Research Article

Yanfang Su, Kanglong Liu, Fengkai Liu, John Lee and Tan Jin*

Lexical complexity in exemplar EFL texts: towards text adaptation for 12 grades of basic English curriculum in China

https://doi.org/10.1515/iral-2022-0236
Received December 11, 2022; accepted July 30, 2023; published online August 14, 2023

Abstract: Lexical complexity has been a key consideration of teaching preparation in determining grade appropriateness of teaching materials. However, the lack of quantified and defined standards for benchmarking lexical complexity has made it difficult for teachers when adapting source texts to target learners. This study has assessed quantitative differences in lexical complexity of exemplar texts at different points of schooling using a range of lexical diversity and sophistication features. The data consists of 2,372 texts from popular curriculum packages adopted from 1 to 12 grades of the English curriculum in China. One-way ANOVAs revealed significant differences in 16 out of 17 lexical complexity indices among different grades. Subsequent post hoc tests identified three lexical diversity features and four sophistication features that helped to differentiate exemplar texts across these 12 grades. These findings on the nature and role of lexical complexity have yielded new insights into the establishment of grade-level benchmarks for material preparation.

Keywords: exemplar texts; lexical complexity; lexical diversity; lexical sophistication; material preparation

***Corresponding author: Tan Jin**, School of Foreign Studies, South China Normal University, No. 55, West of Zhongshan Avenue, Tianhe District, Guangzhou, 510631, China, E-mail: alextanjin@gmail.com. https://orcid.org/0000-0002-6421-0977

Yanfang Su and Kanglong Liu, The Hong Kong Polytechnic University, Hong Kong, China, E-mail: yanfang.su@connect.polyu.hk (Y. Su), kl.liu@polyu.edu.hk (K. Liu). https://orcid.org/0000-0002-4526-4187 (Y. Su). https://orcid.org/0000-0003-3962-2563 (K. Liu)

Fengkai Liu and John Lee, City University of Hong Kong, Hong Kong, China, E-mail: fengkaliu3-c@my.cityu.edu.hk (F. Liu), jsylee@cityu.edu.hk (J. Lee). https://orcid.org/0000-0002-0422-3362 (F. Liu). https://orcid.org/0000-0003-2505-2678 (J. Lee)

1 Introduction

It has been widely recognized that teachers should take on the role of the material analyst and text adaptor to ensure effective language teaching (Crossley et al. 2012b; Li et al. 2021). The primary concern of adaptation involves adjusting linguistic complexity, including syntactic and lexical complexity (Berendes et al. 2018; Hiebert et al. 2018). It has been found that lexical complexity significantly influences the difficulty of comprehending reading materials (Rayner and Duffy 1986), with its impact considered greater than that of syntactic complexity (Arya et al. 2011; Droop and Verhoeven 1998). Consequently, selecting and preparing teaching materials necessitates careful consideration of lexical complexity in order to effectively cater to students' needs. In other words, the ability of teachers to choose materials appropriate for students' lexical complexity level is crucial for successful language learning (Chen and Meurers 2018).

Often teachers adopt an intuitive approach to adapting source materials by relying on past experiences or knowledge (Allen 2009; Simensen 1987). Nevertheless, even experienced language teachers or specialists may make inconsistent or incorrect judgments about the compatibility of text linguistic complexity and student proficiency levels (Carabantes and Paran 2022; Green and Rogar 2012). Recently, researchers have utilized computer techniques that include a range of linguistic complexity indices to examine and quantify the linguistic properties of adapted texts (Crossley et al. 2007; Crossley et al. 2011). However, such quantification of linguistic features does not provide adequate assistance to teachers in determining the grade-appropriateness of English curriculum teaching materials in the absence of appropriate benchmarks of linguistic complexity suitable for different grade levels, nor does it provide direct feedback to teachers' text adaptation processes (Jin et al. 2020).

Previous research has sought to construct benchmark measures of EFL teaching materials in terms of syntactic complexity and vocabulary profiles (Jin et al. 2020; Jin and Lu 2018). Where the vocabulary profile offers a basic word list for a learning grade, lexical complexity offers a more holistic and contextualized approach to developing a functional foreign language and includes lexical density, sophistication, and diversity (Allaw 2021; Bui 2021). Since benchmarks for lexical complexity have yet to be developed, the purpose of this study is to offer a comprehensive analysis of the lexical complexity of sample texts across China's 12 grades of the English curriculum. An in-depth examination of the lexical complexity of exemplar texts can also reveal features that distinguish various grade levels.

2 Literature review

2.1 Approaches to text adaptation

Preparing materials appropriate for the intended learners has always been a central and challenging issue for language teachers. It requires a certain degree of text adaptation to ensure the linguistic appropriateness of the texts for the intended learners (Chen and Meurers 2018; Jin et al. 2020).

Teachers tended to favor the intuitive approach to text adaptation which involves adapting the available or original texts to learner ability based on their prior knowledge or experiences (Allen 2009; Simensen 1987). Several studies have investigated intuitive strategies used by language instructors or experts, such as modifying punctuation, replacing low-frequency words, transforming passive voice into active, adding repetition to lower lexical diversity, replacing polysemous words with unambiguous ones, modifying the proportion of word types, and simplifying complicated syntactic structures (Green and Roger 2012; Noone 2018; Oh 2001; Rets et al. 2022; Ross et al. 1991; Young 1999). Among the different types of adaptations, lexical changes are the primary consideration in text adaptation. Previous research has approved teachers' intuitive approach to text adaptation in identifying texts that are more comprehensible or suitable for the target learners (Oh 2001; Rets and Rogaten 2021). Nevertheless, problems existed. First, earlier research primarily focused on the text adaptation skills of experienced language teachers in adapting original texts to a specific level, rather than on developing benchmark metrics for text adaptation tailored to students of varying proficiency levels. Inexperienced teachers may lack the skills to assess their students' language proficiency levels or learning capacities (Carabantes and Paran 2022), which can result in inconsistent collation of teaching material of inappropriate complexity (Green and Roger 2012; Noone 2018). Moreover, it has been found that the adapted texts might be over-represented by high-frequency words and atypical or simplified sentence structures, which cannot facilitate students' language development (Hiebert and Mesmer 2013). Therefore, more guidance for teachers is needed to mitigate any negative influence the intuitive judgment may cause.

Compared with the subjective intuitive approach, a more objective approach to text adaptation is the structural approach. One method typically adopted by graded readers is to follow specified wordlists and structure lists for different grade levels (Hill 2008). However, the wordlists or structure lists are not considered prescriptive since publishers still have a degree of flexibility in setting the language standards

of their graded readers (Claridge 2012). Another structural method is to use a range of quantifiable linguistic features, such as referring to readability formulas (Crossley et al. 2012b; Simensen 1987). Traditional readability formulas, such as the Flesch-Kincaid (Kincaid et al. 1975) and Dale-Chall (Chall and Dale 1995), typically take into account the length or number of words, sentences, or syllables. However, these traditional readability formulas demonstrated an oversimplified representation of the text difficulty and therefore cannot be regarded as reliable guidelines for text adaptation. More recent readability formulas, such as Coh-Metrix (Crossley et al. 2008), cover an even wider variety of linguistic features, including lexical and syntactic complexity features, discourse features, and the psycholinguistic and cognitive aspects that might influence text processing. However, it is argued that even with the guidance of various graded lists or the reference of quantitative statistics of readability formula, teachers still need to rely on their intuition during the text adaptation process (Allen 2009). While the structural approach can help mitigate the subjectivity of the intuitive approach in text adaptation, the extensive range of linguistic features and complex formulas may pose challenges for teachers in understanding and implementing them effectively. Moreover, without appropriate alignment of the quantitative statistics with curriculum standards in different educational contexts, teachers may struggle to determine the appropriate level of linguistic complexity suitable for their intended students (Sung et al. 2015).

Some efforts have been made to establish benchmarks for adapting teaching materials in EFL contexts using lexical coverage profiles and syntactic complexity measures. For example, several studies have examined vocabulary coverage in high-stakes English tests in Japan (Chujo 2004) and China (Jin et al. 2017; Jin and Lu 2018) to propose consistent data-driven benchmarks of text features rather than using intuition. Jin et al. (2020) systematically examined the syntactic complexity of textbooks across 12 grades in China's Basic English curriculum, identifying distinguishing features of syntactic complexity that provided direct insights into the establishment of syntactic complexity benchmarks for future text adaptation practices. It is useful to follow up this study of syntactic complexity with one on lexical complexity.

In view of the drawbacks of both the intuitive and structural approaches, along with the lack of benchmark standards, it is crucial to conduct a systematic examination of the lexical complexity of exemplar texts across 12 grade levels. Such an endeavor can offer valuable insights for the development of benchmark references that frontline teachers can utilize when selecting teaching materials suitable for their intended learners.

