英語學系研究助理

1. Introduction

Although researchers in the field of second and foreign language (L2) have disagreed regarding the role of language input in L2 learning, they generally concur that language input is a prerequisite of L2 development (Ellis, 2005). L2 learners acquire knowledge of vocabulary, grammar, and text structure mainly through exposure to and comprehension of oral and written L2 texts. Written texts in particular are "one of the best and most in-depth means of providing or receiving target language input" (Frantzen, 2010, p. 34).

Among various types of written input, school textbooks are usually the principal source of language input for L2 learners, especially those who learn the target language primarily at school. Textbooks thus play a crucial part in determining the content of language lessons and the L2 knowledge students may acquire (Zhang, 2017). In light of the profound influence of textbooks on learning of L2 in instructed settings, researchers have called for analyses of textbook corpora to determine the quality and quantity of input available to learners (Zyzik, 2009); this study responds to this call. Because of the emphasis Asian societies such as Taiwan place on success on high-stakes examinations (Hill, 2010) such as college entrance exams, this study examined the extent to which the text difficulty of senior high school textbooks match that of reading passages on college entrance exams in Taiwan. The results could reveal the characteristics of school textbooks and illuminate how well textbooks in a foreign language learning context prepare students for high-stakes entrance exams.

2. Literature Review

2.1 College Entrance Examinations and Senior High School English Instruction in Taiwan

In Taiwan, high-stakes examinations such as college entrance examinations wield enormous sway over classroom instruction because success on these exams determines "the possibilities of students' academic and career success in the future" (Chen & Huang, 2017, p.6). Teachers tend to "teach to the test"

and prioritize helping students pass or achieve high scores on entrance exams (Chen & Huang, 2017; Reynolds et al., 2018). The influence of the entrance exams on curriculum design and pedagogical practice in English classes may exceed that of the national curriculum guidelines for basic education (from primary to senior high school education), which serve a regulatory function by establishing national goals for the development of school curricula and classroom teaching. Although the curriculum guidelines state that the goal of English education is to cultivate students' English communication skills (including listening, speaking, reading, and writing), most senior high school English teachers focus on the two literacy skills assessed on college entrance exams: reading and writing (Lin, 2018; Wang, 2008). Moreover, because a large portion (72 out of 100 points) of the exam is allocated to the reading section, which assesses grammar, vocabulary knowledge, and reading comprehension, teachers usually prioritize instruction of grammar, vocabulary, and reading skills.

The washback effect of college entrance examinations on high school English instruction also manifests in the word list compiled by the College Entrance Examination Center (CEEC), the institute responsible for developing and administering college entrance exams in Taiwan. The curriculum guidelines for English specify core competences but do not stipulate specific content for instruction (e.g., specific English vocabulary words, grammatical concepts, sentence structures, or text types). The only exception is a list of 2,000 English words provided as an appendix to the Curriculum Guidelines for Primary and Junior High School (Grades 3 to 9) English.[1] For senior high school (Grades 10 to 12), no specific word list is provided in the guidelines. Instead, the CEEC compiles and regularly updates the English word lists.

The CEEC's word list in use at the time of this study, the Senior High School English Reference Word List (SHERWL), was released in 2002. It contains 6,480 words classified into six levels, with each level containing 1,080 words. The vocabulary listed in the higher levels is assumed to be more difficult and less commonly used than those listed in the lower levels. According to the

[1] Among the 2,000 words, 1,200 words are set as the minimum basic English vocabulary that all students in Taiwan should master before graduating from junior high school. However, the order and time in which these words are taught are flexible and left for the teachers and textbook publishers to determine.

introduction to the 2002 version of SHERWL (2002-SHERWL), the word list mainly serves as a reference for test development and "yet does not set limits to the English vocabulary to appear in the English tests of the college entrance examination" (CEEC, 2002), not to mention impinging on English instruction or the development of learning materials. However, because of the high stakes associated with college entrance exams, the CEEC word lists are highly valued by textbook publishers and English teachers and "widely acknowledged as the de facto list of to-be-learned English vocabulary" (Reynolds et al., 2018, p. 51). English teachers prioritize the learning of the vocabulary in the lists, whereas textbook publishers consult the CEEC's word lists to develop English learning materials for senior high school students.

2.2 English Tests on College Entrance Examinations in Taiwan

The CEEC offers two exams for students seeking admission to colleges in Taiwan: the General Scholastic Ability Test (GSAT) in January and the Advanced Subjects Test (AST) in July. At the time of this study, both the GSAT and the AST mandated English as a test subject. The English tests on the GSAT and the AST normally consist of two sections: a reading section (in the form of multiple-choice questions) that assesses students' knowledge of English vocabulary and ability to comprehend English passages, and a writing section that measures students' abilities to translate from Chinese to English and write short essays in English based on a given topic or a set of pictures. Except for the vocabulary questions, the reading section of the GSAT-English and AST-English involves texts adapted from various authentic sources, including newspapers, magazines, and books, in four types of questions: rational cloze, banked cloze, sentence gap filling, and reading comprehension; however, the GSAT-English normally does not contain sentence gap filling (CEEC, n.d.-a, -b).

The CEEC claims that the scope of the GSAT-English corresponds to the materials covered in the required English courses for the first two years of senior high school (roughly approximating to those covered in the first four volumes of school textbooks; CEEC, n.d.-a), whereas the scope of the AST-English corresponds to the materials covered in the three years of senior high school study (roughly approximating to those covered in all six

volumes of school textbooks; CEEC, n.d.-b). The GSAT-English requires a vocabulary of approximately 4,500 words (i.e., words in Levels 1-4 of the 2002-SHERWL; CEEC, n.d.-a), and the AST-English requires all six levels of the 2002-SHERWL (CEEC, n.d.-b). As a result, the AST-English is assumed to be more difficult than the GSAT-English.

The national curriculum guidelines do not specify content or specific vocabulary to be included in English courses or textbooks. Therefore, it is difficult to evaluate how the GSAT-English and AST-English tests correspond to course content or how well the course textbooks prepare senior high school students for the two high-stakes college entrance exams. One solution would be to scrutinize the linguistic features and difficulty of the texts in the English tests and school textbooks through corpus-based methods. Such an analysis could reveal the extent to which school textbooks prepare students for the texts on the tests.

2.3 Corpus-Based Analysis of Text Difficulty

Advances in computational linguistics and better accessibility to powerful, easy-to-use software tools have contributed to the growth of corpus-based research in the field of L2 learning and teaching.[2] These advancements enable L2 researchers and educators to analyze and compute various linguistic features and patterns of texts (spoken and written) more accurately and quickly with computers; the results of such research have significant implications for materials design, language testing, and classroom pedagogy. The development of corpus-based, automated measures of text difficulty is

[2] A distinction between corpus-based and corpus-driven language studies was first addressed by Tognini-Bonelli (2001), who favors corpus-driven studies. According to Biber (2010), corpus-based research "assumes the validity of linguistic forms and structures derived from linguistic theory" (p. 162) and usually aims to reveal the systematic patterns of variation and use for predefined linguistic features in a given corpus. Corpus-driven research "makes minimal a priori assumptions regarding the linguistic features that should be employed for the corpus analysis" (p. 162) and seeks to "identify new linguistic constructs through inductive analysis of corpora" (p. 169). According to Biber's definitions, the approach adopted in this study is corpus based because the goal of this study was not to discover new linguistic features or constructs but to uncover the use and patterns of predefined linguistic features in school textbooks and test papers. Nevertheless, not all corpus linguists accept the binary distinction between corpus-based and corpus-driven linguistics; some may consider all corpus linguistics to be corpus based (e.g., McEnery & Hardie, 2011).

an example of these advancements, which have empowered L2 researchers and educators to pursue various objectives, such as identifying differences among types of texts (e.g., Biber, 2006; Crossley et al., 2007; McNamara et al., 2012), assigning texts of an appropriate difficulty level to L2 learners (e.g., Chen & Meurers, 2019; Graesser et al., 2019; Lexile Framework for Reading, 2016; Sung et al., 2015), and validating the appropriacy of texts on reading tests in terms of difficulty (Green et al., 2010).

