Tabaran Institute of Higher Education
International Journal of Language Testing
Vol. 11, No. 2, October 2021

ISSN 2476-5880

Text Complexity of Reading Comprehension Passages in the National Matriculation English Test in China: The Development from 1996 to 2020

Xiaoli Yu[1]

Received: 7 June 2021 Accepted: 9 August 2021

Abstract

This study examined the development of text complexity for the past 25 years of reading comprehension passages in the National Matriculation English Test (NMET) in China. Text complexity of 206 reading passages at lexical, syntactic, and discourse levels has been measured longitudinally and compared across the years. The natural language processing tools used in the study included TAALES, TAALED, TAASSC, and TAACO. To compare the differences across the years at various levels of text complexity, ANOVA and MANOVA tests were conducted. The results suggested that lexical level text complexity revealed the most evident changes throughout the years, lexical sophistication, density, and diversity levels of the most recent years of reading passages have increased remarkably compared to the early years. The syntactic level text complexity indicated a moderate elevation toward the recent years of reading passages. For the discourse level text complexity, regarding cohesion, insignificant fluctuation occurred throughout the years and the general trend was not necessarily increasing. Combined, the results indicated that text complexity of the reading comprehension passages in the NMET over the past 25 years had been steadily increasing by including more low frequency and academic vocabulary, diversifying vocabulary in the passages, and complicating sentence and grammatical structures. The results were further examined against the general curriculum standards and guidelines to analyze whether the changes were reflected in the policies. It showed that the exams required a much larger vocabulary size than the number indicated in the guidelines, suggestions for test designers and pedagogical practices were provided accordingly.

Keywords: Corpus Linguistics; High Stakes Exam; Natural Language Processing; Reading Comprehension; Text Complexity

1. Introduction

Standardized high-stakes exams have been widely used in China to assess learners' English proficiency and play as the key gatekeeper for their academic development. The National Higher Education Entrance Examination, known as *Gaokao* in Chinese, is taken annually by

[1] Department of Foreign Language Education, Faculty of Education, Middle East Technical University, Turkey. Email: xiaoli@metu.edu.tr

millions of high school graduates across China. As the major gatekeeper, *Gaokao* is the most visible and important exam in China which predominately determines whether a high school graduate is able to continue his/her higher education in a prestige university (Qi, 2007). For most provinces in China, the current structure of *Gaokao* can be described as "3+X." The 3 refers to the three compulsory subjects that each testee needs to take, namely Chinese, Mathematics, and English. The X differs according to the testee's disciplinary choice. For testees pursuing the Humanities stream, the subjects in the X are History, Politics, and Geography; whereas for the Sciences stream, the subjects in the X are Physics, Chemistry, and Biology. In sum, most testees need to take six subject exams in *Gaokao* in total.

Regarding the National Matriculation English Test (NMET), as one of the three compulsory exams in *Gaokao*, it plays a considerable role in influencing one's *Gaokao* final result. Although some provinces implement different exam structures depending on the local educational circumstances, the major content of the NMET is uniformly prescribed by China's Ministry of Education. The National Education Examinations Authority (NEEA), operating directly under the Ministry of Education, issues a guideline for the NMET each year, which provides an essential guide about the exam for testees' preparation as well as authorities of individual provinces in designing their own exams (Cheng & Qi, 2006; Farley & Yang, 2020). Throughout the years, the NMET has been gradually reformed toward the direction of examining learners' communicative skills. For instance, listening has been added as an essential component since 1999 (Cheng & Qi, 2006). In addition, despite a relatively low share of the total score, speaking has been adopted as a separate section beyond the written sections in different cities and provinces (Cheng, 2008). By reforming and diversifying the content of the NMET, the intention of the test constructors is to improve Chinese English learners' language-use ability rather than mere linguistic knowledge of English (Ministry of Education, 2017; Qi, 2007); nevertheless, as *Gaokao* has been remaining as the predominant gatekeeper for most students to pursue higher education and testing grammatical knowledge on paper-based exams continues to take up the major weighting in the NMET, the reality of teaching and learning to the test and focusing on rote memorization stays in the practice of English education in China (Qi, 2005; Yan, 2015).

Despite various reforms of the NMET, reading comprehension has always been an essential component of the exam, which takes approximately 25 – 30% of the total points. According to the general guideline from the National Education Examinations Authority (NEEA, 2019a), designing the exams in *Gaokao* should be based on specific requirements from different universities and the national curriculum standards. The major objectives suggested in the general guideline include being able to understand common topics like announcement, instruction, and advertisement in books, newspapers, and magazines. The testees are expected to obtain useful information from the passage to understand the main ideas and specific details, grasp the major structure, comprehend the author's purpose, attitude, and opinions, and infer meanings of specific vocabulary and phrases (NEEA, 2019b). Regarding the General High School English Curriculum Standards, the 2003 and 2017 editions both confirm these objectives (Ministry of Education, 2003, 2017).

Although the NMET has been acting as the baton of English teaching and learning in China and understanding the longitudinal test development may greatly influence pedagogical practices, few empirical studies have investigated the development of the reading comprehension passages included in the exam from a longitudinal perspective. In particular, to the best of the researcher's knowledge, no study has used natural language processing tools to analyze the changes of the text complexity of the passages via a systematic and comprehensive manner, including longitudinal analyses at lexical, syntactic, and discourse levels. Therefore, employing corpus-based analyzing methods, the current study aims to provide a comprehensive picture of how the text complexity of the reading comprehension passages in the NMET has changed throughout the past 25 years, from 1996 to 2020. The results of the study are expected to fill the gap in relevant research fields and offer insights for test designing and pedagogical activities. The following research questions led to the investigation of the study:

1. How has the text complexity of the reading comprehension passages at the lexical level changed over the past 25 years (i.e., 1996-2020)?
2. How has the text complexity of the reading comprehension passages at the syntactic level changed over the past 25 years (i.e., 1996-2020)?
3. How has the text complexity of the reading comprehension passages at the discourse level changed over the past 25 years (i.e., 1996-2020)?

2. Review of Literature

In this section, previous studies that are related to text complexity and corpus-based measurement of linguistic features are introduced.

2.1. Reading Comprehension, Readability, and Text Complexity

The widely accepted RAND[1] Reading Study Group (Snow, 2002) model of reading comprehension suggests that embedded in a sociocultural context, reader, activity, and text are the three major components that crucially influence comprehension outcome. In particular, features of the text, such as the vocabulary load, linguistic structure, discourse style, and genre, may have a direct impact on readers' constructing different representations embedded within the text, including the surface code (i.e., the exact wording of the text), the text base (i.e., idea units representing the meaning of the text), and the mental models (i.e., the way in which information is processed for meaning). This notion is in line with Alderson's (2000) view that linguistic features of a text, text length, type, organization, genre, and so on all affect readers' comprehension. In addition, Hornof (2008) and Guthrie, Klauda, and Ho (2013) note that for L2 reading, the text itself, including characteristics of rhetoric, genre, and text complexity, plays a critical role in influencing readers' understanding, especially for intermediate-level readers. Hence, in order to better comprehend the relationship between text, reader, and task, understanding text complexity should be a priority for research in L2 reading comprehension.

Readability assessment has been developed based on text complexity in order to determine the difficulty of a text through automated means rather than mere human judgment. A large number of readability formula have been created since the beginning of the last century. Traditional readability formulas tend to examine easily measured units such as sentence length, word length, and word frequency; whereas more recent readability formulas are inspired by cognitive theory and incorporate measures revealing the relationships between elements in a text rather than only the counts of surface level measures (Benjamin, 2012). For instance, the readability formula proposed by Crossley, Greenfield, and McNamara (2008) incorporates measures regarding vocabulary frequency, similarity of syntax across sentences, and referential cohesion. Research synthesis suggests that although some traditional readability measures may still work well for typical texts, with more types of variables at different levels, newly developed readability formulas and analysis tools have proven to be more reliable and valid for wide-ranging ages and abilities (Benjamin, 2012). Thus, it is crucial to conduct more comprehensive analyses that measure different aspects of a text to better reveal its readability.