2.2 Lexical complexity

Lexical complexity is an important construct in language education research. A variety of research has been conducted to understand the relationship between lexical complexity and language development, reading and writing quality or text comprehension (Allaw 2021; Cumming et al. 2005; Engber 1995; Lu 2012). By definition, lexical complexity is a multidimensional construct, including lexical density, sophistication, and variation (Allaw 2021; Bui 2021).

Lexical density quantifies the ratio of lexical items or content words to running words in a text (Ure 1971). Specifically, lexical items, in contrast with grammatical items, typically include "nouns, adjectives, verbs (excluding modal verbs, auxiliary verbs), and adverbs with an adjectival base" (Lu 2012: 192) that could "contribute with content to the meaning of a sentence" (Vicente 2018: 947). Although lexical density is known as one of the dimensions of lexical complexity, researchers question the utility of this construct since neither a significant relationship between the lexical density score and the readability of texts (To et al. 2013) nor a correlation between lexical density and L2 oral and written performances (Engber 1995; Lu 2012) has been found.

Lexical diversity, also known as lexical variation or lexical range, measures the variety of words used in texts (Lu 2012). Scholars have discovered a close correlation between lexical diversity and students' language proficiency and vocabulary knowledge (Crossley et al. 2012a; Housen et al. 2011). The most straightforward measurements of lexical diversity include the number of different words in a text (NDW) and the type-token ratio (TTR) (Lu 2012; Jarvis 2013). However, the simple type-token ratio is sensitive to the length of the text (Engber 1995). To mitigate the impact of text length, researchers have proposed transformed versions of TTR, such as Corrected TTR (CTTR; Carroll 1964) and Root TTR (RTTR; Guiraud 1960). Despite the controversies surrounding TTR and its transformations, previous research has recognized them as discriminative features for assessing the speaking and writing abilities at different proficiency levels (Cumming et al. 2005; Daller et al. 2003; Lu 2012). Jarvis (2013) argued that TTR primarily focuses on lexical repetition rather than the actual lexical diversity of a text. He thus advocated using the Measure of Textual Lexical Diversity (MTLD) developed by McCarthy (2005), which takes into consideration the influence of text length and the evenness of vocabulary distribution. However, MLTD contains relatively complex algebraic formulas which may present challenges for teachers in understanding the concept and applying it to their text adaptation practices. Other measures of lexical diversity focus on the variation of specific word types, such as computing the ratio of verb types to total verb tokens or using a corrected version to mitigate the impact of text length

(Harley and King 1989). McClure (1991) calculated the ratio of specific word types, including lexical words, nouns, and modifiers, moving beyond the sole calculation of overall word variety. However, consensus has yet to be reached on whether these features can effectively discriminate texts of different levels (McClure 1991; Lu 2012).

The third dimension, lexical sophistication, also known as lexical rareness, focuses on infrequently used words that require a higher level of contextual knowledge (Linnarud 1987). Traditionally, researchers adopted a frequency-based approach by examining frequency bands (Morris and Cobb 2004) or calculating the mean frequency scores of words in a text relative to the reference corpus (Jarvis 2013; Lu and Hu 2022). The frequency-based lexical sophistication measures describe the percentage of words that are uncommon or advanced in a text (Lu and Hu 2022). It has been demonstrated that texts with more frequent words tend to be simpler (Crossley et al. 2007). Recently, researchers have advocated for the inclusion of word sense data when evaluating the lexical sophistication of texts for multiple reasons (Deane et al. 2006; Lu and Hu 2022). First, when encountering polysemous words, students exert more effort to select the context-specific meaning from a word's multiple senses in order to correctly comprehend the text (González-Fernández and Schmitt 2020). Different word senses may present varying levels of difficulty for students (Lu and Hu 2022). Students may find it challenging when encountering low-frequency senses of high-frequency words typically employed in specific subject fields (Skoufaki and Petri 2021). Recent scholars have therefore attempted to disambiguate word senses in texts and calculate the frequency of particular senses to indicate lexical sophistication (Hu et al. 2019; Lu and Hu 2022).

3 The present study

This study aims to examine the variations in lexical complexity of exemplar texts across various grade levels and identify the lexical complexity features that could distinguish between the grade levels. It is hoped that the findings of this study would offer insights into the establishment of lexical complexity benchmarks. The following two research questions aim to assist in developing suitable benchmarks for text adaptation for the 12 EFL levels in China.

1. What are the overall patterns of lexical complexity of exemplar texts across 12 grade levels?
2. What are the features that best distinguish lexical complexity of different grade levels?

4 Methods

4.1 Dataset

In EFL contexts, textbooks are the primary input for students in language learning (Chen and Meurers 2018), and thus textbooks can serve as compilations of exemplar texts for exploring benchmarks for teachers in material preparation (Jin et al. 2020). For this reason, the data in this study consists of exemplar texts from 12 grades of the Basic English Curriculum in China. The exemplar texts were systematically sampled from various textbook series used across different regions of the Chinese mainland. The data collection process spanned five years, resulting in the collection of a total of 50 packages of exemplar texts published by 16 publishers from 2006 to 2021. The development of textbooks follows the guiding principle outlined in the Regulations on Textbook Management for Primary and Secondary Schools, which mandates adherence to the Curriculum Standards set by the National Textbook Committee Office (Ministry of Education 2019). Therefore, although the textbooks are published by different publishers, they all adhere to two English curriculum standards, i.e., The English Curriculum Standards for Compulsory Education (Ministry of Education 2012) (for textbooks of Grade 1–9); and The English Curriculum Standards for Senior High Schools (Ministry of Education 2003, 2020), both of which follow standardized procedures for textbook compilation. A team of experts in English, faculty and staff members of teaching and research, and frontline teachers of primary and secondary schools oversee the compilation of each series of textbooks (Ministry of Education 2019). First drafts of the textbooks are reviewed by an audit team of English experts, and then trialed and revised multiple times following feedback from teachers and students. The entire compilation and publication process is designed to ensure that the complexity of the textbooks is suitable for students of varying grade levels (Jin et al. 2020).

Due to the considerable variation in the number of exemplar texts among different grade levels, a balanced dataset was created by randomly selecting two hundred documents from the 50 packages of exemplar texts for each grade. To ensure the reliability of the lexical complexity analysis, only texts that contained at least one clause were included in the analysis. The final corpus used in this study comprises a total of 2,372 exemplar texts spanning Grades 1–12 (see Table 1).

Table 1: Details of the dataset.

Grade	Number of texts	Total number of words	MIN words per text	MAX words per text	Average number of words per text	Standard deviation (SD) of words per text
1	186	5,085	5	221	27.34	21.99
2	194	6,978	5	260	35.97	27.71
3	193	6,422	6	191	33.27	19.70
4	200	12,062	7	208	47.47	28.94
5	199	14,569	7	619	73.21	57.19
6	200	18,369	5	917	91.85	79.12
7	200	29,451	46	303	147.26	47.58
8	200	48,454	53	341	208.94	50.61
9	200	51,311	102	403	256.56	59.82
10	200	81,181	61	1,190	405.91	171.88
11	200	102,141	46	1,448	510.71	277.17
12	200	119,140	73	1,398	595.70	242.43
Total	**2,372**	**550,456**	**5**	**1,448**	**204.86**	**228.22**

The bold values in the "Total" column represent the aggregate data for all 12 levels as a whole dataset across the specific variable or dimension being measured in the dataset.

4.2 Operationalization of lexical complexity

In this study, we carefully selected a set of lexical complexity measures for analysis based on three criteria: (1) their relevance to teachers' actual text adaptation practices, (2) their ability to capture the multidimensional nature of lexical complexity, and (3) their widespread popularity in the examination of L2 texts. In addition, to make sure that the lexical complexity measures in our study are understandable for teachers to guide their text adaptation practices, we selected relatively simple measures that do not involve complex algebraic formulas. Since there is insufficient evidence to establish a clear relationship between lexical density and the readability of texts and considering that teachers' text adaptation practices do not primarily focus on lexical density, this study has solely concentrated on two dimensions: lexical diversity and lexical sophistication. A total of 17 indices were selected for analysis.

4.2.1 Lexical diversity

Lexical diversity encompasses the range of words used in a text (Lyashevskaya et al. 2021). To customize texts according to learner proficiency levels, teachers sometimes make adjustments to the text length, manipulate word repetition to modify the overall vocabulary diversity (e.g., Green and Roger 2012; Oh 2001; Ross et al. 1991), and

revise the proportion of words belonging to different parts of speech, such as nouns, verbs, and modifiers (Rets et al. 2022). As a result, three sub-categories of lexical diversity indices were included in the analysis. The first category is concerned with the number of different words in a text (NDW). The second category analyzed the diversification of the total vocabulary by utilizng the type-token ratio (TTR), which contrasts the number of unique words with the total number of words in a text. However, the simple type-token ratio is sensitive to the text length (Engber 1995). Therefore, a relatively simple transformed form of TTR, i.e., root TTR (RTTR), that is mediate the influence of the text length (Guiraud 1960), is also used in this study. The third category consists of eight indices related to the variation of specific lexical word types (Lu 2012; McClure 1991), including the variation of lexical words, verbs, nouns, adjectives, adverbs, and modifiers (in this study, the sum of adjectives and adverbs), see Table 2. Statistics for these lexical diversity indices were retrieved based on Python scripts of the L2 Lexical Complexity Analyzer (L2LCA) (Lu 2012).