2.3.1 Lexical Coverage

Estimating lexical coverage (or text coverage) and text readability is a common corpus-based approach to evaluating text difficulty in research on L2 teaching and learning materials. Lexical coverage refers to "the percentage of the running words in a text known by the readers" (Nation, 2006, p. 1). For example, 95% lexical coverage means that readers know 95% of the words in a written text; of twenty words, one would be unknown. The ease of reading a text generally increases as the percentage of unknown words decreases (Carver, 1994). In lexical coverage research, the 95% and 98% coverage levels are commonly recognized as the threshold required to adequately comprehend a text (Schmitt et al., 2011). These figures are crucial because they enable estimation of the vocabulary size required for acceptable comprehension of specific texts, that is, the vocabulary load of texts. An analysis of vocabulary loads can then be performed to measure text difficulty (Webb & Nation, 2008). For example, by using fourteen 1,000 word family lists from the British National Corpus (BNC), Nation (2006) estimated that a vocabulary of 8,000 to 9,000 word families would be required to reach 98% coverage of (and thus to understand) such authentic written texts as newspapers and novels, whereas 3,000 word families would be sufficient to cover 98% of the running words in a simplified graded reader. The vocabulary load of a simplified graded reader is apparently lighter than that of a novel. The lexical coverage and vocabulary load of texts must be estimated against a word list. The word family lists of Nation (2012) were derived from the BNC and the Corpus of Contemporary America (COCA); they have been the most comprehensive word lists and thus commonly adopted in research on lexical coverage. To facilitate a comparison of research findings across studies, this study

adopted Nation's BNC/COCA word family lists to analyze lexical coverage. Because of the crucial role of the CEEC's word lists in senior high school English education in Taiwan, the 2000-SHERW was also used to calculate lexical coverage.

2.3.2 Text Readability

Text difficulty can be measured in terms of text readability. Text readability is traditionally measured unidimensionally by scaling text difficulty or ease on a single metric, such as the Flesch–Kincaid Grade Level (Klare, 1974), Degrees of Reading Power (Koslin et al., 1987), and Lexile scores (Stenner, 1996). McNamara et al. (2014) acknowledged the value of traditional unidimensional metrics, especially in terms of their simplicity and correspondence to grade level, but noted that these metrics "consider only the superficial characteristics of text" such as words and sentences, "tend to be predictive of readers' surface understanding" of the texts (p. 79), and do not account for multiple levels of comprehension, which would be "aligned with theories of text and discourse comprehension" (p. 84).

To address the limitations of traditional metrics of readability, Graesser and colleagues developed the readability index for L2 texts (RDL2; Crossley et al., 2008). RDL2 is unique in that it assesses text difficulty not only at the word and sentence levels but also in terms of cohesion between sentences, whereas the traditional readability measures only account for word- and sentence-level difficulty. The RDL2 value is calculated using the following formula (Crossley et al., 2008):

$$-45.032 + (52.23 \times \textit{content word overlap value}) + (61.306 \times \textit{sentence syntax similarity value}) + (22.205 \times \textit{CELEX word frequency value}).$$

The content word overlap index measures the extent to which content words overlap between two adjacent sentences. The more content words overlap (i.e., the higher the content word overlap index is), the higher the textual comprehension and reading speed are. The sentence syntax similarity index measures the uniformity and consistency of parallel syntactic constructions in texts. The higher the uniformity and consistency of parallel syntactic constructions are (i.e., the higher the sentence syntax similarity index is), the lighter the required cognitive burden is, and the

more a reader can concentrate on understanding the meaning. The CELEX word frequency index is based on frequency norms from the CELEX database (Baayen et al., 1993), a 17.9-million-word corpus. The higher the vocabulary frequency in the CELEX database is (i.e., the higher the CELEX word frequency value is), the faster the interpretation and processing of words is. Higher values for these three indices indicate that the text is easier to read. Consequently, a higher RDL2 would indicate greater ease for reading a text. Crossley et al. (2011) provided evidence of the advantages of RDL2 over two traditional readability formulas (the Flesch–Kincaid Grade Level and Flesch Reading Ease scores) in classifying levels of simplified readers for L2 learners.

In addition to developing RDL2, Graesser and colleagues proposed adopting a multidimensional approach to readability by scaling text difficulty or ease on the basis of a multilevel theoretical framework for language and discourse processing (Graesser & McNamara, 2011; McNamara et al., 2014). Graesser et al. (2011) generated eight text easability components based on 53 indices produced by Coh-Metrix, an automated computational tool that provides numerous metrics of text characteristics on multiple levels of language and discourse (including words, syntax, and discourse relationships between ideas). The eight components are (a) narrativity (PCNAR); (b) syntactic simplicity (PCSYN); (c) word concreteness (PCCNC); (d) referential cohesion (PCREF), which reflects the extent to which overlapping words and ideas are used across sentences and an entire text; (e) deep cohesion (PCDC), which is achieved by using explicit connectives to demonstrate the causal and logical relationships between ideas; (f) verb cohesion, which reveals the degree to which verbs are repeated in a text; (g) connectivity, which reflects the number of explicitly conveyed logical relations in a text; and (h) temporality. Studies have often used the first five components (PCNAR, PCSYN, PCCNC, PCREF, and PCDC) because "they are most directly associated with the ease of a text and because they account for a largest portion of the variance among the 37,520 texts" from which the eight components were generated (McNamara et al., 2014, p. 86).

Because unidimensional and multidimensional measures of readability both have merits, this study adopted both measures to examine the

readability of passages from the textbooks and college entrance exam papers. RDL2 was used as the unidimensional measure of readability. As for the multidimensional measure, this study only used PCNAR, PCSYN, PCCNC, PCREF, and PCDC. As recommended by the research team of Coh-Metrix, this study adopted the *z* scores for statistical analysis.

2.4 Research Questions

The main objective of this study was to determine how well senior high school English textbooks prepare students for reading the passages on the high-stakes college entrance exams in terms of text difficulty. The following two research questions were addressed: (a) Against the BNC/COCA frequency-based word lists and the CEEC Senior High School English Reference Word List (2002-SHERWL), to what extent does the lexical coverage of senior high school English textbooks correspond to that of GSAT-English and AST-English tests? (b) Does readability differ significantly between the textbook reading passages and those on the GSAT-English and AST-English tests?

3. Method

This study analyzed and compared the lexical coverage and readability metrics of the texts in senior high school textbooks and college entrance examination papers. The corpora and data analysis procedures are explained below.

3.1 The Corpora

Two corpora were compiled for this study: a textbook corpus, which consists of texts from the five editions of MOE–authorized senior high school textbooks in Taiwan (abbreviated FC, FS, LT, NI, and SM), and a test corpus, which consists of all reading passages from the English section of college entrance exams from 2002 to 2017. During this period, the textbooks and exams were developed following the same set of curriculum guidelines and with reference to the same edition of the CEEC word list, 2002-SHERWL.

The textbook corpus contains 2,384 texts, comprising the main reading passages, dialogue sections, and comprehension exercises from each lesson. Many of the comprehension exercises were in the form of gapped texts, for which readers fill in certain words and phrases. To create complete texts for the corpus analysis, the removed words and phrases were added back into the text during the text extraction process. All texts were included in the calculation of lexical coverage to estimate the amount of vocabulary students were likely exposed to either after studying the first four volumes of an edition of the textbooks for the GSAT-English tests or after reading all volumes (Volumes 1-6) of an edition for the AST-English tests. Only texts longer than 100 words were included in the Coh-Metrix analysis of readability (i.e., RDL2 and text easability principal component scores) because McNamara et al. (2014) cautioned that short texts (<100 words) could affect each Coh-Metrix index score. In addition, to address the second research question, only reading passages from the textbooks were included in the Coh-Metrix analysis to compare readability between passages from textbooks and those on the GSAT-English and AST-English tests.

The test corpus comprises all reading passages on the GSAT-English and AST-English tests administered from 2002 to 2017. A total of 16 regular GAST-English tests and 16 regular AST-English tests were collected. The reading passages on the tests were in the form of rational cloze, banked cloze, sentence gap filling, reading comprehension, and short answer. Because the rational cloze, banked cloze, and sentence gap filling questions corresponded to passages with gaps in the text, the missing words and phrases were filled during the text extraction process using the answers provided by the CEEC. A total of 260 texts were extracted. The same criteria for selecting texts from the textbook corpus for analysis were applied to selecting texts from the test corpus. All 260 test texts were included in the analysis of lexical coverage, but only texts longer than 100 words were included in the Coh-Metrix analysis, leading to the exclusion of one rational cloze text in GAST 2004. Tables 1 and 2 present the textbook and test corpora in the lexical coverage and Coh-Metrix analyses, respectively.