Different from text difficulty which depends on readers' performance of a task, text complexity can be understood as the independent variables of a text that can be analyzed, studied, or manipulated (Mesmer, Cunningham, & Hiebert, 2012). A large number of empirical studies have consistently revealed the strong relationship between vocabulary features of a text and reading comprehension, readers' vocabulary size and coverage of the entire words in a text are among the strongest indicators of reading comprehension outcome (Wright & Cervetti, 2017). In a similar way, syntactic features also influence readers' processing of texts as more complex sentences and grammatical structures make the sentences more difficult to be parsed (Mesmer et al., 2012; Kyle, 2016). Shiotsu and Weir's (2007) study suggests that readers' syntactic knowledge actually accounts more for the variance in L2 English reading test results compared to lexical knowledge. Therefore, the examination of linguistic features at both lexical and syntactic levels should be of critical consideration regarding text complexity analysis. With respect to discourse level textual characteristics, among different variables, cohesion has continuously been considered as a key property and extensively researched especially with the help of computer-assistant technology. Research supports that more cohesive texts are easier to be comprehended (Gernsbacher, 2013; Graesser, McNamara, & Louwerse, 2003). Thus, with strong empirical evidence in terms of the roles of lexical and syntactic level textual features as well as cohesion at discourse level in shaping text complexity and influencing reading comprehension, the current study examines these target linguistic features of the reading comprehension passages in the past 25 years of the NMET in order to demonstrate the longitudinal development of text complexity and shed light on reading comprehension test designing and pedagogical practices.

2.2. Corpus-based Measurements of Linguistic Features

In terms of linguistic features at the lexical level, lexical sophistication, diversity, density, and errors are the major properties that have been examined widely in various studies (Read, 2000). Relevant to the purpose of the current study, levels of lexical sophistication, diversity, and density of the reading passages are determined to be computed to conduct the target comparison across the years. Regarding lexical sophistication, as it measures the percentage of advanced vocabulary in a text, most previous studies have considered frequency features of the words in a text as the principal factor determining the lexical sophistication level of the text (Laufer & Nation, 1995); in this regard, the more frequent the vocabulary of a text is, the less sophisticated the text is regarding lexis. Laufer (1994), Laufer and Nation (1995), and Nation and Waring (1997) employed frequency bands in revealing lexical sophistication levels of texts. Currently, VocabProfilers in Compleat Lexical Tutor (https://www.lextutor.ca/) is a common venue for conducting analysis based on frequency bands, the result reveals the percentage of different word frequency bands of a text (Cobb, n.d.). Nevertheless, measurement of lexical sophistication has developed from merely focusing on word frequency bands to frequency counts to word range, cognitive and psycholinguistic features of the vocabulary items. Kim, Crossley, and Kyle (2018) extended the analysis of lexical sophistication beyond word frequency by employing the Tool for the Automatic Analysis of Lexical Sophistication (TAALES; Kyle & Crossley, 2015). Through TAALES, Kim et al.'s analysis included relevant domains of lexical sophistication such as word range, contextual distinctiveness, word neighborhood, academic language, and so forth.

For lexical diversity, the construct is related to the range of vocabulary and avoidance of repetition (Read, 2000). The ratio between the number of different words and the total number of words (i.e., type-token ratio, TTR) is the traditional measurement of lexical diversity. To avoid the influence of text length on the final TTR result, several other more advanced mathematic models were developed based on TTR, for example, vocd-D by Durán, Malvern, Richards, and Chipere (2004) and MTLD by McCarthy and Jarvis (2010). Most recently, the Tool for the Automatic Analysis of Lexical Diversity (TAALED) unites a range of measurements of lexical diversity, including the classic TTR and other more robust indices such as MTLD, MATTR, HD-D, and so on (Kyle, Crossley, & Jarvis, 2020). Finally, regarding lexical density, another important indicator of text complexity (Fang & Pace, 2013), it refers to the percentage of content words in a text, higher lexical density contributes to the increasing difficulty in comprehension (Read, 2000). TAALED also calculates lexical density levels for both types and tokens.

Common measures of syntactic complexity include mean length of clause (MLC), T-unit (MLTU), sentence (MLS), which measure the average number of words per clause, T-unit, and sentence respectively; also, the number of complex and elaborated clausal and phrasal structures (e.g., dependent clauses, complex T-units, complex nominals, verb phrases) per clause, T-unit and sentence have also been involved in measuring syntactic complexity in various studies (Biber, Gray, & Poonpon, 2011; Lu, 2011; Ortega, 2003; Wolfe-Quintero, Inagaki, & Kim, 1998). Regarding the corpus-based automatic analysis of syntactic complexity, Coh-Metrix (Graesser, McNamara, Louwerse, & Cai, 2004) is a computational system that has been widely used in different studies to calculate various indices of syntactic complexity. Some indices of syntactic complexity that Coh-Metrix calculates are the number of words before the main verb, modifiers per noun phrase, the incidence of all clauses, subjective relative clauses, -that verb component, and so forth. In addition, the L2 Syntactic Complexity Analyzer (L2SCA; Lu, 2010) incorporates 14 syntactic complexity measures categorized into five types: (a) length of production unit; (b) sentence complexity ratio; (c) the amount of subordination; (d) the amount of coordination; and (e) particular syntactic structures. Most recently, Kyle (2016) brought up the necessity of examining syntactic sophistication, namely the relative complexity, together with the traditional measures of syntactic complexity (i.e., absolute complexity, Bulté

& Housen, 2012) to reveal a fuller picture of syntactic characteristics of a text. Kyle's (2016) Tool for the Automatic Analysis of Syntactic Sophistication and Complexity (TAASC) includes the 14 L2SCA indices as well as fine-grained clausal and phrasal complexity and syntactic sophistication measures. In TAASC, 190 indices regarding syntactic sophistication are developed according to user-based theories of language acquisition.

Lastly, cohesion refers to the specific elements of a text that indicates the coherent feature of the text and facilitates readers' comprehension (Graesser et al., 2003; Louwerse, 2004). Cohesive devices are common elements that contribute to the cohesion of a text. Compared to the manual approach in measuring cohesion, computational approaches present less fallibility of hand counts and the subjective nature of intuitive judgment (Crossley & McNamara, 2009). Aforementioned Coh-Metrix (Graesser et al., 2004) is one of the most commonly accepted tools analyzing lexical, syntactic, and semantic properties of texts that are related to cohesion. Coh-Metrix is built upon various existing resources and databases, including WordNet (Miller, Beckwith, Fellbaum, Gross, & Miller, 1990), the MRC Psycholinguistics Database (Coltheart, 1981), and the CELEX Database (Baayen, Piepenbrock, & van Rijn, 1993). However, Coh-Metrix has limitations such as limited number of cohesion indices and a lack of batch processing. The Tool for the Automatic Analysis of Cohesion (TAACO) developed by Crossley, Kyle, and McNamara (2016) is a relatively newly developed text cohesion analysis tool that allows for batch processing and incorporates more than 150 indices to examine text cohesion. The measures of cohesion that TAACO provides include local (i.e., sentence-level), global (paragraph-level), and overall text (i.e., text-level) cohesion.

3. Method

This section introduces the corpus used for the current study and the detailed procedures used for analyzing the textual data.

3.1. Corpus

The corpus used in this study comprises all of the reading comprehension passages in the national version of the National Matriculation English Test (NMET) from 1996 to 2020. The National Education Examinations Authority (NEEA), operating directly under the Ministry of Education, has been in charge of designing the NMET. Although for certain relatively developed regions and provinces, the provincial education authorities design and administer province-specific exams based on the guideline provided by the NEEA, the nation-wide version of the NMET has been employed by the largest number of provinces. For certain years, the NEEA has designed more than one set of the NMET to serve the needs of multiple provinces. For each set of the NMET, the reading comprehension component contains four to five short passages following by multiple-choice questions. In general, the reading comprehension section takes 25 – 30% of the total score. Each year, after the nationwide university entrance exam, different media and platforms publish the exam to the public. Thus, most of the exams are able to be retrieved freely from the Internet. In total, a total of 206 passages from 46 sets of the NMET were collected from the past 25 years to form the corpus for the current study[2] (Table 1).

Table 1.
Corpus of reading comprehension passages from the NMET, 1996-2020.