4.2.2 Lexical sophistication

The most frequently mentioned strategy in teacher text adaption practices is substituting low-frequency words with high-frequency synonyms (Green and Roger

Table 2: Lexical diversity indices (Lu 2012).

Feature	Code	Formula*	Source
Number of different words	NDW	T	Klee (1992); Miller (1991)
Type-token ratio	TTR	$\frac{T}{N}$	Templin (1957)
	RTTR	$\frac{T}{\sqrt{N}}$	Guiraud (1960)
Lexical word diversity	LV	$\frac{T(\text{lex})}{N(\text{lex})}$	Casanave (1994)
Verb diversity	VV1	$\frac{T(\text{verb})}{N(\text{verb})}$	Harley and King (1989)
	CVV1	$\frac{T(\text{verb})}{\sqrt{2N(\text{verb})}}$	Lu (2012)
	VV2	$\frac{T(\text{verb})}{N(\text{lex})}$	Lu (2012)
Noun diversity	NV	$\frac{T(\text{noun})}{N(\text{lex})}$	McClure (1991)
Adjective diversity	AdjV	$\frac{T(\text{adj})}{N(\text{lex})}$	McClure (1991)
Adverb diversity	AdvV	$\frac{T(\text{adv})}{N(\text{lex})}$	McClure (1991)
Modifier diversity	ModV	$\frac{T(\text{adj+adv})}{N(\text{lex})}$	McClure (1991)

Note: *T, number of word types; T(lex), number of lexical word types; T(verb), number of verb types; T(noun), number of noun types; T(adj), number of adjective types; T(adv), number of adverb types; T(adj + adv), number of adjective and adverb types; N, number of word tokens; N(lex), lexical word tokens; N(verb), number of verb tokens; N(noun), number of noun tokens; N(adj), number of adjective tokens; N(adv), number of adverb tokens; N(adj + adv), number of adjective and adverb tokens.

2012; Noone 2018; Oh 2001; Rets et al. 2022). In some cases, teachers might also substitute polysemous words that students find difficult to understand (Young 1999). These text adaptation strategies pertain to another dimension of lexical complexity, lexical sophistication, which engages both word frequencies and word senses. In other words, both word-form-based indices and sense-aware indices are included in analyzing the lexical sophistication of texts.

To retrieve the data of language sophistication measures, language models were trained to recognize and label word senses and frequency data in the exemplar texts using natural language processing technology. Based on Lu and Hu's (2022) Python scripts, sense-aware lexical sophistication indices were retrieved. First, the Oxford Dictionary Online was used to retrieve all the word senses and corresponding example sentences for polysemous words, which were then fed as input for training the word sense labeling system. With reference to the Corpus of Historical American English (Davies 2012), the frequency distribution of each word sense was estimated. Following Lu and Hu (2022), we determined whether a word or word sense is sophisticated based on an empirically-determined reference frequency level termed as "*K*". When the frequency of a monosemous word or the frequency of a word sense for a polysemous word in the reference corpus falls below the reference level *K*, the word or word sense is classified as sophisticated (Lu and Hu 2022). Informed by the reference frequency of 3,000 in the study of Lu and Hu (2022), we explored various levels of "*K*", ranging from 3,000 to 6,000 in increments of 500. At the reference frequency of 5,500, all six lexical sophistication indices were strongly correlated with grade levels, as shown by Spearman's correlation analysis. Therefore, "*K*" was ultimately set at 5,500. The six lexical sophistication indices are listed in Table 3.

Table 3: Lexical sophistication indices.

Definition	Code	Formula	Source
Mean logarithmic frequency of sophisticated word tokens	MW	$\sum_{j=1}^{n} \log f(w_j)$	Kyle and Crossley (2015); Lu and Hu (2022)
Root-ratio of sophisticated word tokens	RSN	$\frac{N(sw)}{\sqrt{N(sw)}}$	Kyle and Crossley (2015); Lu and Hu (2022)
Root-ratio of sophisticated word types	RST	$\frac{T(sw)}{\sqrt{T(sw)}}$	Kyle and Crossley (2015); Lu and Hu (2022)
Mean logarithmic frequency of sophisticated word senses	MS	$\sum_{i=1}^{n} \log f(s_i)$	Lu and Hu (2022)
Root-ratio of sense-aware sophisticated word tokens	RSSN	$\frac{N(sa\_sw)}{\sqrt{N(sa\_sw)}}$	Lu and Hu (2022)
Root-ratio of sense-aware sophisticated word types	RSST	$\frac{T(sa\_sw)}{\sqrt{T(sa\_sw)}}$	Lu and Hu (2022)

4.3 Statistical analysis

After retrieving word sense data for exemplar texts, statistical analysis was conducted to answer the research questions. Descriptive statistics regarding the lexical complexity indices of various grade levels were provided to address the first research question. To address the second research query into the distinguishing features of LC at each grade level, data transformation and alternative non-parametric analyses have been taken into consideration since some data was skewed and did not meet the homogeneity assumption. However, we decided to conduct one-way ANOVAs, following Iwashita et al. (2008), when analyzing the lexical complexity of oral texts of varying proficiency levels. The reasons are as follows: (1) the challenge of utilizing transformed data; (2) inappropriateness of non-parametric tests when the groups do not share the same spread; and, (3) ANOVA robustness even in conditions without normal distribution, equal variances and given a large sample size (Blanca et al. 2017). The one-way ANOVA helped to determine whether lexical complexity indices differed significantly across grade levels, with grade level serving as the independent variable and the lexical complexity indices serving as the dependent variables. In addition to one-way ANOVAs, Games-Howell post hoc tests were performed to detect adjacent grades with significant differences in lexical complexity.

5 Results

5.1 RQ1: patterns of lexical complexity across 12 grades

5.1.1 Lexical diversity

The descriptive statistics of lexical diversity measures that attend to all types of vocabularies are presented in Table 4. As shown in Figure 1, the average number of different words in a text gradually increased over 12 grades. Gradually, the growth rate accelerated across the four stages (Grades 1–3, Grades 4–6, Grades 7–9 (junior high school), and Grades 10–12 (senior high school). Figure 2 demonstrates the distinct variation patterns of the type-token ratio and its transformed form. TTR fluctuated across 12 grade levels, while its transformed form, i.e., RTTR, grew nearly linearly from lower to higher grade levels, signifying a gradual increase of lexical diversity across the 12 grades. The different patterns observed between TTR and RTTR can be attributed to TTR's sensitivity to text length (Lu 2012).

Table 4: Descriptive lexical diversity measures applied to total vocabulary.

Grade	*N*	NDW		TTR		RTTR	
		M	SD	M	SD	M	SD
1	186	13.62	7.31	0.578	0.193	2.682	0.743
2	194	19.34	10.73	0.601	0.177	3.259	0.936
3	193	18.82	8.83	0.615	0.176	3.281	0.898
4	200	27.22	13.88	0.610	0.139	3.915	1.032
5	199	39.93	20.19	0.604	0.135	4.667	1.186
6	200	50.69	28.26	0.610	0.119	5.268	1.220
7	200	80.68	24.51	0.555	0.072	6.606	1.114
8	200	114.92	24.10	0.557	0.058	7.939	0.902
9	200	137.15	26.63	0.542	0.055	8.557	0.884
10	200	197.63	65.93	0.509	0.072	9.800	1.415
11	200	235.87	97.36	0.502	0.094	10.416	1.847
12	200	274.81	88.72	0.482	0.067	11.237	1.624

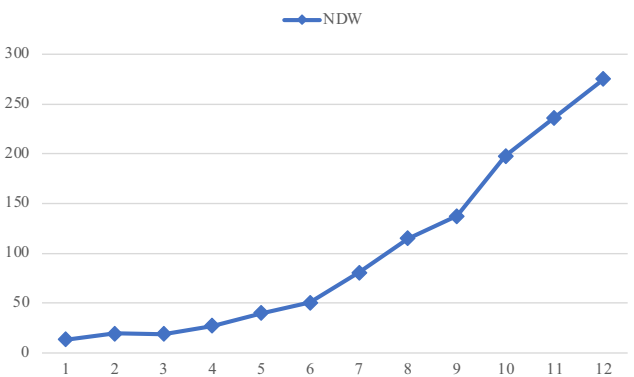

Figure 1: Number of different words.