Table 1. Composition of Textbook and Test Corpora for Lexical Coverage Analysis

Textbook Corpus

Subcorpus / Edition	Volumes 1 to 4		Volumes 1 to 6	
	Token	Text Number	Token	Text Number
FC	57,997	363	89,549	483
FS	45,619	236	74,534	360
LT	46,119	326	70,803	465
NI	56,203	436	89,084	656
SM	49,992	305	79,239	420
Total	255,930	1,666	403,209	2,384

CEEC English Test Corpus

Subcorpus / Year	GSAT-English Tests		AST-English Tests	
	Token	Text Number	Token	Text Number
2002	1,603	7	2,011	10
2003	1,409	8	1,491	7
2004	1,631	8	1,533	8
2005	1,534	8	1,937	9
2006	1,660	8	1,823	9
2007	1,630	8	1,828	9
2008	1,607	8	1,787	9
2009	1,760	8	2,003	8
2010	1,873	8	1,947	8
2011	1,605	8	1,985	8
2012	1,800	8	2,101	8
2013	1,896	8	2,139	8

(continued)

Table 1. Composition of Textbook and Test Corpora for Lexical Coverage Analysis (continued)

CEEC English Test Corpus				
Subcorpus / Year	GSAT-English Tests		AST-English Tests	
	Token	Text Number	Token	Text Number
2014	1,883	8	1,984	8
2015	1,842	8	2,012	8
2016	1,926	8	2,237	8
2017	1,831	8	2,165	8
Total	27,490	127	30,983	133

Table 2. Composition of Textbook and Test Corpora for Coh-Metrix Analysis

Textbook Corpus				
Subcorpus / Edition	Volumes 1 to 4		Volumes 1 to 6	
	Token	Text Number	Token	Text Number
FC	18,315	47	30,934	67
FS	25,337	54	44,060	80
LT	20,828	49	33,093	69
NI	23,817	47	39,065	70
SM	22,156	48	35,901	69
Total	110,453	245	183,053	355

CEEC English Test Corpus				
Subcorpus / Year	GSAT-English Tests		AST-English Tests	
	Token	Text Number	Token	Text Number
2002	1,603	7	2,011	10
2003	1,409	8	1,491	7

(continued)

Table 2. Composition of Textbook and Test Corpora for Coh-Metrix Analysis (continued)

CEEC English Test Corpus				
Subcorpus / Year	GSAT-English Tests		AST-English Tests	
	Token	Text Number	Token	Text Number
2004	1,539	7	1,533	8
2005	1,534	8	1,937	9
2006	1,660	8	1,823	9
2007	1,630	8	1,828	9
2008	1,607	8	1,787	9
2009	1,760	8	2,003	8
2010	1,873	8	1,947	8
2011	1,605	8	1,985	8
2012	1,800	8	2,101	8
2013	1,896	8	2,139	8
2014	1,883	8	1,984	8
2015	1,842	8	2,012	8
2016	1,926	8	2,237	8
2017	1,831	8	2,165	8
Total	27,398	126	30,983	133

3.2 Data Analysis Procedure

The corpora were run through two computer programs, one for calculating the lexical coverage of the texts and the other for estimating text readability. The obtained values for the textbooks and entrance exam tests were compared descriptively or with inferential statistics to answer the research questions.

The program for computing lexical coverage was written using CLAWS tagger by a programmer working with the first author's colleague, a specialist in computational linguistics. The analysis of lexical coverage was performed on lemmatized texts against two sets of word lists: Paul Nation's 34 BNC/COCA word family lists and the CEEC's 2002-SHERWL. Proper nouns were excluded from the texts when lexical coverage was calculated because the 2002-SHERWL does not include proper nouns. Because 95% lexical coverage is generally regarded as the minimum threshold for adequate comprehension of texts (Hsu, 2011; Laufer, 1989; Laufer & Ravenhorst-Kalovski, 2010; Nation, 2001), it was set as the target coverage.

The BNC/COCA word family lists mainly consist of 25,000 word families, with every 1,000 word families constituting a sublist. In addition to the 25 sublists, Nation created another nine sublists to accommodate new words (Sublists 26-30), proper nouns (Sublist 31), marginal words (Sublist 32), transparent compounds (Sublist 33), and abbreviations (Sublist 34; Nation, 2012). All 34 word family lists were incorporated into the program to analyze lexical coverage. The 2002-SHERWL lists 6,480 lemmas grouped into six levels (sublists), with each level (sublist) containing 1,080 lemmas. The size of the 2002-SHERWL is actually larger than 6,480 lemmas because some words are hidden in the list as a result of the compilation principles specified in CEEC (2002). These hidden words largely consist of transparent derivatives of the listed lemmas (e.g., *fearless* vs. *fear*; *impossible* vs. *possible*). They are not clearly listed in the 2002-SHERWL because they are assumed to be easily recognized and likely learned in tandem with their associated lemmas because of the regularity and semantic transparency of the affixation process. To fully represent the lexical items covered in the 2002-SHERWL, we added these hidden words back into the list under their associated lemmas in accordance with the compilation principles before performing lexical coverage analysis. To answer the first research question, regarding the extent to which the lexical coverage of English textbooks corresponds to that of the English tests on the entrance exams, the two sources were compared descriptively.

To address the second research question, regarding differences in readability between the passages from the textbooks and college entrance exams, Coh-Metrix was employed. Both unidimensional (i.e., RDL2) and multidimensional metrics of readability (i.e., PCNAR, PCSYN, PCCNC,

PCREF, and PCDC) were adopted to evaluate text difficulty. Differences in readability between the two sources of texts were analyzed through an analysis of variance (ANOVA).

4. Results and Discussion

4.1 Lexical Coverage

4.1.1 Coverage of BNC/COCA Word Family Lists

To determine the extent to which the vocabulary size of the passages on the English tests match that of the textbooks, the lexical coverage of the passages from the two sources was calculated against the BNC/COCA word lists. Table 3 presents the results regarding the lexical coverage of the textbooks (the first four volumes of a set, the supposed scope of the GSAT-English test, vs. a complete set of six volumes, the supposed scope of the AST-English test).

The first 3,000 word families provide 95% coverage of the first four volumes of every edition of high school English textbooks (Table 3). By contrast, the vocabulary size required to reach 95% coverage of a complete set (six volumes) of textbooks varies by edition, with two editions (FS and NI) requiring 3,000 word families and three (FC, LT, and SM) requiring 4,000 word families.

Tables 4 and 5 display the results regarding the lexical coverage of the GSAT- and AST-English tests against the BNC/COCA word lists, respectively. The vocabulary size required for 95% coverage of the GSAT-English tests ranges from 3,000 to 5,000 word families (Table 4). The earlier tests tended to require a lower vocabulary load than did later tests. Specifically, tests prior to 2004 required a vocabulary size of 3,000 word families to reach 95% coverage. Most (six of the eight) of the tests from 2004 to 2011 required a vocabulary size of 4,000 word families to achieve acceptable comprehension, and most (four of the six) of the tests from 2012 to 2017 required a vocabulary size of 5,000 word families.

Table 3. Lexical Coverage of Textbooks Against BNC/COCA Word Lists

Edition	Accumulative Percentage			
	1,000	2,000	3,000	4,000
FC (Vol. 1-4)	85.49	92.57	95.02[a]	96.34
FS (Vol. 1-4)	87.27	93.29	95.42[a]	96.44
LT (Vol. 1-4)	85.97	92.57	95.21[a]	96.31
NI (Vol. 1-4)	86.32	93.47	96.29[a]	97.22
SM (Vol. 1-4)	85.93	92.20	95.00[a]	95.92
FC (Vol. 1-6)	84.47	91.76	94.69	96.10[a]
FS (Vol. 1-6)	86.67	92.94	95.40[a]	96.48
LT (Vol. 1-6)	85.44	92.03	94.92	96.13[a]
NI (Vol. 1-6)	85.07	92.64	95.98[a]	97.00
SM (Vol. 1-6)	85.11	91.70	94.82	95.94[a]

Note. [a] reaching 95% coverage.

Table 4. Lexical Coverage of GSAT-English Tests Against BNC/COCA Word Lists

Year	Accumulative Percentage				
	1,000	2,000	3,000	4,000	5,000
2002	83.78	92.76	96.32[a]	97.19	97.50
2003	78.28	89.85	95.88[a]	97.30	98.30
2004	78.54	88.53	93.81	95.65[a]	96.20
2005	80.31	89.57	94.07	95.83[a]	97.52
2006	81.45	92.11	95.06[a]	96.33	97.05
2007	82.15	90.98	94.72	95.71[a]	97.24
2008	80.58	89.55	94.59	95.52[a]	97.20
2009	78.92	88.64	93.41	94.66	95.45[a]

(continued)

Table 4. Lexical Coverage of GSAT-English Tests Against BNC/COCA Word Lists (continued)

Year	Accumulative Percentage				
	1,000	2,000	3,000	4,000	5,000
2010	78.16	89.38	93.65	95.30[a]	96.10
2011	79.25	89.35	93.89	95.33[a]	96.88
2012	79.39	88.61	92.78	94.67	95.83[a]
2013	78.59	88.82	93.51	94.73	95.99[a]
2014	79.71	88.95	94.32	95.27[a]	95.96
2015	75.84	88.55	93.16	94.57	95.11[a]
2016	78.76	88.79	95.02[a]	96.47	97.46
2017	77.12	87.98	93.39	94.70	95.25[a]

Note. [a] reaching 95% coverage.