Year	# of	# of	Year	# of	# of	Year	# of	# of
	Exams	passages		Exams	passages		Exams	passages
1996	1	4	2006	2	5+5	2016	3	4+4+4
1997	1	5	2007	2	5+5	2017	3	4+4+4
1998	1	5	2008	2	5+5	2018	3	4+4+4
1999	1	5	2009	2	5+5	2019	3	4+4+4
2000	1	5	2010	2	5+5	2020	3	4+4+4
2001	1	5	2011	2	5+4			
2002	1	5	2012	2	4+5	Total # of exams: 46		
2003	1	5	2013	2	4+4	Total # of passages: 206		
	2	5+5		2	4+4	Tokens per passage: 132 – 575		
2004			2014			words		
2005	1	5	2015	2	4+4	Total tokens: 55291		

3.2. Data Analysis

3.2.1. Lexical level text complexity. Based on the existing measurements, in the current study, the frequency bands of the words in each reading passage were firstly examined to reveal the percentage of high-frequency words in each text and the variation throughout the years. VocabProfilers in Compleat Lexical Tutor (https://www.lextutor.ca/vp/comp/) was employed to examine the percentage of each frequency band in each passage based on the BNC-COCA 1-25K word frequency lists (Cobb, n.d.). Considering the nature of the target corpus, the cutting line was set as the 10th most frequent 1000 words (i.e., 10K). A series of ANOVA tests were conducted to compare the differences in text coverages throughout the past 25 years of exams.

Following the frequency bands, the frequency counts method was employed to further examine the lexical sophistication levels of the passages. The freely accessible Tool for the Automatic Analysis of Lexical Sophistication (TAALES) was used to investigate the relevant indices of lexical sophistication through a more fine-grained approach (Kyle & Crossley, 2015). TAALES not only calculates the frequency-based indices of the target text compared to a reference corpus, but also provides other measures of lexical sophistication that may influence the level of text complexity. The major index types in TAALES 2.0 include word frequency, word range, psycholinguistic word information, age of acquisition/exposure, academic words, contextual distinctiveness, word recognition norms, semantic network, N-gram frequency and range, N-gram strength of association, word neighbors, and others (Kyle, Crossley, & Berger, 2018). Multiple empirical studies have been conducted to reveal the reliability and validity of this tool (e.g., Balyan et al., 2019; Crossley, Skalicky, Kyle, & Monteiro, 2019). Including all of the 424 indices of TAALES for investigation is beyond the scope of the current study; thus, based on the needs of the current study and nature of the corpus[3], the following indices presented in Table 2 were selected to further examine the lexical sophistication levels of the target reading comprehension passages. After obtaining the initial results from TAALES, a

MANOVA test was conducted to compare the mean differences occurred throughout the past 25 years for each index.

Table 2.
Lexical sophistication measures from TAALES.

Category	Index Name	Description
Word frequency	BNC_Written_Freq_AW_Log	BNC Written Frequency AW Logarithm
	BNC_Written_Freq_CW_Log	BNC Written Frequency CW Logarithm
	BNC_Written_Freq_FW_Log	BNC Written Frequency FW Logarithm
Word range	BNC_Written_Range_AW	BNC Written Range AW
	BNC_Written_Range_CW	BNC Written Range CW
	BNC_Written_Range_FW	BNC Written Range FW
Academic language	All_AWL_Normed	Academic Word List All
Word recognition norms	LD_Mean_RT_Zscore	Lexical Decision Time (z-score)
	LD_Mean_Accuracy	Lexical Decision Accuracy
	WN_Zscore	Word Naming Response Time (z-score)
	WN_Mean_Accuracy	Word Naming Response Accuracy
Contextual distinctiveness	lsa_average_all_cosine	LSA Contextual Distinctiveness (all cosine)
Age of acquisition/exposure	aoe_inverse_average	LDA Age of Exposure (inverse average)

Note. AW = all words; CW = content words; FW = function words.

To examine the lexical diversity level of the reading passages, the freely accessible Tool for the Automatic Analysis of Lexical Diversity (TAALED) was used. According to Zenker and Kyle's (2021) investigation of the influence of text length on the lexical diversity indices, MATTR, MTLD Original, and MTLD-MA-Wrap presented the smallest degree of text length effect and were recommended for examining short texts. Thus, results of these three indices for all words, content words, and function words were obtained for further comparisons across the years. In addition, TAALED also calculates lexical density levels for both types and tokens. In sum, Table 3 presents the selected indices to reveal the lexical diversity and density levels of the target corpus. A MANOVA test was conducted to compare the changes of these measurements across the five-year intervals from 1996 to 2020.

Table 3.
Lexical density and diversity measures from TAALED.

Category	Index Name	Description
Lexical density	lexical_density_types	The number of content word types divided by the total number of word types
	lexical_density_tokens	The number of content word tokens divided by the total number of tokens
MATTR	mattr50_aw	Moving average type-token ratio (50-word window)
	mattr50_cw	Moving average type-token ratio for content words (50-word window)
	mattr50_fw	Moving average type-token ratio for function words (50-word window)
MTLD original	mtld_original_aw	MTLD is based on the average number of tokens it takes to reach a given TTR value (.720).
	mtld_original_cw	MTLD is based on the average number of content word tokens it takes to reach a given TTR value (.720).
	mtld_original_fw	MTLD is based on the average number of function word tokens it takes to reach a given TTR value (.720).
MTLD-MA-Wrap	mtld_ma_wrap_aw	A version of MTLD (all words) that takes a moving-average approach to calculate the index. The final factor is calculated by wrapping back to the beginning of the text.
	mtld_ma_wrap_cw	A version of MTLD (content words) that takes a moving-average approach to calculate the index. The final factor is calculated by wrapping back to the beginning of the text.
	mtld_ma_wrap_fw	A version of MTLD (function words) that takes a moving-average approach to calculate the index. The final factor is calculated by wrapping back to the beginning of the text.

3.2.2. Syntactic level text complexity. To measure the complexity of the reading comprehension passages at the syntactic level, the Tool of the Automatic Analysis of Syntactic Sophistication and Complexity (TAASSC) was employed as it measures both classic indices of syntactic complexity and fine-grained indices of phrasal and clausal complexity (Kyle, 2016). To control the scope and operability of the current study, first, the 14 indices from the classic measurement of syntactic complexity (i.e., L2SCA, Lu, 2010) were examined and the results were compared across the years (see Table 4). Second, selected indices that measure clausal and phrasal complexity levels were examined to further reveal the syntactic level complexity of the reading passages. Due to the large number of indices provided by TAASSC and the limited scope of the current study, only variables that met the assumption of normality were included in the examination, Table 5 presents the final indices.

Table 4.
Syntactic complexity measures from L2SCA (Lu, 2010, p. 479).

Measure	Code	Description
Type 1: Length of the production unit		
Mean length of clause	MLC	Number of words per clause
Mean length of sentence	MLS	Number of words per sentence
Mean length of T-unit	MLT	Number of words per T-unit
Type 2: Sentence complexity		
Sentence complexity ratio	C/S	Number of clauses per sentence
Type 3: Subordination		
T-unit complexity ratio	C/T	Number of clauses per T-unit
Complex T-unit ratio	CT/T	Number of complex T-units divided by T-units
Dependent clause ratio	DC/C	Number of dependent clauses per clause
Dependent clause per T-unit	DC/T	Number of dependent clauses per T-unit
Type 4: Coordination		
Coordinate phrases per clause	CP/C	Number of coordinate phrases per clause
Coordinate phrases per T-unit	CP/T	Number of coordinate phrases per T-unit
Sentence coordination ratio	T/S	Number of T-units per sentence
Type 5: Particular structures		
Complex nominals per clause	CN/C	Number of complex nominals per clause
Complex nominals per T-unit	CN/T	Number of complex nominals per T-unit
Verb phrases per T-unit	VP/T	Number of verb phrases per T-unit

Table 5.
Syntactic sophistication and complexity measures from TAASSC.

Category	Index Name	Description
Clause Complexity	aux_per_cl	auxiliary verbs per clause
	nsubj_per_cl	nominal subjects per clause
	cl_av_deps	dependents per clause
	av_dobj_deps	dependents per direct object
	av_dobj_deps_NN	dependents per direct object (no pronouns)
	av_pobj_deps_NN	dependents per object of the preposition (no pronouns)
	det_dobj_deps_struct	determiners per direct object
Noun Phrase Complexity	det_dobj_deps_NN_struct	determiners per direct object (no pronouns)
	prep_pobj_deps_NN_struct	prepositions per object of the preposition (no pronouns)

3.2.3. Discourse level text complexity. The tool selected for the current study is the Tool for the Automatic Analysis of Cohesion (TAACO). Developed based on Coh-Metrix (Graesser et al., 2004), TAACO is also freely accessible, it computes 150 classic and recently developed indices that measure local, global, and overall text cohesion (Crossley et al., 2016). Including all of the indices for analysis was beyond the scope of the current study, therefore, only indices met the assumptions of normality and homogeneity were included for the comparison analysis across the years. In addition, since the writing format of many reading passages (e.g., advertisement, poster) does not demonstrate the traditional concept of paragraphs, the indices examining paragraph-level cohesion were also excluded from the current study. The final selected indices for assessing cohesion are displayed in Table 6. After obtaining the initial results of the indices from TAACO, a MANOVA test was again carried out to compare the differences across the five-year intervals over the past 25 years.