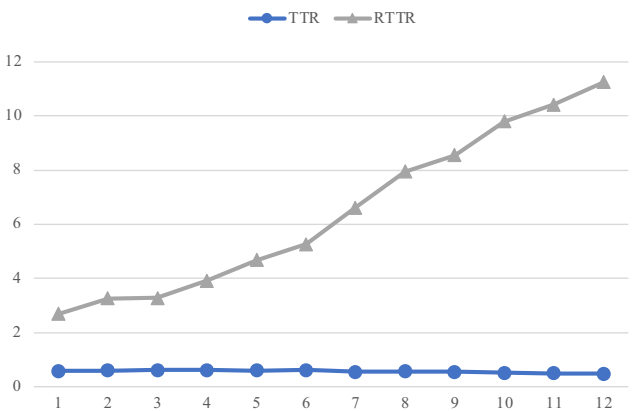

Figure 2: TTR and transformation measures.

The diversity of particular word types was also examined. Figure 3 shows that across the 12 grades, the ratio of verb types to verb tokens or lexical words in a text (VV1 and VV2) did not present a discernable upward or downward trend. Only the transformed form, i.e., CVV1, rose from a lower to a higher level. The reason behind this observation might be that the transformed form reduces the influence of text length. The descriptive statistics for the three verb diversity indices are shown in Table 5.

Table 6 below presents the descriptive results of the diversity of other specific word types.

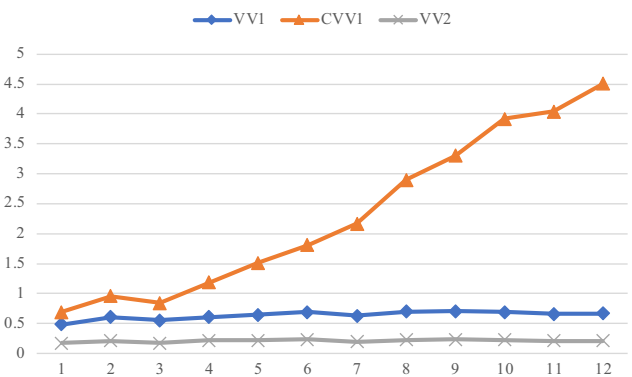

Figure 3: Verb diversity measures.

Table 5: Descriptive results of verb diversity measures.

Grade	*N*	VV1		CVV1		VV2	
		M	SD	M	SD	M	SD
1	186	0.485	0.312	0.688	0.413	0.168	0.147
2	194	0.607	0.267	0.959	0.399	0.208	0.124
3	193	0.552	0.302	0.844	0.440	0.170	0.101
4	200	0.603	0.246	1.185	0.490	0.215	0.129
5	199	0.646	0.202	1.513	0.547	0.218	0.091
6	200	0.686	0.177	1.813	0.617	0.234	0.082
7	200	0.625	0.154	2.170	0.690	0.193	0.062
8	200	0.696	0.103	2.895	0.563	0.224	0.053
9	200	0.705	0.091	3.306	0.494	0.233	0.045
10	200	0.693	0.100	3.919	0.756	0.223	0.049
11	200	0.661	0.158	4.038	1.238	0.206	0.064
12	200	0.666	0.089	4.505	0.940	0.207	0.045

Table 6: Descriptive results of lexical diversity measures applied to other word types.

Grade	*N*	LV		NV		AdjV		AdvV		ModV	
		M	SD	M	SD	M	SD	M	SD	M	SD
1	186	0.620	0.225	0.678	0.263	0.075	0.095	0.006	0.023	0.081	0.097
2	194	0.641	0.198	0.694	0.247	0.062	0.078	0.012	0.031	0.074	0.085
3	193	0.651	0.182	0.739	0.212	0.081	0.080	0.010	0.026	0.091	0.087
4	200	0.673	0.153	0.733	0.201	0.087	0.076	0.013	0.034	0.100	0.083
5	199	0.664	0.152	0.699	0.194	0.078	0.073	0.026	0.042	0.105	0.082
6	200	0.687	0.138	0.694	0.174	0.094	0.065	0.024	0.028	0.118	0.066
7	200	0.655	0.103	0.665	0.140	0.106	0.044	0.027	0.021	0.133	0.051
8	200	0.688	0.082	0.671	0.118	0.117	0.038	0.031	0.019	0.149	0.041
9	200	0.681	0.070	0.655	0.095	0.112	0.030	0.034	0.016	0.146	0.033
10	200	0.674	0.081	0.663	0.100	0.116	0.028	0.034	0.016	0.150	0.030
11	200	0.668	0.096	0.661	0.112	0.111	0.034	0.029	0.016	0.140	0.038
12	200	0.668	0.074	0.663	0.087	0.116	0.027	0.030	0.015	0.146	0.033

As shown in Figure 4, the proportion of lexical word types to lexical word tokens, i.e., LV, remained stable across the 12 grades despite some slight fluctuations. The diversity of nouns increased from Grade 1 to 3, but then showed a downward trend from Grade 4 to 9 and remained relatively stable in Grade 10 to 12. The diversity of adjectives, adverbs, and modifiers showed a slight upward trend from lower to higher grade levels with some fluctuations.

5.1.2 Lexical sophistication

Table 7 summarizes the descriptive statistics of two frequency-based indices, i.e., the mean logarithmic frequency of words and the mean logarithmic frequency of senses.

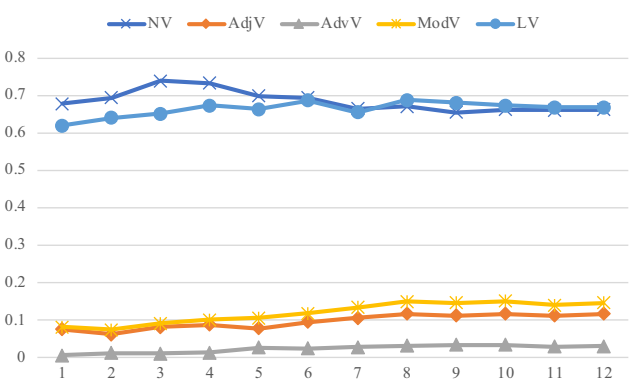

Figure 4: Diversity measures of specific word types.

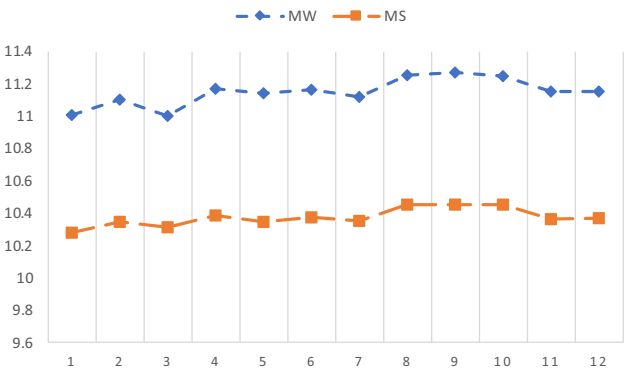

Figure 5: Mean frequency measures of lexical sophistication.

Table 7: Descriptive results of mean frequency sophistication measures.

Grade	*N*	MW		MS	
		M	SD	M	SD
1	186	11.006	1.068	10.277	1.158
2	194	11.104	0.792	10.344	0.778
3	193	11.001	0.853	10.311	0.936
4	200	11.167	0.604	10.388	0.592
5	199	11.139	0.543	10.346	0.566
6	200	11.162	0.440	10.376	0.438
7	200	11.119	0.412	10.351	0.427
8	200	11.255	0.281	10.451	0.279
9	200	11.269	0.268	10.455	0.281
10	200	11.249	0.292	10.453	0.273
11	200	11.153	0.325	10.362	0.300
12	200	11.154	0.246	10.366	0.227

The descriptive statistics revealed that both frequency-based indices fluctuated rather than progressed from lower to higher grade levels (see Figure 5), with the MW ranging from 11.001 to 11.269 and the MS ranging from 10.277 to 10.455.

The results of the root-ratio indices differed from those of the frequency-based indices (see Table 8). As depicted in Figure 6, all four ratio-based indices demonstrated an upward trend from lower to higher grade levels, notwithstanding some fluctuations from Grade 2 to Grade 3. Particularly, the four indices remained relatively stable from Grade 1 to Grade 6, with a slight rise between Grades 4 and 6. From seventh to ninth grade, the four indices increased relatively faster, whereas Grade 10 to 12 witnessed the sharpest increase. It should be noted that the two sense-aware

Table 8: Descriptive results of root ratio sophistication measures.