Table 5. Lexical Coverage of AST-English Tests Against BNC/COCA Word Lists

Year	Accumulative Percentage					
	1,000	2,000	3,000	4,000	5,000	6,000
2002	80.46	88.66	92.89	94.88	95.23[a]	95.33
2003	78.07	89.60	95.24[a]	96.45	97.25	97.45
2004	73.91	85.84	92.50	94.78	96.09[a]	96.87
2005	76.20	87.20	93.96	95.92[a]	96.54	96.75
2006	77.62	89.08	93.75	95.50[a]	96.54	96.65
2007	76.53	87.36	93.82	96.23[a]	96.88	97.43
2008	76.44	87.74	94.18	95.24[a]	96.19	96.36
2009	78.33	89.12	94.46	95.61[a]	96.26	96.80
2010	75.91	86.08	92.71	94.35	95.07[a]	95.27

(continued)

Table 5. Lexical Coverage of AST-English Tests Against BNC/COCA Word Lists (continued)

Year	Accumulative Percentage					
	1,000	2,000	3,000	4,000	5,000	6,000
2011	73.80	85.29	92.75	95.77[a]	96.57	97.13
2012	74.44	86.01	94.67	95.95[a]	96.43	97.14
2013	78.17	88.83	93.60	94.48	95.56[a]	96.17
2014	75.76	86.49	92.49	93.75	95.67[a]	96.42
2015	72.12	84.05	90.41	92.64	94.68	95.33[a]
2016	76.62	87.39	93.38	95.84[a]	96.24	96.60
2017	72.42	84.48	90.85	92.98	94.04	95.43[a]

Note. [a] reaching 95% coverage.

The vocabulary size required to achieve 95% coverage on the AST-English tests fluctuates in the range of 3,000 to 6,000 word families but mostly falls between 4,000 and 5,000 word families (Table 5). Two later tests (from 2015 and 2017) even required 6,000 word families to reach 95% lexical coverage. On average, the vocabulary load of the AST-English tests is larger than that of GSAT-English tests. The results are consistent with the assumption that AST tests are more difficult than the GSAT tests, at least in terms of vocabulary load.

Figure 1 summarizes the results in Tables 3 to 5. With 95% lexical coverage as the minimum criterion, the vocabulary load (3,000 word families) for the first four volumes of the five textbook editions is compatible with only 4 of the 16 GSAT-English tests and 1 of the 16 AST-English tests. For the rest of the tests, a gap of 1,000 to 3,000 word families remained for school textbooks to fill. When a complete set of textbooks is considered, three of the five editions of the textbooks (FC, LT, and SM) exhibited an increase in vocabulary load to 4,000 word families, whereas two (FS and NI) remained the same (i.e., 3,000 word families). The vocabulary load of the complete sets of FC, LT, and SM is compatible with 9 of the 16 AST-English tests. As for the complete sets of FS and NI, the gap in required vocabulary remained even after two more volumes were added. The vocabulary they each offer is

comparable with only 1 of the 16 AST-English test papers. A substantial gap in vocabulary load was observed between the passages from the textbooks and those on the tests.

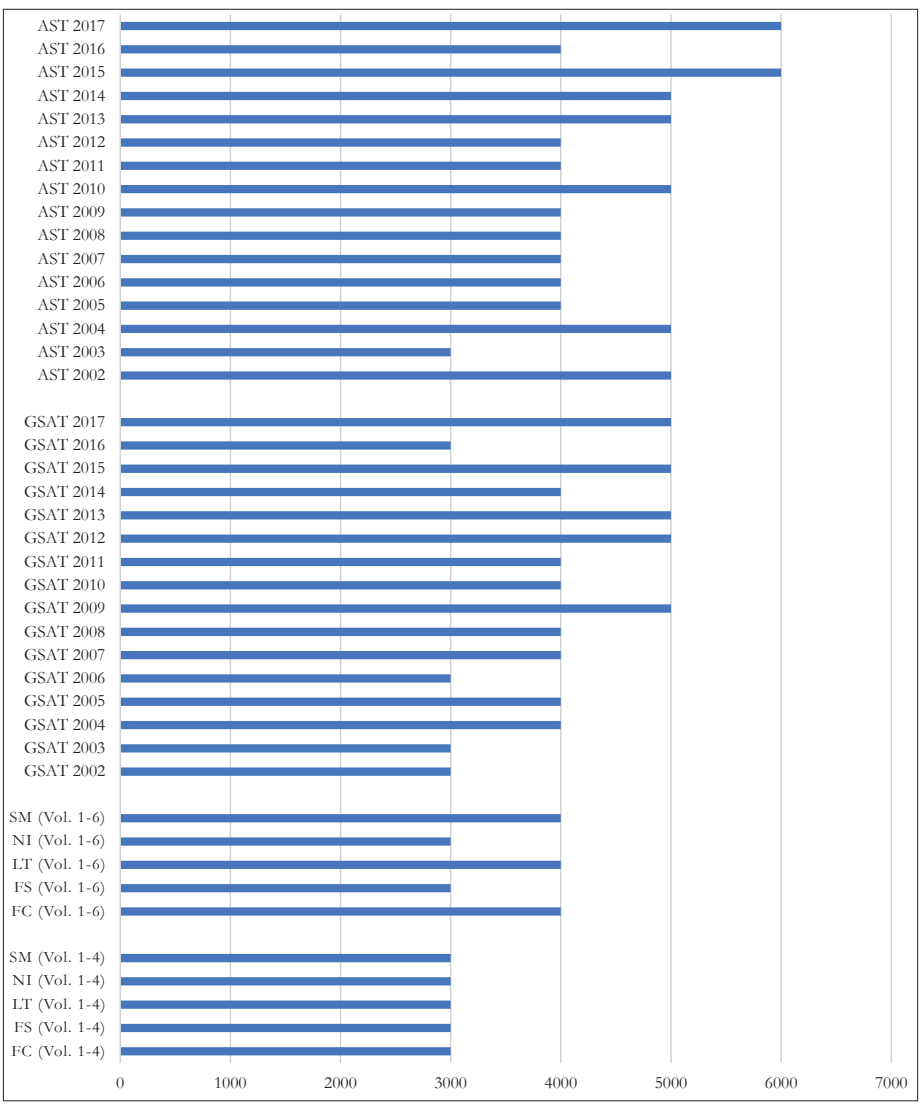

Figure 1. Number of BNC/COCA Word Families Required for 95% Lexical Coverage of Reading Passages in Textbooks and Tests

4.1.2 Coverage of 2002-SHERWL

The vocabulary load was compared between the textbooks and tests by examining the coverage of the 2002-SHERWL in the two sources of texts. Table 6 presents the results regarding the lexical coverage of the textbooks (a complete set vs. the first four volumes of a set). The first four levels of vocabulary on the 2002-SHERWL yielded 95% lexical coverage of all textbook subcorpora both for the complete set of textbooks (Volumes 1 to 6), which is assumed to correspond to the AST-English tests and for the first four volumes of an edition (Volumes 1 to 4), which are expected to match the GSAT-English tests (Table 6). The textbooks for the last year of senior high school did not markedly increase vocabulary size in terms of 2002-SHERWL coverage.

Table 6. Lexical Coverage of Textbooks Against CEEC's 2002-SHERWL

Edition	Accumulative Percentage			
	L1	L2	L3	L4
FC (Vol. 1-4)	82.44	90.37	94.10	95.93[a]
FS (Vol. 1-4)	84.14	91.52	94.76	96.35[a]
LT (Vol. 1-4)	82.78	90.39	93.91	96.08[a]
NI (Vol. 1-4)	82.83	90.95	94.70	96.91[a]
SM (Vol. 1-4)	83.26	90.76	94.22	96.40[a]
FC (Vol. 1-6)	81.09	89.10	93.06	95.33[a]
FS (Vol. 1-6)	83.14	90.67	94.14	96.10[a]
LT (Vol. 1-6)	82.29	89.68	93.22	95.66[a]
NI (Vol. 1-6)	81.42	89.74	93.79	96.32[a]
SM (Vol. 1-6)	82.27	89.69	93.41	95.93[a]

Note. L1 = Level 1; L2 = Level 2; L3 = Level 3; L4 = Level 4.

[a] reaching 95% coverage.

Tables 7 and 8 present the results regarding the lexical coverage of the GSAT-English and AST-English tests against the 2002-SHERWL. To reach 95% lexical coverage, most (13 of the 16) of the GSAT-English tests required at least five levels of vocabulary on the 2002-SHERWL and additional levels since 2015 (Table 7). Not even the entire 2002-SHERWL covered 95% of the vocabulary on the tests from 2015 and 2017, which is

Table 7. Lexical Coverage of GSAT-English Against CEEC's 2002-SHERWL

Year	Accumulative Percentage					
	L1	L2	L3	L4	L5	L6
2002	78.35	87.52	93.14	96.19[a]	96.69	97.50
2003	73.88	86.52	92.76	95.88[a]	97.09	97.73
2004	75.41	85.22	90.37	93.56	95.22[a]	96.75
2005	75.42	85.53	91.33	94.98	96.02[a]	96.74
2006	77.11	87.29	91.87	94.82	95.66[a]	96.27
2007	75.83	87.36	92.82	95.21[a]	96.20	96.99
2008	76.23	85.81	90.11	93.65	95.96[a]	96.45
2009	75.00	84.03	90.80	93.75	95.57[a]	96.53
2010	72.66	84.57	89.05	92.74	94.18	95.30[a]
2011	76.32	85.30	91.28	94.33	95.58[a]	96.64
2012	76.67	85.33	91.06	93.83	94.89	95.83[a]
2013	74.79	85.28	89.87	93.35	95.09[a]	95.99
2014	74.83	84.23	89.48	93.26	94.53	95.27[a]
2015	72.53	82.41	88.49	92.07	93.59	94.73[b]
2016	73.47	85.20	90.34	93.87	95.07[a]	96.31
2017	72.97	83.18	87.93	91.92	93.66	94.98[b]

Note. L1 = Level 1; L2 = Level 2; L3 = Level 3; L4 = Level 4; L5 = Level 5; L6 = Level 6.
[a] reaching 95% coverage; [b] not reaching 95% coverage.