Table 6.
Cohesion measures from TAACO.

Category	Index Name	Description
Lexical overlap (sentence)	adjacent_overlap_2_all_sent	number of lemma types that occur at least once in the next two sentences
	adjacent_overlap_binary_argument_sent	number of sentences with ANY noun and pronoun lemma overlap with next sentence
	adjacent_overlap_2_argument_sent	number of noun and pronoun lemma types that occur at least once in the next two sentences
Semantic overlap	lsa_2_all_sent	Average latent semantic analysis cosine similarity between all adjacent sentences (with a two-sentence span).
Connectives	basic_connectives	number of basic connectives (e.g., for, and, or)
	determiners	number of determiners (e.g., a, an, the)
	all_additive	number of additive connectives (e.g., after all, again, all in all)
	all_positive	number of positive connectives (e.g., actually, after, again)
	all_connective	number of all connectives (actually, admittedly, after)
Giveness	repeated_content_lemmas	number of repeated content lemmas divided by number of words
	repeated_content_and_pronoun_lemmas	number of repeated content and third person pronouns divided by number of words

4. Results

In this section, the results for the three research questions are addressed and discussed.

4.1. Research Question 1: Lexical Level Text Complexity

4.1.1. Lexical sophistication: frequency bands. To compare the changes occurred throughout the past 25 years, the data were organized into five-year intervals to conduct the ANOVA tests. First, regarding the coverage of high-frequency words in each passage, the comparison was conducted for the 2K frequency band (i.e., most frequency 2000 words). Although after deleting the outliers, the assumption of normality was met (N = 203), the assumption of homogeneity was violated, Levene's test of homogeneity showed that the variances across groups were unequal ($p < .05$). Thus, Welch's ANOVA was conducted to compare the mean differences of the 2K coverage across the 25 years. The result suggested that there was a statistically significant difference across the 5-year intervals ($F(4, 86.2) = 15.1$, $p < .001$, $\omega^2 = 0.22$). In particular, for the last five years, namely, from 2016 to 2020, the average coverage of high-frequency words (i.e., 2K) was significantly lower than other five-year intervals from the previous 20 years ($\text{Mean}_{2016\text{-}2020\ 2K} = 90.79$). This suggests that the reading comprehension passages in the past five years of exams included a remarkably higher percentage of low-frequency words compare to the previous 20 years. Figure 1 illustrates the mean differences regarding the 2K frequency band across the past 25 years.

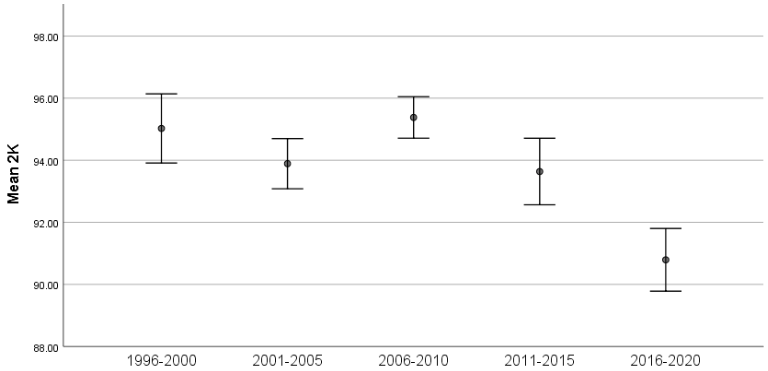


Figure 1. Mean differences of the 2K frequency band, 1996-2020.

According to the description of the National High School English Curriculum Standards (Ministry of Education, 2003, 2017), a vocabulary size of 3000 to 4000 words is considered as the threshold. Hence, in the current study, coverages of 3K and 4K frequency bands (i.e., most frequent 3000 and 4000 words) were selected as the thresholds to compare the differences across the past 25 years. For the 3K band, after discarding the outliers, the assumptions of normality and homogeneity were met, and 186 passages were left for examination (N = 186). No significant differences were detected from the ANVOA test regarding the coverages of the 3K frequency band across the years ($p = 0.12 > .05$). Figure 2 visualizes the mean differences between the five-year intervals. The coverages of the 3K band have gradually lowered throughout the years with the most recent five years reached the lowest percentage ($\text{Mean}_{2016\text{-}2020\ 3K} = 95.89$). In terms of the 4K band, after excluding the outliers to meet the normality assumption, coverages of 134 passages (N = 134) were examined and compared via an ANOVA test. The assumption of homogeneity was also met. The result showed that the differences regarding the coverage of the 4K frequency band were not significant (p = 0.7 > .05). However, it is worth noting that, as the coverage of the 3K band, the last five years also had the lowest mean coverage (Mean $_{\text{2016-2020 4K}}$ = 96.65), and the range between the minimum and maximum coverage values was relatively small (Min $_{\text{2016-2020 4K}}$ = 92.50; Max $_{\text{2016-2020 4K}}$ = 99.10; Range $_{\text{2016-2020 4K}}$ = 6.60). In contrast, from 1996 to 2000, the mean coverage of the 4K frequency band was the highest and the range between the minimum and maximum coverage values was the largest among the years (Mean $_{\text{1996-2000 4K}}$ = 97.27; Min $_{\text{1996-2000 4K}}$ = 92.50; Max $_{\text{1996-2000 4K}}$ = 99.50; Range $_{\text{1996-2000 4K}}$ = 7.00).

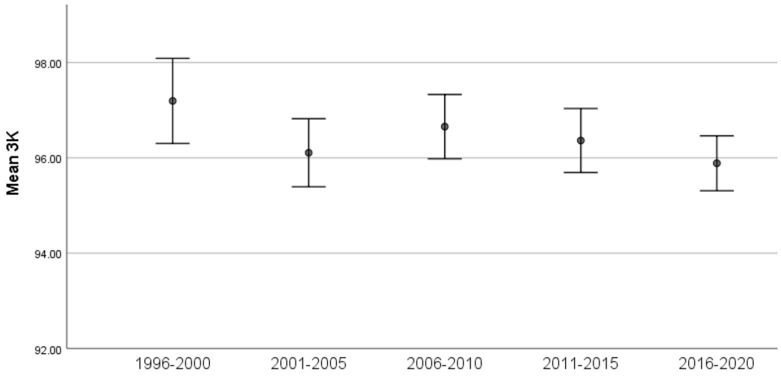


Figure 2. Mean differences of the 3K frequency band, 1996-2020.

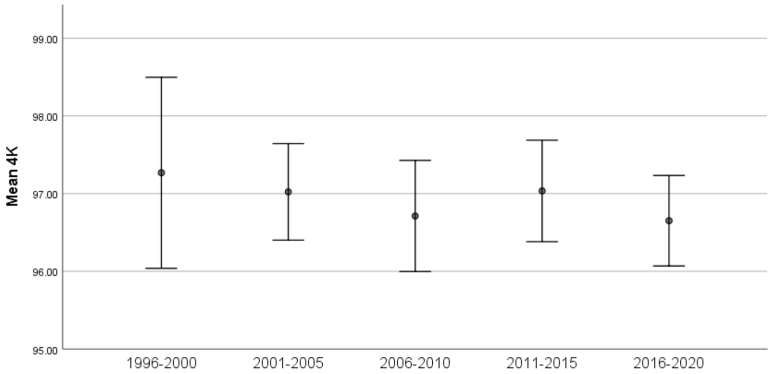


Figure 3. Mean differences of the 4K frequency band, 1996-2020.

Moreover, the results demonstrated that out of the 206 reading comprehension passages, fifty-one of them required a vocabulary size of more than 3K to be able to reach the 95% coverage of the texts. In particular, thirty-five of these passages were found in exams starting from 2010. For the last ten years, it is more common to see such passages that contained more low-frequency words. Fifty-four passages required more than 6K vocabulary size to reach the 95% coverage threshold, and all of these passages were from the exams after the year 2000. Furthermore, eleven passages required more than 10K vocabulary size to reach the 95% coverage and only four of them were from the years before 2010. Three passages from 2016 and 2017 exams showed the lowest coverages with the 10K frequency bands reached lower than 95% of the text coverage.