Grade	*N*	RSN		RSSN		RST		RSST	
		M	SD	M	SD	M	SD	M	SD
1	186	0.685	0.546	1.198	0.848	0.549	0.405	1.004	0.643
2	194	0.863	0.649	1.431	0.916	0.698	0.505	1.242	0.754
3	193	0.862	0.590	1.369	0.849	0.742	0.459	1.243	0.671
4	200	0.932	0.565	1.581	0.715	0.844	0.490	1.446	0.593
5	199	1.205	0.714	1.950	0.974	0.945	0.472	1.658	0.685
6	200	1.262	0.789	2.169	1.050	0.961	0.485	1.822	0.704
7	200	1.754	0.700	2.987	0.922	1.368	0.469	2.570	0.619
8	200	1.890	0.645	3.225	0.770	1.555	0.442	2.952	0.597
9	200	2.106	0.685	3.608	0.865	1.737	0.482	3.340	0.687
10	200	2.734	1.026	4.594	1.272	2.528	0.890	4.523	1.217
11	200	3.577	1.405	5.531	1.822	3.394	1.196	5.508	1.743
12	200	4.049	1.271	6.139	1.617	3.940	1.174	6.266	1.645

sophistication indices increased at a higher rate than the conventional word-based lexical sophistication indices.

5.2 RQ2: distinguishing features of lexical complexity within the 12 grades

5.2.1 Lexical diversity

One-way ANOVAs revealed significant differences between grade levels for all lexical diversity indices. Table 9 summarizes the results for one-way ANOVAs and the post

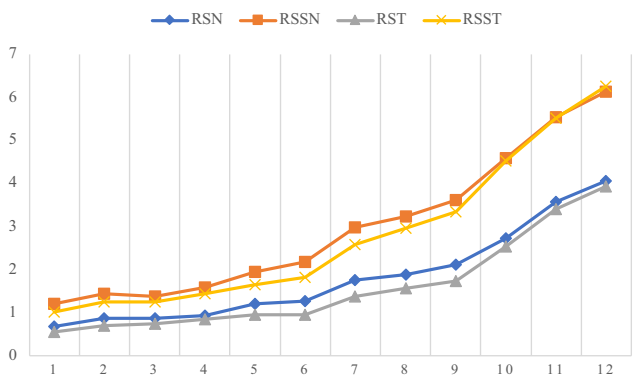

Figure 6: Root-ratio based measures of lexical sophistication.

Table 9: Results of one-way ANOVAs for lexical diversity indices.

Measure	η_p^2	Effect size label	Post hoc tests*
NDW	0.886	Large	1, 2–3, 4, 5, 6, 7, 8, 9, 10, 11, 12
TTR	0.348	Moderate	1-6, 7–9, 10-12
RTTR	0.925	Large	1, 2–3, 4, 5, 6, 7, 8, 9, 10, 11, 12
LV	0.135	Small	1–7, 8–12
VV1	0.300	Moderate	1, 2–6, 7, 8–12
CVV1	0.888	Large	1, 2–3, 4, 5, 6, 7, 8, 9, 10–11, 12
VV2	0.227	Small	1–2, 3, 4–6, 7, 8–12
NV	0.155	Small	1–12
AdjV	0.293	Moderate	1–12
AdvV	0.354	Moderate	1–4, 5–12
ModV	0.384	Moderate	1–7, 8–12

Note: *We use "," to indicate significant differences between adjacent grade levels in the post-test results. For example, regarding the measure "NDW", significant differences were observed between almost all adjacent grade levels, except between Grade 2 and Grade 3.

hoc tests. Researchers have raised concerns regarding the leniency of Cohen's guidelines for interpreting effect size (Olejnik and James 2000). They argue that when a single group exhibits significant deviation from others, the effect size derived from the omnibus test can be notably skewed, leading to an overestimation that fails to accurately reflect the true differences among the groups under study (Levin 1967). In response to this concern, this study adopts Ferguson's more stringent guidelines (2009) for interpreting effect size in social science studies. According to Ferguson, squared association indices between 0.04 and 0.25 can be considered indicative of a small effect size, those between 0.25 and 0.64 as representing a moderate effect, and values exceeding 0.64 as indicative of a large effect.

Two indices measuring the lexical diversity of total vocabularies exhibited a large effect size: NDW: F (11, 2,360) = 637.53, $p < 0.001$, $\eta_p^2 = 0.886$; RTTR: F (11, 2,360) = 1,130.93, $p < 0.001$, $\eta_p^2 = 0.925$. TTR demonstrate a moderate effect size: F (11, 2,360) = 25.47, $p < 0.001$, $\eta_p^2 = 0.348$. Post hoc tests revealed that NDW and RTTR could distinguish almost all the adjacent grade levels, with one exception: Grades 2 and 3 remained stable on all three measures on total vocabulary diversity. TTR stayed virtually unchanged between Grades 1–6, Grades 7–9, and Grades 10–12. Significant changes tended to happen between the transition of different educational stages, e.g., transitioning from primary school to junior high school from Grade 6 to 7, and from junior to senior high school from Grade 9 to 10.

Among the indices that examine the diversity of specific word types, one-way ANOVAs revealed that only the CVV1 exhibited a large effect size: F (11, 2,360) = 649.55, $p < 0.001$, $\eta_p^2 = 0.888$. The one verb diversity measure, and the diversity measures

for adjectives, adverbs, and modifiers exhibited a moderate effect size: VV1: F (11, 2,360) = 20.42, p < 0.001, η_p^2 = 0.300; AdjV: F (11, 2,360) = 18.99, p < 0.001, η_p^2 = 0.293; AdvV: F (11, 2,360) = 30.24, p < 0.001, η_p^2 = 0.354; ModV: F (11, 2,360) = 35.50, p < 0.001, η_p^2 = 0.384. One verb diversity measure and the diversity measures for lexical words and nouns showed a marginal effect size in one-way ANOVAs: VV2: F (11, 2,360) = 20.42, p < 0.001, η_p^2 = 0.227; LV: F (11, 2,360) = 3.97, p < 0.001, η_p^2 = 0.135; NV: F (11, 2,360) = 5.28, p < 0.001, η_p^2 = 0.155. Post hoc tests showed a number of patterns of variation of these measures across 12 grade levels. CVV1 could distinguish between the majority of adjacent grade levels, with the exception of Grades 2 and 3, and 10 and 11. Although significant differences in NV and AdjV were identified across the 12 grade levels, the differences between adjacent grade levels for these two measures were not clear-cut. For the remainder of the measures examining the diversity of specific text types, significant changes were identified at various primary school grade levels (Grades 1–6). Except for CVV1, all measures examining the diversity of specific word types remained stable after the eighth grade.

5.2.2 Lexical sophistication

Table 10 presents the one-way ANOVA results for lexical sophistication indices. As for the two mean-frequency-based lexical sophistication indices, although one-way ANOVAs identified significant differences in MW among different grade levels: MW: F (11, 2,360) = 4.512, p < 0.001, η_p^2 = 0.144; MS: F (11, 2,360) = 1.791, p = 0.05, η_p^2 = 0.091, the effect size of both measures was marginal. Post-hoc tests identified significant differences only between Grade 8 and the adjacent grades in MW, whereas in MS, there were no significant differences between adjacent grade levels.

For the four root-ratio based lexical sophistication indices, results of one-way ANOVAs revealed that there was a significant difference between the 12 grades and all the four measures had large effect size: RSN: F (11, 2,360) = 340.95, p < 0.001,

Table 10: Results of one-way ANOVAs for lexical sophistication indices.

Measure	η_p^2	Effect size label	Post hoc tests*
MW	0.144	Small	1–7, 8–12
RSN	0.783	Large	1-4, 5–6, 7–9, 10, 11, 12
RST	0.843	Large	1–6, 7, 8, 9, 10, 11, 12
MS	0.091	Small	1–12
RSSN	0.826	Large	1–4, 5–6, 7–8, 9, 10, 11, 12
RSST	0.870	Large	1, 2–4, 5–6, 7, 8, 9, 10, 11, 12

Note: *We use "," to indicate significant differences between the adjacent grade levels in the post-test results. For example, regarding the measure "MW", significant differences were identified only between Grade 7 and Grade 8.

η_p^2 = 0.783; RST: F (11, 2,360) = 537.15, p < 0.001, η_p^2 = 0.843; RSSN: F (11, 2,360) = 461.56, p < 0.001, η_p^2 = 0.826; RSST: F (11, 2,360) = 668.10, p < 0.001, η_p^2 = 0.870. Post hoc analyses revealed that the differences between adjacent levels of all four indices were more pronounced at higher grade levels and that the sense-aware indices exhibited similar patterns to their word-form-based counterparts but performed better at discriminating between adjacent grade levels. The RSN and RSSN both remained stable in Grades 1–4, Grades 5–6, and Grades 7–8, but rose significantly in Grades 5, 7, 10, 11, and 12. In addition, both RST and RSST increased significantly from Grade 7–12 However, the word form-based index, i.e., RST, remained stable throughout primary school (Grades 1–6), whereas RSST experienced significant growth in Grades 2–5.