Table 8. Lexical Coverage of AST-English Against CEEC's 2002-SHERWL

Year	Accumulative Percentage					
	L1	L2	L3	L4	L5	L6
2002	76.28	85.03	89.06	92.69	94.13	94.83[b]
2003	73.98	85.78	90.74	94.10	95.04[a]	96.45
2004	69.60	80.63	86.76	91.26	92.95	94.85[b]
2005	71.61	81.00	86.78	92.51	94.48	95.97[a]
2006	73.18	84.48	89.47	92.98	94.35	95.17[a]
2007	72.59	83.32	88.29	92.94	94.26	95.73[a]
2008	72.36	83.16	89.59	93.17	94.74	95.91[a]
2009	73.19	83.18	88.27	93.01	94.86	96.06[a]
2010	72.83	83.67	88.29	92.45	93.32	95.02[a]
2011	69.52	79.45	86.25	91.64	94.26	96.02[a]
2012	67.92	80.10	88.01	93.19	94.91	96.24[a]
2013	73.96	83.92	89.62	92.71	94.11	94.76[b]
2014	71.42	80.90	86.44	89.92	91.83	95.26[a]
2015	66.65	78.73	84.19	90.21	92.10	94.43[b]
2016	70.36	80.82	87.13	92.04	93.52	94.73[b]
2017	67.58	79.54	84.85	89.33	91.73	92.89[b]

Note. L1 = Level 1; L2 = Level 2; L3 = Level 3; L4 = Level 4; L5 = Level 5; L6 = Level 6.
[a] reaching 95% coverage; [b] not reaching 95% coverage.

the minimum requirement for text comprehension. The results cast doubt on the claim that the GSAT-English tests are based on the first four levels of vocabulary in the 2002-SHERWL (CEEC, n.d.-a) although the four levels covered more than 95% of the vocabulary in the textbooks.

The vocabulary demand of the AST-English tests is even greater than that of the GSAT-English (Table 8). Except for that of 2003, all AST-English

tests required all six levels of the 2002-SHERWL vocabulary or above to reach 95% lexical coverage and thus acceptable comprehension. The entire 2002-SHERWL does not provide the minimum (95%) lexical coverage for 6 of the 16 AST-English test papers. This result does not support the claim that AST-English tests mainly involve the six vocabulary levels of the 2002-SHERWL (CEEC, n.d.-b). Among the six most vocabulary-demanding tests, four were administered toward the end of the research period (2013, 2015, 2016, and 2017), indicating increasing difficulty of the tests in the lexical aspect over the years. These results are consistent with the results obtained using the BNC/COCA word family lists; the vocabulary load of the AST-English tests was larger than that of the GSAT-English tests.

Figure 2 summarizes the results of Tables 6 to 8. Except for three GSAT-English tests (those in 2002, 2003, and 2007), a gap of at least one vocabulary level was observed between the vocabulary demand of the textbooks and that of the English tests. Because each level of the 2002-SHERWL consists of 1,080 words, the vocabulary gap between the textbooks and English tests is approximately 1,000 to 2,000 words or more on the 2002-SHERWL. This finding is consistent with the results obtained using the BNC/COCA word family lists.

4.2 Text Readability

RDL2 and text easability principal component scores were used to evaluate the readability and difficulty of the reading passages from the GSAT- and AST-English tests and from the school textbooks. To determine the extent to which the texts from each source differed, ANOVAs were performed for the RDL2 and the five text easability principal component scores. When the *F* test revealed overall significance, Scheffe's test was used for post hoc pairwise comparisons. When the homogeneity of variances assumption was not satisfied, *Welch*'s test was performed during ANOVA, followed by a Dunnett T3 test for post hoc comparisons. To prevent potential inflated type I error rates due to multiple analyses on the same dependent variable, the Bonferroni corrected/adjusted *p* values were used to determine the significance of the differences. Reading passages from various years of the GSAT-English tests were combined in the analysis, as were the reading

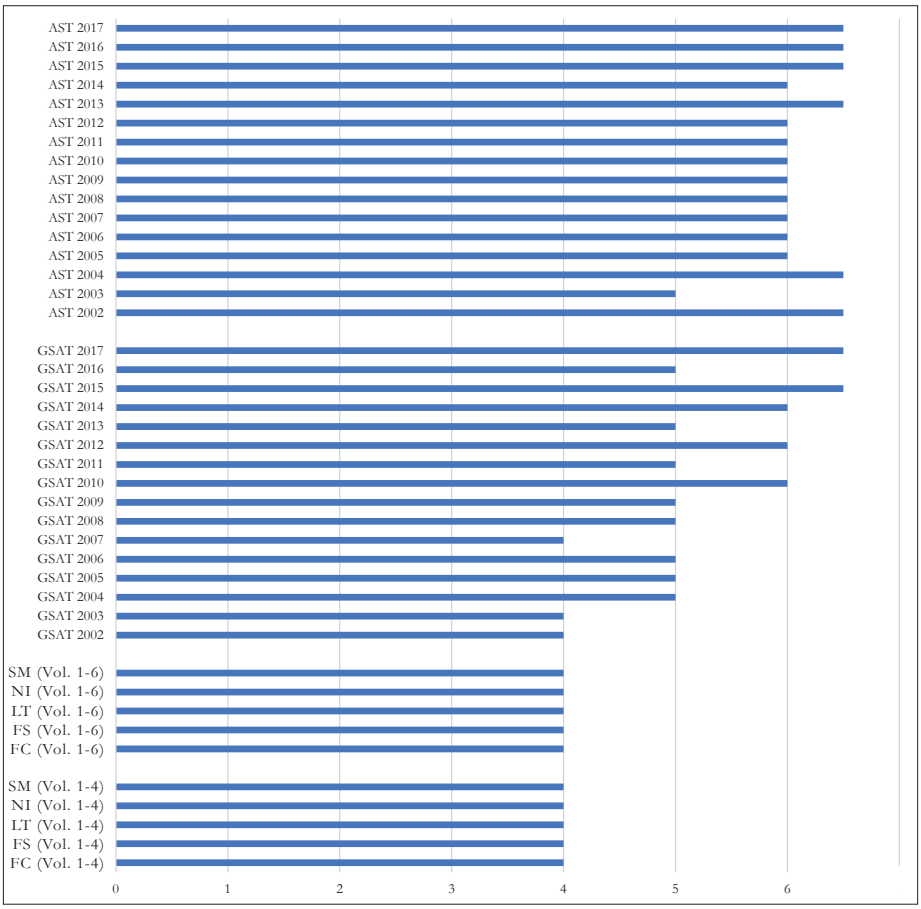

Figure 2. Levels of 2002-SHERWL Required for 95% Lexical Coverage of Reading Passages in Textbooks and Tests

passages from the AST-English tests, to ensure that the sample texts were sufficiently large to represent the GSAT- and AST-English tests and that parametric statistics could be used.