4.1.2. Lexical sophistication: TAALES. Thirteen indices were selected to represent the six different types of measures for lexical sophistication (see Table 2). First, outliers were discarded from the dataset to maximumly meet the assumption of normality before conducting the MANOVA test (*N* = 172). Second, in terms of homogeneity of variances, only two variables did not meet the assumption. Since MANOVA is relatively robust to the assumption of homogeneity of variance-covariance matrices, thus, the following interpretation procedures were proceeded but with caution. Due to the unequal cell sizes and violation of homogeneity, Pillai’s trace was employed to interpret the differences between the five five-year intervals. The result showed that six variables presented statistically significant differences across the past 25 years ($p_{\text{BNC\_Written\_Range\_AW}} < .01$, $p_{\text{BNC\_Written\_Range\_CW}} < .01$, $p_{\text{LD\_Mean\_RT\_Zscore}} = .05$, $p_{\text{BNC\_Written\_Freq\_CW\_Log}} < .01$, $p_{\text{All\_AWL\_Normed}} < .01$, $p_{\text{aoe\_inverse\_average}} < .01$). Figure 4 illustrates the mean differences across the years regarding the thirteen indices in the six categories.

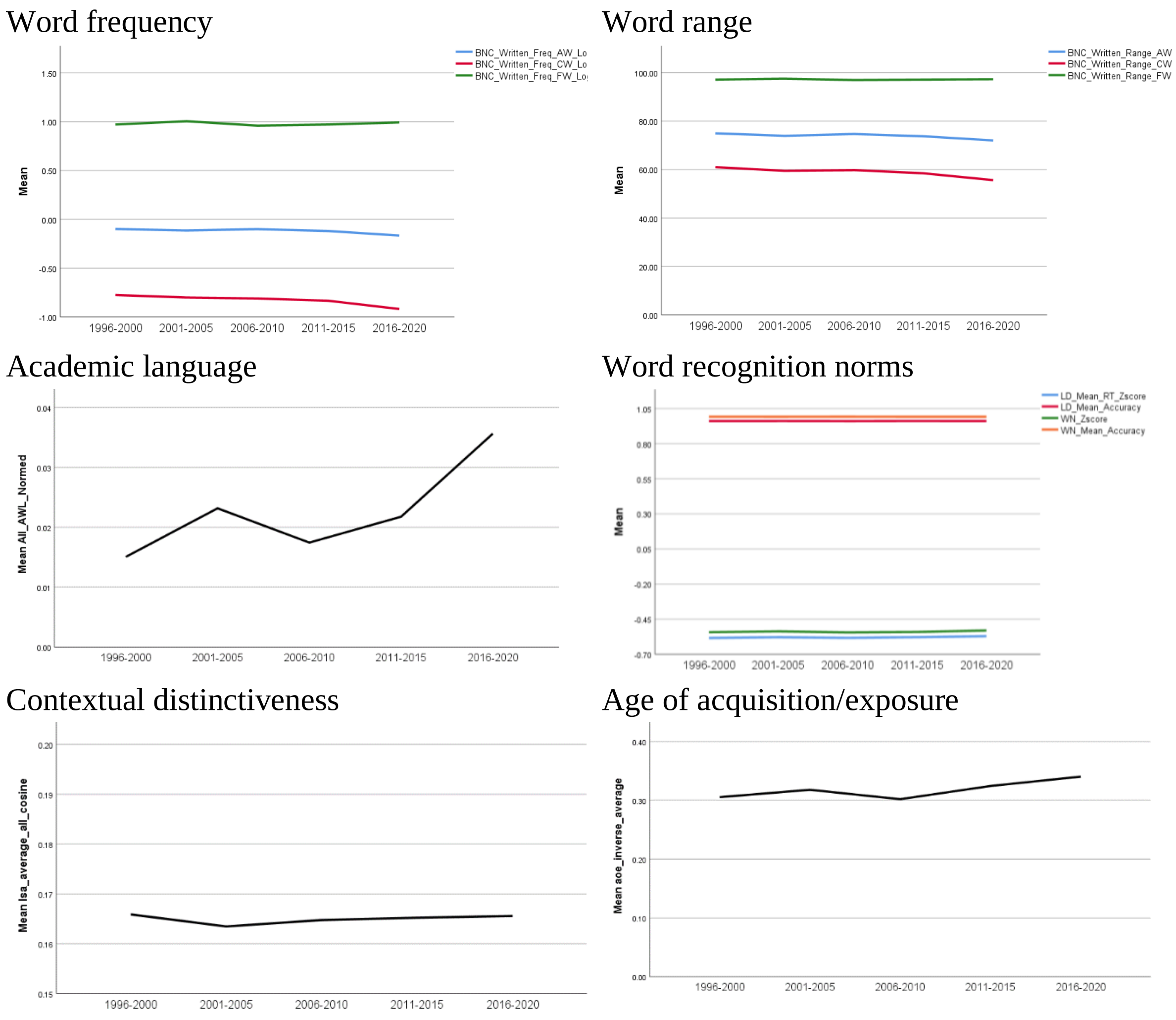
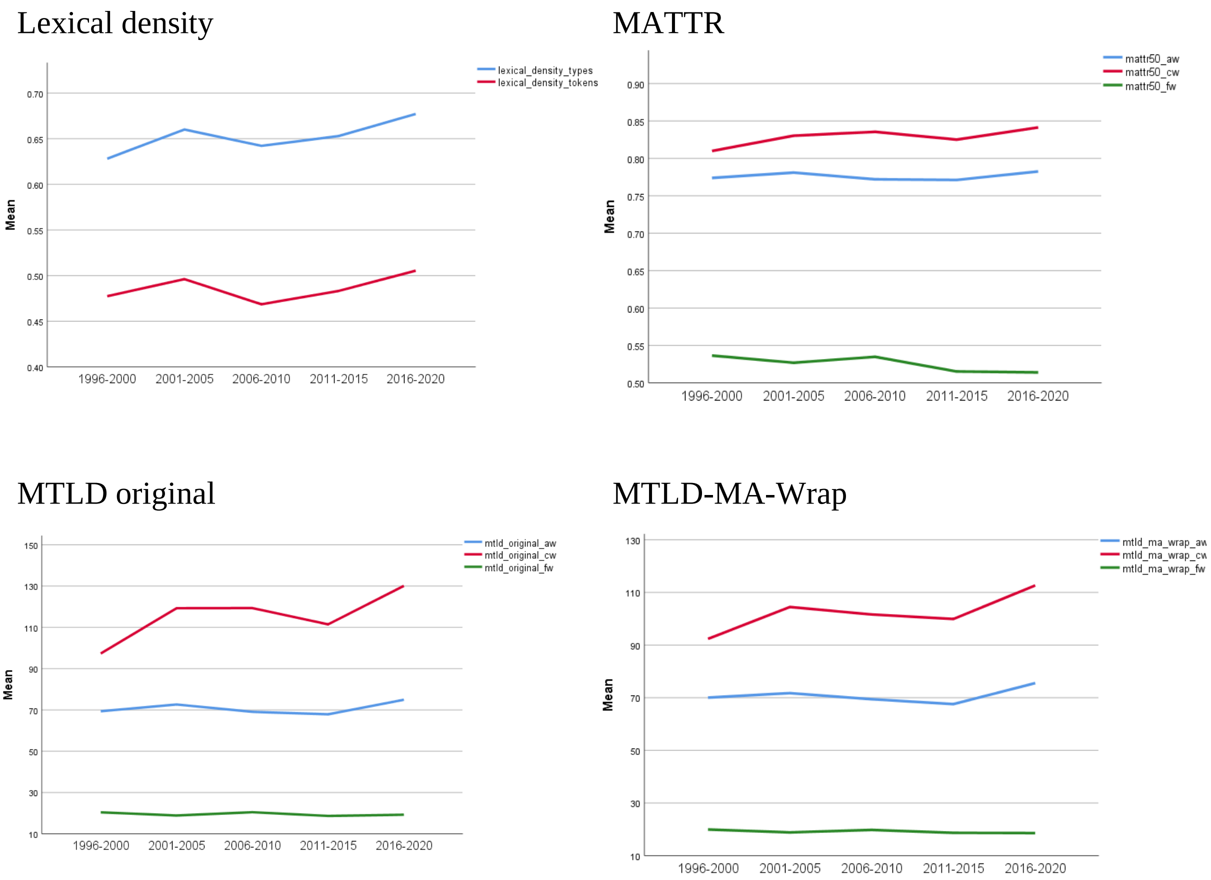

Word frequency
Word range
Academic language
Word recognition norms
Contextual distinctiveness
Age of acquisition/exposure

Figure 4. Mean differences regarding 13 selected measures of lexical sophistication from TAALES, 1996-2020.