5.3 Summary

In general, descriptive statistics of lexical diversity and sophistication indices proved that the lexical complexity of exemplar texts varied across grade levels. Except for the mean logarithmic frequency of word senses, all lexical diversity and sophistication measures revealed significant differences across grade levels. However, the measures did not always demonstrate a gradual increase from lower to higher grade levels, and the differences between adjacent grade levels were occasionally ambiguous. Two lexical diversity measures applicable to total vocabularies (i.e., number of different words and the root-type-token-ratio) gradually increased from lower to higher grade levels. These two measures had a large effect size in one-way ANOVAs and effectively distinguished the lexical diversity of adjacent grade levels. For diversity measures of specific word types, only CVV1 (verb diversity) exhibited an upward trend from lower to higher grade levels and could distinguish the majority of differences between adjacent grade levels. Other measures fluctuated in earlier grades but remained relatively stable after the eighth grade.

The variation patterns for measures of lexical sophistication diverge from the measures of lexical diversity. In contrast to the lexical diversity measures of specific word types, which changed significantly in lower grade levels but remained stable in higher grade levels, the four root-ratio based sophistication indices rose slightly in lower grade levels but increased dramatically after Grade 8. All four indices may be used as distinguishing indices of lexical sophistication. In addition, the growth of the two sense-aware lexical sophistication indices was more pronounced than that of the two word-form based indices at higher grade levels, suggesting that the former has a greater capacity to discriminate between texts of higher grade levels. The two word-form based sophistication indices fluctuated among the twelve grades. Both

indices obtained marginal effect sizes in the one-way ANOVAs and were unable to distinguish between adjacent grade levels.

6 Discussion

In the present study, we examined how lexical complexity in exemplar texts varied across the 12 grades of the basic English curriculum in China, as well as what features could account for differences in lexical sophistication across the grade levels. The findings of the present study are expected to shed light on benchmarking the lexical sophistication of reading materials and should therefore help build an automatic text adaption feedback system in the future.

6.1 RQ1: lexical complexity graduation

The results show that the lexical complexity of exemplar texts varied by grade level. All of the lexical diversity and sophistication indices were significantly different, with the exception of the mean logarithmic frequency of word senses, which was marginally significant (p = 0.05), indicating that the expert textbook compilers and editors considered various aspects of lexical complexity when selecting and adapting reading materials (Green and Roger 2012; Jin et al. 2020). Specifically, the number of unique words in a text demonstrated a pattern of accelerated growth across educational stages. The transformed indices that measure the proportion of unique words in a text, i.e., RTTR, exhibited an ascending pattern. The progression patterns of lexical diversity indices applied to total vocabularies reflect the English Curriculum Standards objectives for teaching vocabulary (Ministry of Education 2020, 2022). According to the English Curriculum Standards, students should acquire approximately 505 words during primary school (Grades 1–6), 1,095 new words during junior high school (Grades 7–9), and around 1,500–1,600 new words during senior high school (Grades 10–12). Regarding the diversity of specific word types, only one corrected measure of verb diversity, namely CVV1, showed an increase from lower to higher grade levels. These variation patterns indicate the crucial role of verbs in vocabulary learning throughout all 12 grades compared to other word types. As a result, it is suggested that, during the adaptation of texts across the 12 grades of the English curriculum in China, teachers should place special emphasis on enhancing the diversity of verbs. On the other hand, the remaining diversity indices related to specific word types exhibited fluctuations during primary school but did not demonstrate noticeable variation patterns in secondary school. This observation can be explained by the requirements outlined in the English Curriculum (Ministry of

Education, 2012, 2022). Students are introduced to different parts of speech during primary school and are expected to comprehend and acquire the meanings and usages of these different parts of speech in secondary school.

Moreover, this study has discovered that the lexical sophistication of exemplar texts increased steadily and gradually throughout each stage of schooling, as indicated by four root-ratio based indices. This suggests that teachers and experts pay heed to the ratio of sophisticated words or word senses in a text during material selection and adaptation. In particular, primary school demonstrated a slow increase in lexical sophistication, while junior high school demonstrated a noticeable boost. The sharpest increase in lexical sophistication occurred during senior high school. Such a progression also reflects the vocabulary learning requirements of the English Curriculum Standards, i.e., high frequency words are the primary focus in compulsory education (primary school and junior high school) and vocabulary knowledge should be broadened and deepened during senior high school. In addition, sense-aware lexical sophistication increased at a more pronounced rate in senior high school, which may be explained by the inclusion of academic texts in senior high school textbooks (Ministry of Education 2020). While academic texts contain a large number of polysemous words, they might be more sensitive to word senses (Skoufaki and Petric 2021).

6.2 RQ2: distinguishing lexical complexity features across the 12 grades

This study examined the indices that may contribute to differentiating lexical complexity within the 12 grade levels. Although 16 out of 17 lexical complexity indices were found to be significantly different across 12 grades, not all indices could differentiate adjacent grade levels. Four lexical diversity indices and four lexical sophistication indices exhibited large effect sizes for one-way ANOVAs and were able to differentiate the majority of adjacent grade groups in post hoc analyses. Two indices pertaining to the lexical diversity of all types of words, namely the number of different words in a text (NDW) and the root type token ratio (RTTR), may be considered as distinguishing features of lexical complexity in exemplar texts. The findings corroborated previous research findings that NDW, the transformed version of TTR, and the transformed version of the ratio between the number of verb types and verb tokens are reliable metrics to gauge second language writing and speaking quality (Kim 2014; Lu 2012; Lyashevskay et al. 2021; Qin and Uccelli 2020). The study also confirmed that the simple type-token ratio is sensitive to text length (Lu 2012; Lyashevskay et al. 2021) and, therefore cannot be used to distinguish the lexical complexity of teaching materials. Only the corrected verb diversity measure,

CVV1, demonstrated a large effect size and differentiated the majority of adjacent grade groups (with the exception of Grades 2–3 and 10–11) in post hoc tests. Other measures examining the diversity of specific word types, such as lexical words, nouns, adjectives, adverbs, and modifiers, could not distinguish between adjacent grade levels. Similar to previous studies (Kim 2014; Lu 2012; McClure 1991), no conclusive evidence can be obtained regarding the debate over whether the diversity of specific word types can serve as reliable indicators for gauging lexical complexity.

Four root-ratio based lexical sophistication indices could also be regarded as discriminative features. Differences between adjacent grade levels in the four indices were significant in the secondary school levels. The findings corroborate with the L2 writing research of Lu and Hu (2022) in that the four root-ratio indices are good indicators of the L2 writing quality. Moreover, both sense-aware indices performed better than the word-form-based ones in discriminating adjacent grade levels of exemplar texts. Studies that adopted a frequency-based approach discovered that the word frequency level might not be able to discriminate between different levels of teaching materials very well (Chujo 2004; Sun and Yen Dang 2020). The differentiating power of sense-aware lexical sophistication indices in this study indicates the necessity of considering word sense information when preparing materials that are appropriate for learners' proficiency levels (Deane et al. 2006).

6.3 Implications for benchmarking

The findings have immediate implications for benchmarking the lexical complexity of EFL reading materials. As the corpus in this study is comprised of exemplar texts selected from EFL textbooks in China, the three lexical diversity indices and four lexical sophistication indices that effectively differentiate adjacent grade levels can serve as discriminative measurements for assessing lexical complexity. By considering the descriptive results and the differences between grade levels in these measurements, lexical complexity benchmark guidelines can be established. Teachers can use these benchmarks to assess the lexical complexity of their chosen materials, adjust their levels accordingly and make adaptations as needed. Moreover, the identified cut-off frequency of sophisticated words and word senses can aid EFL teachers in identifying sophisticated words and adapting the reading materials to align with appropriate benchmarks. Future studies can also focus on developing computational tools to automate the differentiation and comparison process. Such tools would calculate the lexical complexity features of the input text and compare

them with the established benchmarks. Subsequently, the tool would identify areas that require adaptation and provide suggestions to teachers accordingly.

6.4 Implications for pedagogy

The findings of the present study shed light on the future of EFL pedagogy. First, since measures of both lexical diversity and lexical sophistication vary across grade levels, it is necessary for teachers to adapt the lexical complexity levels of reading materials by considering both dimensions. Second, since the three lexical diversity indices measured here increase gradually across the 12 grades, teachers should always pay close attention to the number and proportion of different words in a text, as well as the proportion of different verbs. Furthermore, when students spend time reading inappropriately difficult texts, they may not benefit academically from the reading (Allington 2002). As the findings indicate that the lexical sophistication indices remain relatively stable in primary school, it might be unnecessary to include sophisticated words or sophisticated word sense in vocabulary teaching. In contrast, English teachers should exercise caution when modifying sophisticated words or word meanings at higher grade levels, which has been the focus of teachers' text adaptation practices (Rets et al. 2022; Young 1999).