Table 9 presents the descriptive statistics of the RDL2 and text easability component scores for the reading passages from the English tests and textbooks (the first four volumes vs. the whole set/six volumes of each edition). With a few exceptions, the combined reading passages from Volumes 1 to 4 of each set of textbooks had higher overall readability (RDL2), narrativity, syntactic

Table 9. Descriptive Statistics of Readability Metrics for Reading Passages from Textbooks and Tests

Readability Metrics	Text Source	*N*	*M*	*SD*
RDL2	FC (Vol. 1-4)	47	19.12	3.62
	FC (Vol. 1-6)	67	18.10	4.36
	FS (Vol. 1-4)	54	20.40	4.91
	FS (Vol. 1-6)	80	19.23	5.42
	LT (Vol. 1-4)	49	17.97	4.07
	LT (Vol. 1-6)	69	17.21	4.50
	NI (Vol. 1-4)	47	20.87	5.47
	NI (Vol. 1-6)	70	19.33	5.72
	SM (Vol. 1-4)	48	17.94	4.13
	SM (Vol. 1-6)	69	17.05	4.42
	GSAT-E	126	16.22	4.80
	AST-E	133	13.73	5.11
PCNARz	FC (Vol. 1-4)	47	0.31	0.57
	FC (Vol. 1-6)	67	0.20	0.63
	FS (Vol. 1-4)	54	0.29	0.77
	FS (Vol. 1-6)	80	0.21	0.80
	LT (Vol. 1-4)	49	0.11	0.81
	LT (Vol. 1-6)	69	0.07	0.77
	NI (Vol. 1-4)	47	0.54	0.83
	NI (Vol. 1-6)	70	0.38	0.86
	SM (Vol. 1-4)	48	0.34	0.83
	SM (Vol. 1-6)	69	0.26	0.82
	GSAT-E	126	-0.28	0.77
	AST-E	133	-0.56	0.65

(continued)

Table 9. Descriptive Statistics of Readability Metrics for Reading Passages from Textbooks and Tests (continued)

Readability Metrics	Text Source	*N*	*M*	*SD*
PCSYNz	FC (Vol. 1-4)	47	0.71	0.45
	FC (Vol. 1-6)	67	0.48	0.60
	FS (Vol. 1-4)	54	0.49	0.70
	FS (Vol. 1-6)	80	0.37	0.69
	LT (Vol. 1-4)	49	0.36	0.52
	LT (Vol. 1-6)	69	0.26	0.54
	NI (Vol. 1-4)	47	0.56	0.63
	NI (Vol. 1-6)	70	0.44	0.60
	SM (Vol. 1-4)	48	0.27	0.44
	SM (Vol. 1-6)	69	0.09	0.56
	GSAT-E	126	0.02	0.54
	AST-E	133	-0.05	0.65
PCCNCz	FC (Vol. 1-4)	47	0.68	0.79
	FC (Vol. 1-6)	67	0.68	0.76
	FS (Vol. 1-4)	54	0.69	0.71
	FS (Vol. 1-6)	80	0.63	0.71
	LT (Vol. 1-4)	49	0.62	0.80
	LT (Vol. 1-6)	69	0.56	0.77
	NI (Vol. 1-4)	47	0.33	0.89
	NI (Vol. 1-6)	70	0.31	0.81
	SM (Vol. 1-4)	48	0.49	0.79
	SM (Vol. 1-6)	69	0.49	0.78
	GSAT-E	126	0.80	0.90
	AST-E	133	0.69	0.91

(continued)

Table 9. Descriptive Statistics of Readability Metrics for Reading Passages from Textbooks and Tests (continued)

Readability Metrics	Text Source	*N*	*M*	*SD*
PCREFz	FC (Vol. 1-4)	47	-0.36	0.66
	FC (Vol. 1-6)	67	-0.35	0.69
	FS (Vol. 1-4)	54	-0.30	0.64
	FS (Vol. 1-6)	80	-0.41	0.80
	LT (Vol. 1-4)	49	-0.57	0.63
	LT (Vol. 1-6)	69	-0.58	0.67
	NI (Vol. 1-4)	47	-0.23	0.75
	NI (Vol. 1-6)	70	-0.41	0.77
	SM (Vol. 1-4)	48	-0.29	0.81
	SM (Vol. 1-6)	69	-0.39	0.77
	GSAT-E	126	-0.26	0.91
	AST-E	133	-0.56	0.74
PCDCz	FC (Vol. 1-4)	47	0.50	0.90
	FC (Vol. 1-6)	67	0.53	0.83
	FS (Vol. 1-4)	54	0.37	0.77
	FS (Vol. 1-6)	80	0.35	0.71
	LT (Vol. 1-4)	49	0.52	0.75
	LT (Vol. 1-6)	69	0.54	0.72
	NI (Vol. 1-4)	47	0.76	0.79
	NI (Vol. 1-6)	70	0.73	0.71
	SM (Vol. 1-4)	48	0.96	0.93
	SM (Vol. 1-6)	69	0.86	0.91
	GSAT-E	126	0.71	1.18
	AST-E	133	0.57	1.12

Note. FC, FS, LT, NI, SM = the five sets of senior high school textbooks; GSAT-E = GSAT-English test; AST-E = AST-English test; RDL2 = L2 readability; PCNARz = narrativity score; PCSYNz = syntactic simplicity; PCCNCz = word concreteness; PCREFz = referential cohesion; PCDCz = deep cohesion.

simplicity, word concreteness, referential cohesion, and deep cohesion on average than did the combined reading passages from Volumes 1 to 6 (Table 9). The results suggest that the reading passages from the last two volumes (Volumes 5 and 6) of each textbook series contributed to a higher text difficulty in the entire set of textbooks.

Because the CEEC claims that the GSAT-English assessment largely corresponds to what students learn during their first two years of senior high school, ANOVAs were performed to determine whether there were significant differences in text difficulty between the reading passages on the GSAT-English tests and those from Volumes 1 to 4 of each textbook series written for students in the first two years of senior high school. The results reveal an overall significant difference in RDL2, $F(5, 365) = 10.728$, $p < 0.001$; PCNARz, $F(5, 365) = 11.364$, $p < .001$; PCSYNz, $F(5, 365) = 15.205$, $p < .001$; PCCNCz, $F(5, 365) = 2.656$, $p = .022$; and PCDCz, *Welch*$(5, 147.493) = 3.197$, $p = .009$; between the two sources of texts. No significant difference was observed for PCREFz, *Welch*$(5, 146.172) = 1.791$, $p = .118$; the dimension of referential cohesion. Post hoc comparisons were performed for RDL2, PCNARz, PCSYNz, PCCNCz, and PCDCz between the texts of GSAT-English tests and those in the textbook series (Volumes 1 to 4); Table 10 presents the results.

Table 10. Pairwise Comparisons of Readability Between GSAT-English and First Four Volumes of Textbooks

	Text Source		*MD* (I-J)	*SE*	*p*	95% Confidence Interval	
	I	J				LB	UB
RDL2 (Scheffe)	GSAT-E	FC (Vol. 1-4)	-2.90	0.68	.001	-4.93	-0.87
		FS (Vol. 1-4)	-4.18	0.79	< .001	-6.56	-1.80
		LT (Vol. 1-4)	-1.75	0.72	.223	-3.91	0.41
		NI (Vol. 1-4)	-4.65	0.91	< .001	-7.39	-1.92
		SM (Vol. 1-4)	-1.72	0.73	.264	-3.92	0.48

(continued)

Table 10. Pairwise Comparisons of Readability Between GSAT-English and First Four Volumes of Textbooks (continued)

	Text Source		*MD* (I-J)	*SE*	*p*	95% Confidence Interval	
	I	J				LB	UB
PCNARz (Scheffe)	GSAT-E	FC (Vol. 1-4)	-0.59	0.13	.002	-1.03	-0.15
		FS (Vol. 1-4)	-0.57	0.13	.001	-0.99	-0.15
		LT (Vol. 1-4)	-0.39	0.13	.104	-0.83	0.04
		NI (Vol. 1-4)	-0.82	0.13	< .001	-1.26	-0.38
		SM (Vol. 1-4)	-0.62	0.13	.001	-1.06	-0.18
PCSYNz (Scheffe)	GSAT-E	FC (Vol. 1-4)	-0.69	0.09	< .001	-1.01	-0.38
		FS (Vol. 1-4)	-0.48	0.09	< .001	-0.78	-0.18
		LT (Vol. 1-4)	-0.34	0.09	.022	-0.65	-0.03
		NI (Vol. 1-4)	-0.54	0.09	< .001	-0.86	-0.22
		SM (Vol. 1-4)	-0.25	0.09	.201	-0.57	0.06
PCCNCz (Scheffe)	GSAT-E	FC (Vol. 1-4)	0.13	0.14	.978	-0.35	0.60
		FS (Vol. 1-4)	0.12	0.14	.980	-0.34	0.57
		LT (Vol. 1-4)	0.19	0.14	.881	-0.28	0.66
		NI (Vol. 1-4)	0.48	0.14	.049	0.00	0.95
		SM (Vol. 1-4)	0.31	0.14	.441	-0.16	0.78
PCDCz (Dunnett T3)	GSAT-E	FC (Vol. 1-4)	0.21	0.17	.971	-0.29	0.71
		FS (Vol. 1-4)	0.34	0.15	.276	-0.10	0.78
		LT (Vol. 1-4)	0.19	0.15	.970	-0.26	0.63
		NI (Vol. 1-4)	-0.05	0.16	1.000	-0.52	0.41
		SM (Vol. 1-4)	-0.25	0.17	.888	-0.76	0.26

Note. The significance cutoff was set to .05/5 = .01 on the basis of the Bonferroni correction. FC, FS, LT, NI, SM = the five sets of senior high school textbooks; GSAT-E = GSAT-English test; RDL2 = L2 readability; PCNARz = narrativity score; PCSYNz = syntactic simplicity; PCCNCz = word concreteness; PCDCz = deep cohesion.