Combining the statistical analyses and visualized mean differences, it can be seen that indices related to word frequency, word range, academic language, and age of acquisition/exposure revealed the most apparent changes throughout the years. Towards the most recent years, the word frequency counts and word range in the passages lowered significantly. Moreover, academic words appeared significantly more frequently in recent years' exams. Lastly, the age of exposition was increased, this indicates that more words requiring an older age of acquisition have appeared in the recent years of exams. In sum, indices from TAALES revealed the increase of lexical sophistication throughout the years. This is in line with the previous examination regarding frequency bands.

4.1.3. Lexical diversity and density. After excluding the outliers that influenced the normal distribution of the variables, results of 182 reading passages were included in the comparison (N = 182). The assumption of homogeneity was met, and the variables moderately correlated with each other. The results of MANOVA showed that the two indices of lexical density presented significant differences across the years (both $p < .01$). For the last five years (i.e., 2016-2020), the average lexical density level of the reading comprehension passages was significantly higher than the years from 1996 to 2000 and from 2006 to 2010 (see Figure 5.). Regarding the other indices that are transformations of type-token ratio (i.e., MATTR, MTLD Original, and MTLD-MA-Wrap), although the differences across the years were not statistically significant, it is worth noting that the lexical diversity levels of all words and content words have gradually increased throughout the years despite some minor decrease (see Figure 5). This suggests a moderate increase of the lexical diversity level for the past 25 years, meaning comparing to the late 1990s or early 2000s, more different and less repetitive content words have been included in the recent years of reading comprehension passages.


Lexical density

MATTR

MTLD original

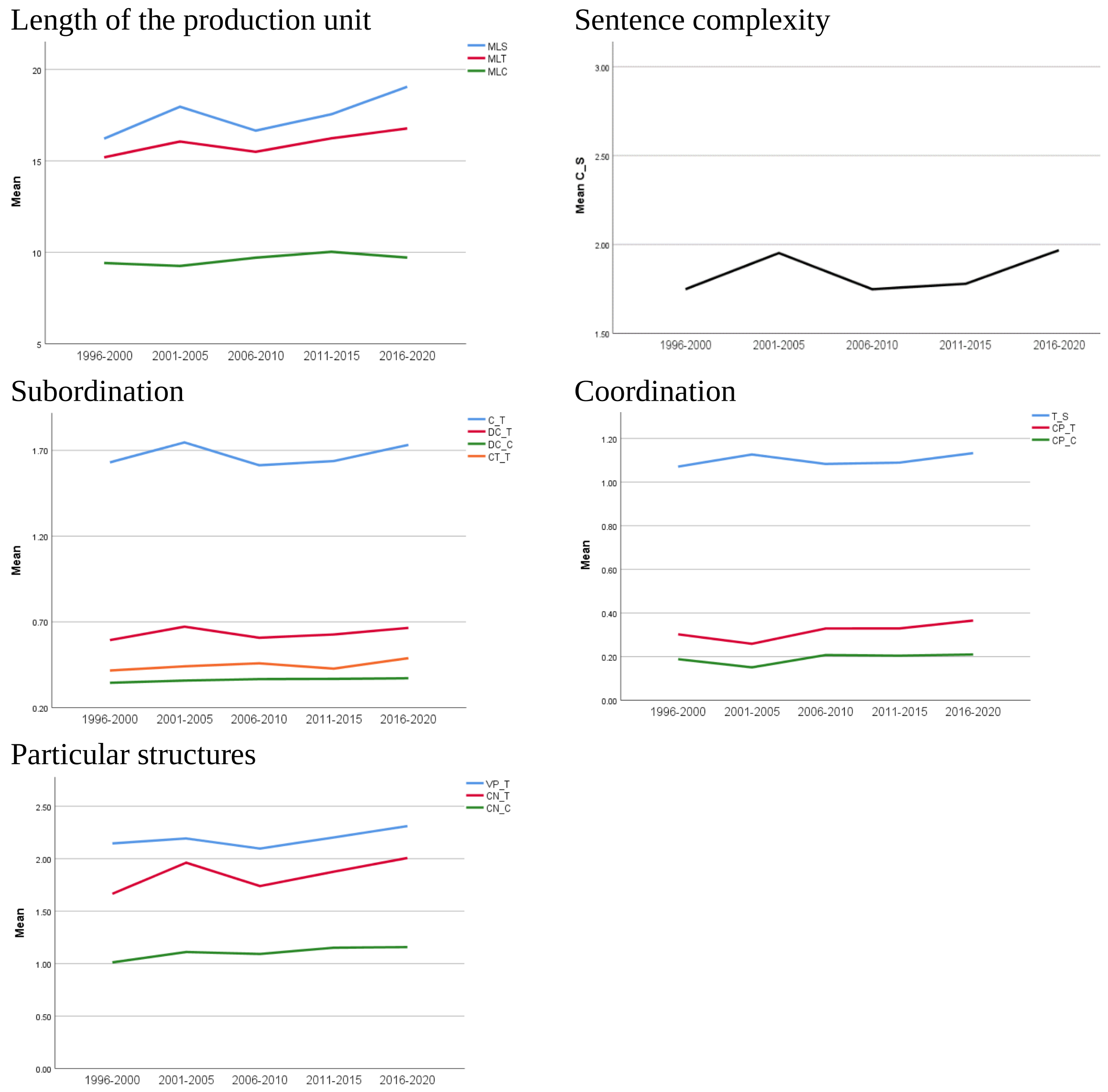
MTLD-MA-Wrap

Figure 5. Mean differences regarding measures of lexical density and diversity from TAALED, 1996-2020.

In sum, for lexical level text complexity, results regarding lexical sophistication measured by frequency bands and TAALES presented an increase toward the most recent years of exams. In other words, the most recent years of exams included a higher percentage of low frequency and academic words in the reading comprehension passages; in addition, the cognitive demand of knowing the words in the passages has been increased. Similarly, regarding lexical density and diversity, the reading texts from the last five years of exams showed a significantly higher lexical density level than the previous years; moreover, the lexical diversity levels of the passages have also been increased gradually.

4.2. Research Question 2: Syntactic Level Text Complexity

Regarding the 14 indices of classic measures of syntactic complexity, after eliminating the outliers that may influence the normality of the variables, 167 passages were finalized for the analysis (N = 167). The assumption of homogeneity was also met. The MANOVA test demonstrated that only one index (i.e., mean length of sentences, MLS) revealed significant differences across the years. The average MLS in the last five years of exams (i.e., 2016-2020) was significantly higher than the years between 2006 and 2010 ($p < .05$). Judging from the mean differences of each index (see Figure 6), although the differences were mainly not statistically significant, the major trend of syntactic complexity can be seen as gradually increasing despite some moderate fluctuations over the years. The results of all 14 indices for the most recent five years were always higher than the first ten years (i.e., 1996 – 2000 & 2001 – 2005).


Length of the production unit

Sentence complexity

Subordination

Coordination

Particular structures

Figure 6. Mean differences regarding measures of syntactic complexity from L2SCA, 1996-2020.

Next, regarding the nine selected clausal and phrasal complexity indices measured by TASSC, the results of the MANOVA test showed that no statistically significant differences were found regarding any of the measures ($N = 206$, all $p > .05$). Based on the visualization of the mean differences of the measures throughout the years (see Figure 7), all of the indices displayed very mild changes.