Several pedagogical implications specific to higher grade levels are provided. As sense-aware lexical sophistication indices rise sharply in secondary school levels, particularly during senior high school, teachers should be aware of the need to strengthen high school students' mastery of various word senses that they may encounter in particular contexts. The English curriculum also outlines specific requirements for using phrases in specific contexts at both junior and senior high school levels. However, many previous studies have found that students tend to rely on preconceived or biased notions when attempting to comprehend the meaning of a word when reading (Bensoussan and Laufer 1984; Frantzen 2003). As a first step, secondary school teachers must instruct students on the polysemous nature of numerous words. Teachers could, for instance, directly teach students the different meanings and applications of a word and design vocabulary learning tasks that explicitly require students to differentiate between different word senses in context (Skoufaki and Petric 2021). In addition, teachers could teach students how to infer the meaning of a word from the context in reading by using a variety of cues, such as morphological cues and background knowledge (Levesque et al. 2019), to improve their comprehension of word senses in specific contexts (Bensoussan and Laufer 1984; Skoufaki and Petric 2021).

7 Conclusions and future research

This study has examined the lexical sophistication of exemplar texts across the 12 grade levels of China's basic English curriculum. It has further investigated the indices that contribute to the differentiation of lexical complexity across these grade levels, broadening our understanding of the lexical complexity of exemplar texts adapted for EFL in China. The study also provides recommendations for pedagogy, benchmarking and textbook design.

There are several limitations in this study which could be investigated in future research. The first is the singular focus adopted here on the lexical complexity of textbooks without examining the lexical requirements of reading tests. Since it has been determined that there is a vocabulary gap between teaching and testing (Jin et al. 2016), a systematic analysis of the lexical complexity of reading texts could help to identify potential gaps between teaching and testing and provide insights into teaching practices. Secondly, this study does not take into account the variation in lexical complexity across different topics or genres. Since texts of various genres or topics may exhibit different proportions of lexically sophisticated words or word senses, future research could examine the interaction between the topical or genre variable and the lexical complexity measures of EFL learning materials. Despite these limitations, there are important avenues for future exploration and application of this research. For instance, designing sense-labelling systems and integrating lexical complexity indices into an online text adaptation system could assist teachers in adapting texts by automatically grading the lexical complexity level of a text and labeling word senses that are too complex for students.

References

Allaw, Elissa. 2021. A learner corpus analysis: Effects of task complexity, task type, and L1 & L2 similarity on propositional and linguistic complexity. *International Review of Applied Linguistics in Language Teaching* 59(4). 569–604.

Allen, David. 2009. A study of the role of relative clauses in the simplification of news texts for learners of English. *System* 37(4). 585–599.

Allington, Richard L. 2002. What I've learned about effective reading instruction: From a decade of studying exemplary elementary classroom teachers. *Phi Delta Kappan* 83(10). 740–747.

Arya, Diana J., Elfrieda H. Hiebert & P. David Pearson. 2011. The effects of syntactic and lexical complexity on the comprehension of elementary science texts. *International Electronic Journal of Environmental Education* 4(1). 107–125.

Bensoussan, Marsha & Batia Laufer. 1984. Lexical guessing in context in EFL reading comprehension. *Journal of Research in Reading* 7(1). 15–32.

Berendes, Karin, Sowmya Vajjala, Detmar Meurers, Doreen Bryant, Wolfgang Wagner, Maria Chinkina & Trautwein Ulrich. 2018. Reading demands in secondary school: Does the linguistic complexity of textbooks increase with grade level and the academic orientation of the school track? *Journal of Educational Psychology* 110(4). 518–543.

Blanca, María J., Rafael Alarcón Postigo, Jaume Arnau Gras, Roser Bono Cabré & Rebecca Bendayan. 2017. Non-normal data: Is ANOVA still a valid option? *Psicothema* 29(4). 552–557.

Bui, Gavin. 2021. Influence of learners' prior knowledge, L2 proficiency and pre-task planning on L2 lexical complexity. *International Review of Applied Linguistics in Language Teaching* 59(4). 543–567.

Carabantes, Luis & Amos Paran. 2022. 'It may also be our own fault to think so, to limit them before even trying': Assuming learner limitations during materials design in English language teacher education. *Tesol Quarterly* 56(4). 1266–1289.

Carroll, John B. 1964. *Language and thought*. Englewood Cliffs, NJ: Prentice-Hall.

Casanave, Christine Pearson. 1994. Language development in students' journals. *Journal of Second Language Writing* 3. 179–201.

Chall, Jeanne Sternlicht & Edgar Dale. 1995. *Readability revisited: The new Dale-Chall readability formula*. Brookline, MA: Brookline Books.

Chen, Xiaobin & Detmar Meurers. 2018. Word frequency and readability: Predicting the text-level readability with a lexical-level attribute. *Journal of Research in Reading* 41(3). 486–510.

Chujo, Kiyomi. 2004. Measuring vocabulary levels of English textbooks and tests using a BNC lemmatized high frequency word list. *Language and Computers* 51. 231–249.

Claridge, Gillian. 2012. Graded readers: How the publishers make the grade. *Reading.in a Foreign Language* 24(1). 106–119.

Crossley, Scott A., Jerry Greenfield & Danielle S. McNamara. 2008. Assessing text readability using cognitively based indices. *Tesol Quarterly* 42(3). 475–493.

Crossley, Scott A., Tom Salsbury & Danielle S. McNamara. 2012a. Predicting the proficiency level of language learners using lexical indices. *Language Testing* 29(2). 243–263.

Crossley, Scott A., David Allen & Danielle S. McNamara. 2012b. Text simplification and comprehensible input: A case for an intuitive approach. *Language Teaching Research* 16(1). 89–108.

Crossley, Scott A., David F. Dufty, Phillip M. McCarthy & Danielle S. McNamara. 2007. Toward a new readability: A mixed model approach. In *Proceedings of the Annual Meeting of the Cognitive Science Society*, vol. 29, 197–202.

Crossley, Scott A., Tom Salsbury, Danielle S. McNamara & Jarvis. Scott. 2011. Predicting lexical proficiency in language learner texts using computational indices. *Language Testing* 28(4). 561–580.

Cumming, Alister, Robert Kantor, Kyoko Baba, Keare Eouanzoui, Usman Erdosy & Mark Jamse. 2005. Analysis of discourse features and verification of scoring levels for independent and integrated prototype written tasks for the new TOEFL®. ETS Research Report Series. i–77.

Daller, Helmut, Roeland Van Hout & Jeanine Treffers-Daller. 2003. Lexical richness in the spontaneous speech of bilinguals. *Applied Linguistics* 24(2). 197–222.

Davies, Mark. 2012. Expanding horizons in historical linguistics with the 400-million-word Corpus of Historical American English. *Corpora* 7(2). 121–157.

Deane, Paul, Kathleen M. Sheehan, John Sabatini, Yoko Futagi & Irene Kostin. 2006. Differences in text structure and its implications for assessment of struggling readers. *Scientific Studies of Reading* 10(3). 257–275.

Droop, Mienke & Ludo Verhoeven. 1998. Background knowledge, linguistic complexity, and second-language reading comprehension. *Journal of Literacy Research* 30(2). 253–271.

Engber, Cheryl A. 1995. The relationship of lexical proficiency to the quality of ESL compositions. *Journal of Second Language Writing* 4(2). 139–155.

Ferguson, Christopher J. 2009. An effect size primer: A guide for clinicians and researchers. *Professional Psychology: Research and Practice* 40. 532–538.

Frantzen, Diana. 2003. Factors affecting how second language Spanish students derive meaning from context. *The Modern Language Journal* 87(2). 168–199.

González-Fernández, Beatriz & Norbert Schmitt. 2020. Word knowledge: Exploring the relationships and order of acquisition of vocabulary knowledge components. *Applied Linguistics* 41(4). 481–505.

Green, Anthony & Hawkey Roger. 2012. Re-Fitting for a different purpose: A case study of item writer practices in adapting source texts for a test of academic reading. *Language Testing* 29(1). 109–129.

Guiraud, Pierre. 1960. *Problémes et méthodes de la statistique linguistique* [*Problems and methods of statistical linguistics*]. Dordrecht, The Netherlands: D. Reidel.

Harley, Birgit & Mary Lou King. 1989. Verb lexis in the written compositions of young L2 learners. *Studies in Second Language Acquisition* 11. 415–440.

Hiebert, Elfrieda H., Amanda P. Goodwin & Gina N. Cervetti. 2018. Core vocabulary: Its morphological content and presence in exemplar texts. *Reading Research Quarterly* 53(1). 29–49.