Despite the overall significant *F* statistic for PCDCz (deep cohesion) and PCCNCz (word concreteness), no significant difference was identified in the pairwise comparison (Table 10).[3] Together with the nonsignificant *F*-test result for PCREFz (referential cohesion), the results reveal that the texts from the GSAT-English tests and all textbook series (Volumes 1 to 4) are similar in terms of text cohesion and word concreteness. Nevertheless, the GSAT-English had significantly lower level of narrativity (PCNARz) than did all but one set of textbooks (i.e., LT). The syntactic simplicity (PCSYNz) and overall readability (RDL2) of the English tests were also significantly lower than those of three of the textbook series: FC, FS, and NI. To conclude, the reading passages on the GSAT-English tests were more difficult than those in the first four volumes of FC, FS, and NI in terms of the overall readability metric, narrativity, and syntactic simplicity, although they were similar in word concreteness, referential cohesion, and deep cohesion. SM differed from the tests in only one dimension, narrativity (PCNARz), whereas LT exhibited no significant differences in any aspect. The reading passages in LT and SM were similar to those on the GSAT-English tests in terms of readability and difficulty.

To determine whether the text difficulty of an entire set of textbooks (Volumes 1 to 6) corresponded to those of the AST-English tests, ANOVAs were performed to compare the English tests with the five sets of textbooks. The analysis revealed an overall significant difference in RDL2, $F(5, 482) = 18.37$, $p < .001$; PCNARz, $Welch(5, 203.388) = 25.76$, $p < .001$; PCSYNz, $F(5, 482) = 10.53$, $p < .001$; PCCNCz, $F(5, 482) = 2.45$, $p = .033$; and PCDCz, $Welch(5, 211.473) = 3.73$, $p = .003$; but not in PCREFz, $F(5, 482) = 1.38$, $p = .232$. Post hoc pairwise comparisons were thus performed for RDL2, PCNARz, PCSYNz, PCCNCz, and PCDCz; Table 11 summarizes the results.

As shown in Table 11, in terms of easability, the results for the whole sets of textbooks (vs. the AST-English tests) are similar to those for the first four volumes of textbooks (vs. the GSAT-English tests). Despite an overall significant *F* for PCDCz (deep cohesion), the results of the pairwise comparisons were

[3] Nonsignificance in post hoc comparisons after a significant overall *F* is normal because "the hypotheses tested by the overall test and a multiple-comparison test are quite different, with quite different levels of power" (Howell, 2009, p. 366).

Table 11. Pairwise Comparisons of Readability Between AST-English Tests and Complete Sets of Textbooks

	Text Source		*MD* (I-J)	*SE*	*p*	95% Confidence Interval	
	I	J				LB	UB
RDL2	AST-E (Scheffe)	FC (Vol. 1-6)	-4.38	0.75	< .001	-6.87	-1.88
		FS (Vol. 1-6)	-5.50	0.71	< .001	-7.86	-3.14
		LT (Vol. 1-6)	-3.48	0.74	.001	-5.96	-1.01
		NI (Vol. 1-6)	-5.60	0.74	< .001	-8.06	-3.14
		SM (Vol. 1-6)	-3.33	0.74	.001	-5.80	-0.86
PCNARz	AST-E (Dunnett T3)	FC (Vol. 1-6)	-0.77	0.10	< .001	-1.05	-0.48
		FS (Vol. 1-6)	-0.78	0.11	< .001	-1.09	-0.46
		LT (Vol. 1-6)	-0.63	0.11	< .001	-0.96	-0.31
		NI (Vol. 1-6)	-0.94	0.12	< .001	-1.29	-0.59
		SM (Vol. 1-6)	-0.82	0.11	< .001	-1.16	-0.48
PCSYNz	AST-E (Scheffe)	FC (Vol. 1-6)	-0.52	0.09	< .001	-0.83	-0.21
		FS (Vol. 1-6)	-0.41	0.09	.001	-0.71	-0.12
		LT (Vol. 1-6)	-0.31	0.09	.050	-0.61	0.00
		NI (Vol. 1-6)	-0.48	0.09	< .001	-0.79	-0.18
		SM (Vol. 1-6)	-0.14	0.09	.820	-0.44	0.17
PCCNCz	AST-E (Scheffe)	FC (Vol. 1-6)	0.01	0.12	1.000	-0.40	0.41
		FS (Vol. 1-6)	0.06	0.11	0.998	-0.32	0.44
		LT (Vol. 1-6)	0.13	0.12	0.951	-0.27	0.53
		NI (Vol. 1-6)	0.38	0.12	0.079	-0.02	0.77
		SM (Vol. 1-6)	0.20	0.12	0.752	-0.20	0.60

(continued)

Table 11. Pairwise Comparisons of Readability Between AST-English Tests and Complete Sets of Textbooks (continued)

	Text Source		*MD* (I-J)	*SE*	*p*	95% Confidence Interval	
	I	J				LB	UB
PCDCz	AST-E (Dunnett T3)	FC (Vol. 1-6)	0.04	0.14	1.000	-0.38	0.45
		FS (Vol. 1-6)	0.22	0.13	0.710	-0.15	0.59
		LT (Vol. 1-6)	0.02	0.13	1.000	-0.36	0.41
		NI (Vol. 1-6)	-0.16	0.13	0.967	-0.55	0.22
		SM (Vol. 1-6)	-0.30	0.15	0.472	-0.73	0.14

Note. The significance cutoff was set to .05/5 = .01 on the basis of the Bonferroni correction. FC, FS, LT, NI, SM = the five sets of senior high school textbooks; GSAT-E = GSAT-English test; AST-E = AST-English test; RDL2 = L2 readability; PCNARz = narrativity score; PCSYNz = syntactic simplicity; PCCNCz = word concreteness; PCDCz = deep cohesion.

nonsignificant. The nonsignificant pairwise comparisons for PCDCz (deep cohesion) and the nonsignificant *F*-test results for PCREFz (referential cohesion) indicate that the reading passages from the English tests and textbooks (Volumes 1 to 6) were similar in terms of cohesion. This is also true of PCCNCz (word concreteness); the English tests and textbooks were similar in terms of word concreteness. Moreover, the AST-English tests were more syntactically complex than three textbook series (FS, FC, and NI), corresponding to the results for the GSAT-English tests.

Somewhat different results were observed for PCNARz (narrativity) and overall readability (RDL2). When only four volumes of textbooks were included in the analysis, the GSAT-English tests exhibited significantly lower narrativity than did four textbook series (Volumes 1 to 4 of FC, FS, NI, and SM), and they were less readable than three textbook series (Volumes 1 to 4 of FS, FC, and NI). The difference between the AST-English tests and textbook series was larger in these two aspects; the AST-English tests had lower narrativity and overall readability than did all five complete sets textbooks (Volumes 1 to 6). In brief, the reading passages on the AST-

English tests were more difficult and had higher syntactic complexity, lower narrativity, and lower overall readability than did the school textbooks.

5. Conclusions, Implications, and Limitations

This study aimed to reveal how sufficiently senior high school English textbooks in Taiwan prepare students for English passages on high-stakes college entrance exams in terms of text difficulty. A corpus-based approach was adopted to compare the vocabulary load and readability of passages from senior high school textbooks with those of the English sections of college entrance exams. The results indicate that the passages from the English textbooks do not correspond to those on the tests in terms of vocabulary load and several Coh-Metrix readability/difficulty metrics. The passages on the English tests generally have lower overall readability (RDL2), less narrativity, and higher syntactic complexity than do those in the textbooks. The passages from the English tests also demand a larger vocabulary size than do the textbooks; this gap in vocabulary load between the textbooks texts and CEEC English tests has been widening in recent years.

The findings have several implications. To students, the findings suggest sole reliance on the input provided by textbooks does not prepare them well for reading passages on the CEEC English tests. To comprehend these passages, students must extensively read texts of various genres and difficulty levels, including those with low narrativity and high syntactic complexity. Extensive reading can also increase students' exposure to certain words' usage in different meaningful contexts, thereby facilitating vocabulary acquisition and vocabulary expansion. This can also enhance students' readings skills, which will help them master materials above their competence levels. To English teachers, the findings suggest that they should supplement textbooks with online or offline learning materials and resources to help students bridge the gap between school textbooks and the CEEC English tests in terms of vocabulary demand and readability. For textbook writers, the findings indicate a need to adjust the difficulty level and vocabulary load of textbooks to match those of the CEEC English tests. Alternatively, with the wide disparity in English competence among students

in each class, textbook writers may consider adding difficult passages (e.g., with a higher vocabulary load, lower narrativity, and/or higher syntactic complexity) as supplementary materials for students ready or motivated to read challenging texts. Finally, the Ministry of Education should consider the extent to which authorized textbooks should correspond to the CEEC English tests in terms of text difficulty and vocabulary load and provide publishers and the CEEC with guidelines for developing textbook materials and tests.