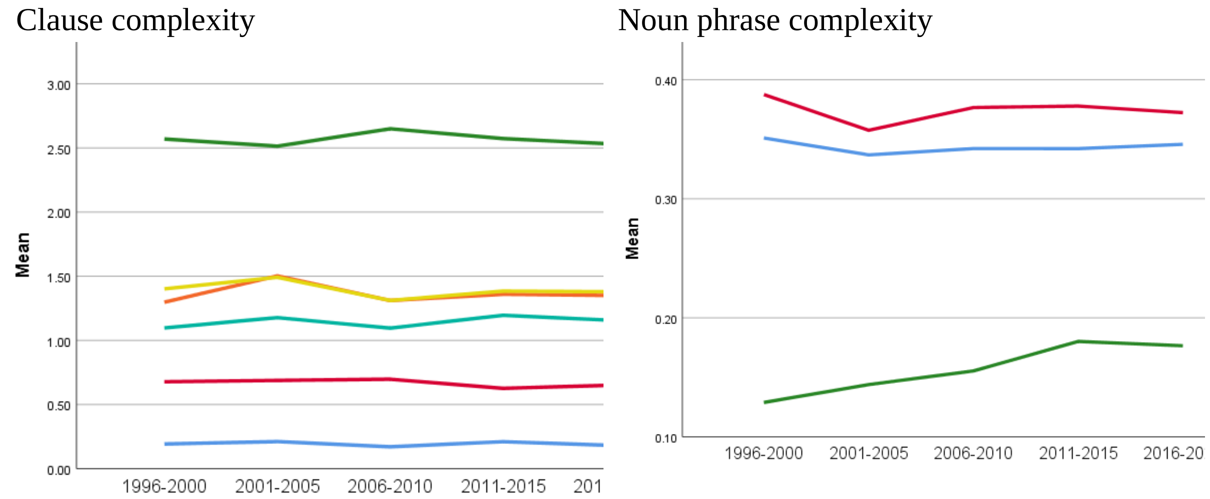
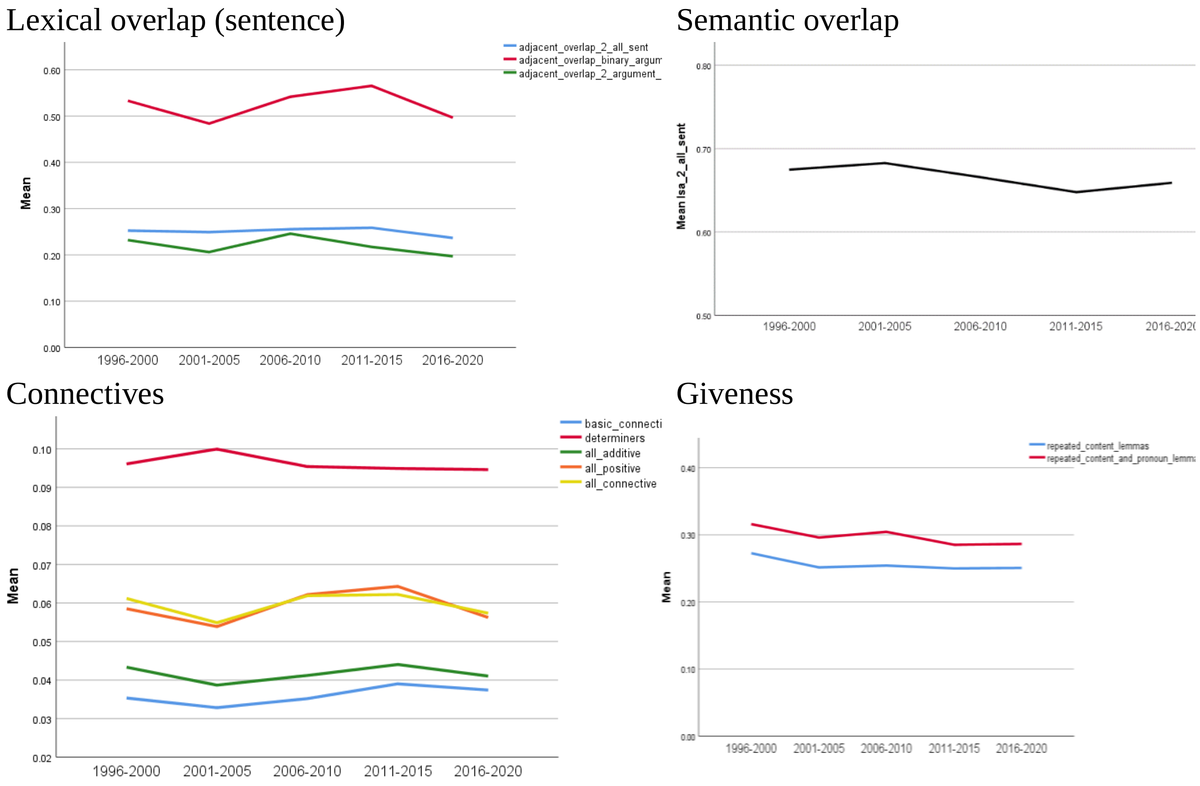


Figure 7. Mean differences regarding measures of clausal and phrasal complexity from TASSC, 1996-2020.

Combined, for syntactic level text complexity of the reading comprehension passages, the changes across the past 25 years have been less drastic. The only significant change among the syntactic measures was the mean length of sentences, the most recent five years (i.e., 2016-2020) included longer sentences on average compared to the years between 2006 and 2010. As for the other measures of syntactic sophistication and complexity, including clausal and phrasal complexity, there have been no significant changes throughout the years.

4.3. Research Question 3: Discourse Level Text Complexity

Regarding the eleven selected variables assessing the cohesion levels of the reading passages, the results from the MANOVA test showed that significant differences across the years were only identified regarding the index adjacent_overlap_2_argument_sent (i.e., number of noun and pronoun lemma types that occur at least once in the next two sentences). The average value of the years between 2006 and 2010 was significantly higher than the years between 2016 to 2020 (N = 206, p = 0.02). For the rest of the indices, no significant changes were detected. Figure 8 illustrates the changes throughout the years for different aspects of cohesion. It can be observed that the changes regarding cohesion in all aspects have been quite mild. The result also differed from the lexical and syntactic levels of text complexity. The reading comprehension passages from the most recent years of exams mainly displayed higher levels of lexical and syntactic complexity compared to the earlier years despite some statistically insignificant differences; however, at the discourse level, in particular for cohesion, the recent five years of reading passages presented slightly lower average values regarding most of the indices compared to the previous two decades.



Lexical overlap (sentence)

Semantic overlap

Connectives

Giveness

Figure 8. Mean differences regarding measures of cohesion from TAACO, 1996-2020.

5. Discussions and Conclusion

In this section, the results of the study are summarized. In addition, further discussions are provided regarding the connection between the results and the curriculum standards and exam guidelines. Finally, pedagogical implications and the limitations of the study are addressed.

5.1. Summary of Results

This study examined the longitudinal development of text complexity at various levels in the reading comprehension passages of the past 25 years of NMET. The results suggested that lexical level text complexity has experienced the most noticeable changes throughout the years. Lexical sophistication, density, and diversity levels of the most recent years of reading comprehension passages have revealed a remarkable increase compared to the previous years, in particular, the late 1990s and early 2000s. This result is in line with Wang (2018), which also suggests the increase of TTR and vocabulary requirement in terms of the provincial university entrance English exam in Jiangsu between 2008 and 2017. The syntactic level text complexity also indicated a general elevation toward the recent years of exams regarding the traditional and more fine-grained measures of syntactic complexity despite some insignificant changes. Lastly, in terms of the discourse level text complexity, the measures of cohesion in different aspects did not reveal apparent changes for the past 25 years. The fluctuation throughout the years has been minor for most of the measures, and unlike the lexical and syntactic level text complexity, reading passages from the recent years did not necessarily show an increase regarding cohesion compared to the earlier years.

5.2. Comparison against General Curriculum Standards and Exam Guidelines

With the analyzed results, it is worth comparing the changes of text complexity of the target reading passages and the General High School English Curriculum Standards (Ministry of Education, 2003, 2017) as well as the exam guideline provided by the National Education Examinations Authority (NEEA, 2019b). Both editions of the curriculum standards underline the requirements with respect to vocabulary in English education. The 2003 edition specifies that 2400 – 2500 words are the minimum requirement for all high school graduates; as for those who would like to further improve English proficiency, the vocabulary size of 3300 – 4500 words should be the target. Regarding the 2017 edition of the curriculum standards, the required vocabulary size does not differ considerably from the earlier edition, it specifies that grasping the target vocabulary items should not be limited to memorization; instead, students should be able to understand and use the vocabulary in different contexts with various topics. The latest version of the guideline for the NMET also provides a word list of around 3000 words. Based on the results from the coverages of the frequency bands throughout the years, a vocabulary size of the most frequency 3000 words is sufficient to cover 95% of the words appear in most of the reading passages. Although Laufer (1989) found that knowing 95% of the words in a text may provide reasonable understanding, 98% coverage has been more widely accepted as the threshold for an acceptable level of comprehension (Nation, 2006; Webb & Paribakht, 2015). Nevertheless, the examination of the past 25 years of reading comprehension passages indicates that knowing the most frequent 3000 – 4000 words would not be able to reach the 98% threshold; moreover, the general trend is that the coverage is decreasing gradually, and the lexical and syntactic complexity of the passages is increasing steadily. This may create a mismatch between the curriculum standards and the reality of the exams as the reading passages in the exams require a much larger vocabulary size than the number suggested in the exam guideline and national curriculum standards. Under this circumstance, high school graduates who are preparing for the NMET need to go well beyond the textbooks and classroom instruction to reach a higher vocabulary size, this may require a lot of extra investment from the students both mentally and financially. Although it is understandable that the NMET is designed as a placement test that distinguishes students with various levels of English proficiency, the increasing expectations as indicated by the changes of the text complexity over the years and the mismatch between the national curriculum standards and the real exam reading passages may add to the growing gap between students because of their access to external learning resources.