Hiebert, Elfrieda H. & Heidi Anne E. Mesmer. 2013. Upping the ante of text complexity in the common core state standards: Examining its potential impact on young readers. *Educational Researcher* 42(1). 44–51.

Hill, David, R. 2008. Graded readers in English. *ELT Journal* 62(2). 184–204.

Housen, Alex, Els Schoonjans, Sonja Janssens, Aurélie Welcomme, Ellen Schoonheere & Michel Pierrard. 2011. Conceptualizing and measuring the impact of contextual factors in instructed SLA–the role of language prominence. *International Review of Applied Linguistics in Language Teaching* 49(2). 83–112.

Hu, Renfen, Li Shen & Shichen Liang. 2019. Diachronic sense modeling with deep contextualized word embeddings: An ecological view. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. 3899–3908. Association for Computational Linguistics.

Iwashita, Noriko, Annie Brown, Tim McNamara & O'Hagan Sally. 2008. Assessed levels of second language speaking proficiency: How distinct? *Applied Linguistics* 29(1). 24–49.

Jarvis, Scott. 2013. Defining and measuring lexical diversity. In Jarvis Scott & Michael Daller (eds.), *Vocabulary knowledge: Human ratings and automated measures*, 13–44. Amsterdam & Philadelphia: John Benjamins.

Jin, Tan, Kai Guo, Barley Mak & Qiuping Wu. 2017. Lexical profiles of reading texts in high-stakes tests: Where are the benchmarks? *International Journal of Computer-Assisted Language Learning and Teaching* 7(1). 34–49.

Jin, Tan, Yunting Li & Baichuan Li. 2016. Vocabulary coverage of reading tests: Gaps between teaching and testing. *TESOL Quarterly* 50(4). 955–964.

Jin, Tan & Xiaofei Lu. 2018. A data-driven approach to text adaptation in teaching material preparation: Design, implementation, and teacher professional development. *Tesol Quarterly* 52(2). 457–467.

Jin, Tan., Xiaofei Lu & Ni. Jing. 2020. Syntactic complexity in adapted teaching materials: Differences among grade levels and implications for benchmarking. *The Modern Language Journal* 104(1). 192–208.

Kim, Ji-Young. 2014. Predicting L2 writing proficiency using linguistic complexity measures: A corpus-based study. *English teaching* 69(4). 27–50.

Kincaid, J. Peter, Robert P. Fishburne, Richard L. Rogers & Brad S. Chissom. 1975. *Derivation of new readability formulas (Automated Readability Index, Fog Count and Flesch Reading Ease Formula) for Navy enlisted personnel, Research branch report 8–75, millington, TN: Naval technical training*. Memphis, TN: U. S. Naval Air Station.

Klee, Thomas. 1992. Developmental and diagnostic characteristics of quantitative measures of children's language production. *Topics in Language Disorders* 12. 28–41.

Kyle, Kristopher & Scott A. Crossley. 2015. Automatically assessing lexical sophistication: Indices, tools, findings, and application. *Tesol Quarterly* 49(4). 757–786.
Levin, Joel R. 1967. Misinterpreting the significance of "explained variation". *American Psychologist* 22. 675–676.
Levesque, Kyle C., Michael J. Kieffer & S. Hélène Deacon. 2019. Inferring meaning from meaningful parts: The contributions of morphological skills to the development of children's reading comprehension. *Reading Research Quarterly* 54(1). 63–80.
Li, Jiao, Xuesong Gao & Xuehai Cui. 2021. Language teachers as materials developers. *RELC Journal*. https://doi.org/10.1177/0033688221106162.
Linnarud, Moria. 1987. Lexis in composition: A performance analysis of Swedish learners' written English. *Studies in Second Language Acquisition* 9. 254–256.
Lu, Xiaofei. 2012. The relationship of lexical richness to the quality of ESL learners' oral narratives. *The Modern Language Journal* 96(2). 190–208.
Lu, Xiaofei & Renfen Hu. 2022. Sense-aware lexical sophistication indices and their relationship to second language writing quality. *Behavior Research Methods* 54(3). 1444–1460.
Lyashevskaya, Olga, Irina Panteleeva & Olga Vinogradova. 2021. Automated assessment of learner text complexity. *Assessing Writing* 49. 1–16.
McCarthy, Philip M. 2005. *An assessment of the range and usefulness of lexical diversity measures and the potential of the measure of textual, lexical diversity (MTLD)* [*Microfiche*]. Doctoral Dissertation, University of Memphis.
McClure, Erica. 1991. A comparison of lexical strategies in L1 and L2 written English narratives. *Pragmatics and Language Learning* 2. 141–154.
Miller, Jon F. (1991). Quantifying productive language disorders. In Jon F. Miller (ed.), *Research in child language disorders: A decade of progress*, 211–220. Austin, TX: Pro-Ed.
Ministry of Education, China. 2003. *English curriculum standards for high school* (*experimental version*). Beijing, China: People's Education Press.
Ministry of Education, China. 2012. *English curriculum standards for compulsory education*. Beijing, China: Beijing Normal University Press.
Ministry of Education, China. 2019. *Regulations on textbook management for primary and secondary schools*. http://www.moe.gov.cn/srcsite/A26/moe_714/202001/t20200107_414578.html (accessed 9 May 2021).
Ministry of Education, China. 2020. *General Senior high School curriculum Standards 2017 edition*. Beijing, China: People's Education Press.
Ministry of Education, China. 2022. *English curriculum standards for compulsory education* (*2022 version*). Beijing, China: Beijing Normal University Press.
Morris, Lori & Tom Cobb. 2004. Vocabulary profiles as predictors of the academic performance of Teaching English as a Second Language trainees. *System* 32(1). 75–87.
Noone, Richard L. 2018. Developing texts that match readers' abilities in the sciences: Seeking an appropriate instructional level. In Rahma Al-Mahrooqi & Christopher Denman (eds.), *English education in Oman: Current scenarios and future trajectories*, 35–57. Singapore: Springer.
Oh, Sun-Young. 2001. Two types of input modification and EFL reading comprehension: Simplification versus elaboration. *TESOL Quarterly* 35(1). 69–96.
Olejnik, Stephen & Algina James. 2000. Measures of effect size for comparative studies: Applications, interpretations, and limitations. *Contemporary Educational Psychology* 25(3). 241–286.
Qin, Wenjuan & Paola Uccelli. 2020. Beyond linguistic complexity: Assessing register flexibility in EFL writing across contexts. *Assessing Writing* 45. 1–14.

Rayner, Keith & Susan A. Duffy. 1986. Lexical complexity and fixation times in reading: Effects of word frequency, verb complexity, and lexical ambiguity. *Memory & Cognition* 14(3). 191–201.

Rets, Irina & Jekaterina Rogaten. 2021. To simplify or not? Facilitating English L2 users' comprehension and processing of open educational resources in English using text simplification. *Journal of Computer Assisted Learning* 37(3). 705–717.

Rets, Irina, Lluisa Astruc, Tim Coughlan & Ursula Stickler. 2022. Approaches to simplifying academic texts in English: English teachers' views and practices. *English for Specific Purposes* 68. 31–46.

Ross, Steven, Michael H. Long & Yasukata Yano. 1991. Simplification or elaboration? The effects of two types of text modifications on foreign language reading comprehension. University of Hawai'i Working Papers in ESL, vol. 10, 1–32.

Simensen, A. Marit. 1987. Adapted readers: How are they adapted? *Reading in a Foreign Language* 4(1). 41–57.

Skoufaki, Sophia & Bojana Petrić. 2021. Academic vocabulary in an EAP course: Opportunities for incidental learning from printed teaching materials developed in-house. *English for Specific Purposes* 63. 71–85.

Sun, Ye & Thi Ngoc Yen Dang. 2020. Vocabulary in high-school EFL textbooks: Texts and learner knowledge. *System* 93. 1–13.

Sung, Yao–Ting, Wei–Chun Lin, Scott Benjamin Dyson, Kuo–En Chang & Yu–Chia Chen. 2015. Leveling L2 texts through readability: Combining multilevel linguistic features with the CEFR. *The Modern Language Journal* 99(2). 371–391.

Templin, Mildred. 1957. *Certain language skills in children: Their development and interrelationships*. Minneapolis: The University of Minnesota Press.

To, Vinh, Si Fan & Damon Thomas. 2013. Lexical density and readability: A case study of English textbooks. *Internet Journal of Language, Culture and Society* 37. 61–71.

Ure, Jean. 1971. Lexical density and register differentiation. *Applications of Linguistics* 23(7). 443–452.

Vicente, Agustin. 2018. Polysemy and word meaning: An account of lexical meaning for different kinds of content words. *Philosophical Studies* 175(4). 947–968.

Young, Dolly N. 1999. Linguistic simplification of SL reading material: Effective instructional practice? *The Modern Language Journal* 83(3). 350–366.