Certain limitations of this study should be noted. First, this study analyzed the vocabulary load and readability of school textbooks only. Although textbooks are the main source of input for foreign language learners, they may not be the sole input. In addition, although input is a prerequisite of L2 development, learners do not necessarily internalize input into their language knowledge. Several other factors (e.g., learner factors, contextual factors, and instructional factors) beyond the scope of this study play a role in determining learning outcomes. Hence, caution should be exercised in interpreting the lexical coverage of the textbooks in this study as the vocabulary size students acquire during high school. Second, the learning materials analyzed in this study only comprised textbooks. The results may differ if other learning materials are included in the analysis, such as supplementary reading passages in teachers' manuals and reference books accompanying textbooks; future corpus-based studies can consider including these learning materials. Third, the comparisons made in this study were limited to lexical coverage and text readability. Comparing other linguistic features (e.g., lexical diversity, n-grams/multiword expressions, and high-frequency vocabulary coverage) between school textbooks and CEEC tests would be a compelling direction. Finally, this study compared the textbooks and CEEC English tests written in accordance with the curriculum guidelines implemented prior to 2019, when new curriculum guidelines were implemented and a revised CEEC reference word list was released. The gap between these two sources of texts should not be generalized to the textbooks and CEEC English tests from other periods. Researchers are advised to examine whether the gap is smaller between the textbooks and English tests developed in accordance with the new curriculum guidelines.

References

Baayen, R. H., Piepenbrock, R., & van Rijn, H. (Eds.). (1993). *The CELEX lexical database* (CD-ROM). Linguistic Data Consortium.

Biber, D. (2006). *University language: A corpus-based study of spoken and written discourse*. John Benjamin.

Biber, D. (2010). Corpus-based and corpus-driven analyses of language variation and use. In B. Heine & H. Narrog (Eds.), *The Oxford handbook of linguistic analysis* (pp. 153-191). Oxford University Press.

Carver, R. P. (1994). Percentage of unknown vocabulary words in text as a function of the relative difficulty of the text: Implications for instruction. *Journal of Literacy Research*, *26*(4), 413-437. https://doi.org/10.1080/10862969409547861

Chen, H.-L. S., & Huang, H.-Y. (2017). *Advancing 21st century competencies in Taiwan*. National Taiwan Normal University.

Chen, X., & Meurers, D. (2019). Linking text readability and learner proficiency using linguistic complexity feature vector distance. *Computer Assisted Language Learning*, *32*, 418-447. https://doi.org/10.1080/09588221.2018.1527358

College Entrance Examination Center. (n.d.-a). *GSAT-English*. https://www.ceec.edu.tw/en/xmdoc/cont?xsmsid=0J180519944235388511

College Entrance Examination Center. (n.d.-b). *AST-English*. https://www.ceec.edu.tw/en/xmdoc/cont?xsmsid=0J180520414679660023

College Entrance Examination Center. (2002). *Introduction to reference word list for senior high English education*. https://www.ceec.edu.tw/SourceUse/ce37/3.pdf

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

Crossley, S. A., Louwerse, M. M., McCarthy, P. M., & McNamara, D. S. (2007). A linguistic analysis of simplified and authentic texts. *The Modern Language Journal*, *91*(1), 15-30. https://doi.org/10.1111/j.1540-4781.2007.00507.x

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

Ellis, R. (2005). Principles of instructed language learning. *System*, *33*(2), 209-224. https://doi.org/10.1016/j.system.2004.12.006

Frantzen, D. (2010). Incremental gains in foreign language programs: The role of reading in learning about other cultures. *Reading in a Foreign Language*, *22*(suppl. 1), 31-37.

Graesser, A. C., Greenberg, D., Olney, A., & Lovett, M. W. (2019). Educational technologies that support reading comprehension for adults who have low literacy skills. In D. Perin

(Ed.), *The Wiley handbook of adult literacy* (pp. 471-493). John Wiley & Sons.

Graesser, A. C., & McNamara, D. S. (2011). Computational analyses of multilevel discourse comprehension. *Topics in Cognitive Science*, *3*(2), 371-398. https://doi.org/10.1111/j.1756-8765.2010.01081.x

Graesser, A. C., McNamara, D. S., & Kulikowich, J. M. (2011). Coh-Metrix: Providing multilevel analyses of text characteristics. *Educational Researcher*, *40*(5), 223-234. https://doi.org/10.3102/0013189X11413260

Green, A., Ünaldi, A., & Weir, C. (2010). Empiricism versus connoisseurship: Establishing the appropriacy of texts in tests of academic reading. *Language Testing*, *27*(2), 191-211. https://doi.org/10.1177/0265532209349471

Hill, P. (2010). *Asia-Pacific secondary education system review series No. 1: Examination systems*. United Nations Educational, Scientific and Cultural Organization.

Howell, D. C. (2009). *Statistical methods for psychology* (7th ed.). Cengage Learning.

Hsu, W. (2011). The vocabulary thresholds of business textbooks and business research articles for EFL learners. *English for Specific Purposes*, *30*(4), 247-257. https://doi.org/10.1016/j.esp.2011.04.005

Klare, G. R. (1974). Assessing readability. *Reading Research Quarterly*, *10*(1), 62-102. https://doi.org/10.2307/747086

Koslin, B. I., Zeno, S., & Koslin, S. (1987). *The DRP: An effective measure in reading*. College Entrance Examination Board.

Laufer, B. (1989). What percentage of text lexis is essential for comprehension? In C. Lauren & M. Nordman (Eds.), *Special language: From human thinking to thinking machines* (pp. 316-323). Multilingual Matters.

Laufer, B., & Ravenhorst-Kalovski, G. C. (2010). Lexical threshold revisited: Lexical text coverage, learners' vocabulary size and reading comprehension. *Reading in a Foreign Language*, *22*(1), 15-30.

Lexile Framework for Reading. (2016). *Lexile analyzer*. https://www.lexile.com/tools/lexile-analyzer/step-3-type-or-scan-your-text/

Lin, M.-C. (2018). From skill to competence of English language teaching: The contextualized communicative approach. *Journal of Education Research*, *294*, 72-90. https://doi.org/10.3966/168063602018100294005

McEnery, T., & Hardie, A. (2011). *Corpus linguistics: Method, theory and practice*. Cambridge University Press.

McNamara, D. S., Graesser, A. C., & Louwerse, M. M. (2012). Sources of text difficulty: Across genres and grades. In J. P. Sabatini, E. Albro, & T. O'Reilly (Eds.), *Measuring up: Advances in how we assess reading ability* (pp. 89-116). Rowman & Littlefield Education.

McNamara, D. S., Graesser, A. C., McCarthy, P. M., & Cai, Z. (2014). *Automated evaluation*

of text and discourse with Coh-Metrix. Cambridge University Press.

Nation, I. S. P. (2001). *Learning vocabulary in another language*. Cambridge University Press.

Nation, I. S. P. (2006). How large a vocabulary is needed for reading and listening? *The Canadian Modern Language Review*, *63*(1), 59-82. https://doi.org/10.3138/cmlr.63.1.59

Nation, I. S. P. (2012). *The BNC/COCA word family lists*. http://www.victoria.ac.nz/lals/about/staff/publications/BNC_COCA_25000.zip

Reynolds, B. L., Shih, Y. C., & Wu, W. H. (2018). Modeling Taiwanese adolescent learners' English vocabulary acquisition and retention: The washback effect of the College Entrance Examination Center's reference word list. *English for Specific Purposes*, *52*, 47-59. https://doi.org/10.1016/j.esp.2018.08.001

Schmitt, N., Jiang, X., & Grabe, W. (2011). The percentage of words known in a text and reading comprehension. *The Modern Language Journal*, *95*(1), 26-43. https://doi.org/10.1111/j.1540-4781.2011.01146.x

Stenner, A. J. (1996). *Measuring reading comprehension with the Lexile framework*. https://files.eric.ed.gov/fulltext/ED435977.pdf

Sung, Y.-T., Lin, W.-C., Dyson, S. B., Chang, K.-E., & Chen, Y.-C. (2015). Leveling L2 texts through readability: Combining multilevel linguistic features with the CEFR. *The Modern Language Journal*, *99*(2), 371-391. https://doi.org/10.1111/modl.12213

Tognini-Bonelli, E. (2001). *Corpus linguistics at work*. John Benjamins.

Wang, C.-C. (2008). Communicative language teaching in Taiwan: Teacher conceptions and major challenges. *Journal of Applied English*, *1*, 31-58. https://doi.org/10.29691/JAE.200810.0003

Webb, S., & Nation, I. S. P. (2008). Evaluating the vocabulary load of written text. *TESOLANZ Journal*, *16*, 1-10.

Zhang, X. (2017). A critical review of literature on English language teaching textbook evaluation: What systemic functional linguistics can offer. *Journal of Language and Cultural Education*, *5*(1), 78-102. https://doi.org/10.1515/jolace-2017-0005

Zyzik, E. (2009). The role of input revisited: Nativist versus usage-based models. *L2 Journal*, *1*(1), 42-61. https://doi.org/10.5070/l2.v1i1.9056