Hence, test designers of the NMET are recommended to pay attention to the continuous increase of text complexity of the reading passages, especially at the lexical level. To adhere to the national curriculum standards, the reading comprehension component of the NMET should focus on including more diverse genres of texts that may reflect real-life reading scenarios rather than only adding more advanced and academic vocabulary to increase the lexical level text complexity. Similar control is also suggested to be applied to syntactic features of the reading passages. Instead of merely increasing the length of the sentences, more varied sentence and grammatical structures can be considered to be included in the passages. Furthermore, despite potential difficulties in implementing, test designers and policymakers of the NMET may also consider involving dynamic assessment in evaluating the learners' reading ability as it may reflect a more comprehensive picture of the learners' language capabilities compared to merely using multiple-choice questions (Birjandi, Estaji, & Deyhim, 2013).

5.3. Pedagogical Implications

In terms of pedagogical practices, based on the current trend, the textual complexity of the reading comprehension passages may not decrease within a short time. Classroom English teachers may consider introducing more assorted and authentic reading materials for students to practice their reading ability and increase their vocabulary size through extended reading activities. In this way, students' performance in the NMET may not be crucially influenced due to a lack of vocabulary knowledge and/or their family financial situations to reach more learning materials. As Guo (2012) also noted, extensive reading is closely related to learners' vocabulary development as well as their overall English ability and knowledge. In addition, various empirical studies conducted in different contexts also showed language learners' considerable vocabulary gain through reading authentic English materials. For instance, Pellicer-Sánchez and Schmitt (2010), Shakibaei, Namaziandost, and Shahamat (2019), and Wong and Looi (2010) have investigated the vocabulary learning and general English improvement of English learners from Spanish, Iranian, and Singaporean contexts respectively. The results have generally suggested the promising value of authentic materials and extensive reading in language gain. In addition, improving learners' reading repertoire has also been shown as beneficial for enlarging their relevant background knowledge of the texts, which is further helpful for improving their reading comprehension (Roohani, Dayeri, & Farhang-Ju, 2017).

5.4. Limitations and Future Studies

Finally, this study is not without limitations. The current study examined various linguistic levels of the yearly reading comprehension passages to reveal the development of text complexity. This is only one part of the whole picture. Genre, text structure, authenticity, and test item designing also play an important role in influencing reading comprehension. Further studies may investigate other aspects to provide more insights regarding test designing for reading comprehension. In addition, this study only used computational tools in examining the linguistic features of the passages, more qualitative evaluation from different stakeholders such as experts and classroom teachers may be conducted in the future to triangulate the results and reveal a more comprehensive picture.

Notes

1. RAND is a nonprofit research organization. https://www.rand.org/about.html
2. Although all of the exams were published online every year, it is possible that some exams from the early years are not retrievable. The researcher tried her best to retrieve each year's exams as many as possible from the Internet. Chinese translations of certain vocabulary items were deleted in the corpus.
3. The vocabulary items used in the passages show that the texts were written in British English.

References

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

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

Balyan, R., Crossley, S. A., Brown III, W., Karter, A. J., McNamara, D. S., Liu, J. Y., Lyles, C. R., & Schillinger, D. (2019). Using natural language processing and machine learning to classify health literacy from secure messages: The ECLIPPSE study. *PloS one*, *14*(2), e0212488. Advance online publication. https://doi.org/10.1371/journal.pone.0212488

Benjamin, R. G. (2012). Reconstructing readability: Recent developments and recommendations in the analysis of text difficulty. *Educational Psychology Review*, *24*(1), 63-88. http:// 10.1007/s10648-011-9181-8

Biber, D., Gray, B., & Poonpon, K. (2011). Should we use characteristics of conversation to measure grammatical complexity in L2 writing development? *TESOL Quarterly*, *45*(1), 5–35. https://doi.org/10.5054/tq.2011.244483

Birjandi, P., Estaji, M., & Deyhim, T. (2013). The impact of dynamic assessment on reading comprehension and metacognitive awareness of reading strategy use in Iranian high school learners. *International Journal of Language Testing*, *3*(2), 60-77.

Bulté, B., & Housen, A. (2012). Defining and operationalising L2 complexity. In A. Housen, F. Kuiken, & I. Vedder (Eds.), *Dimensions of L2 performance and proficiency: Complexity, accuracy and fluency in SLA* (pp. 21-46). John Benjamins Publishing Company.

Cheng, L. (2008). The key to success: English language testing in China. *Language Testing*, *25*(1), 15-37. https://doi.org/10.1177%2F0265532207083743

Cheng, L., & Qi, L. (2006). Description and examination of the national matriculation English test. *Language Assessment Quarterly: An International Journal, 3*(1), 53-70. https://doi.org/10.1207/s15434311laq0301_4

Cobb, T. (n.d.). *Compleat lexical tutor*. http://www.lextutor.ca/

Coltheart, M. (1981). The MRC psycholinguistic database. *Quarterly Journal of Experimental Psychology*, *33*(4), 497-505. https://doi.org/10.1080%2F14640748108400805

Crossley, S. A., & McNamara, D. S. (2009). Computational assessment of lexical differences in L1 and L2 writing. *Journal of Second Language Writing, 18*(2), 119-135. https://doi.org/10.1016/j.jslw.2009.02.002

Crossley, S. A., Greenfield, J., & McNamara, D. S. (2008). Assessing text readability using cognitively based indices. *TESOL Quarterly*, *42*(3), 475-493. https://doi.org/10.1002/j.1545-7249.2008.tb00142.x

Crossley, S. A., Kyle, K., & McNamara, D. S. (2016). The tool for the automatic analysis of text cohesion (TAACO): Automatic assessment of local, global, and text cohesion. *Behavior Research Methods*, *48*(4), 1227-1237. https://doi.org/10.3758/s13428-015-0651-7

Crossley, S. A., Skalicky, S., Kyle, K., & Monteiro, K. (2019). Absolute frequency effects in second language lexical acquisition. *Studies in Second Language Acquisition*, *41*(4), 721-744. https://doi.org/10.1017/S0272263118000268

Durán, P., Malvern, D., Richards, B., & Chipere, N. (2004). Developmental trends in lexical diversity. *Applied Linguistics, 25*(2), 220-242. https://doi.org/10.1093/applin/25.2.220

Fang, Z., & Pace, B. G. (2013). Teaching with challenging texts in the disciplines: Text complexity and close reading. *Journal of Adolescent & Adult Literacy*, *57*(2), 104-108. https://doi.org/10.1002/JAAL.229

Farley, A., & Yang, H. H. (2020). Comparison of Chinese Gaokao and western university undergraduate admission criteria: Australian ATAR as an example. *Higher Education Research & Development*, *39*(3), 470-484. https://doi.org/10.1080/07294360.2019.1684879

Gernsbacher, M. A. (2013). *Language comprehension as structure building*. Psychology Press.

Graesser, A. C., McNamara, D. S., & Louwerse, M. M. (2003). What do readers need to learn in order to process coherence relations in narrative and expository text. In A. P. Sweet & C. E. Snow (Eds.), *Rethinking reading comprehension* (pp. *82*-98). Guilford Press.

Graesser, A. C., McNamara, D. S., Louwerse, M. M., & Cai, Z. (2004). Coh-Metrix: Analysis of text on cohesion and language. *Behavior Research Methods, Instruments, & Computers, 36*(2), 193–202. http://doi.org/10.3758/BF03195564

Guo, S. (2012). Using authentic materials for extensive reading to promote English proficiency. *English Language Teaching, 5*(8), 196-206. http://dx.doi.org/10.5539/elt.v5n8p196

Guthrie, J. T., Klauda, S. L., & Ho, A. N. (2013). Modeling the relationships among reading instruction, motivation, engagement, and achievement for adolescents. *Reading Research Quarterly*, *48*(1), 9-26. https://doi.org/10.1002/rrq.035

Hornof, M. (2008). Reading tests as a genre study. *The Reading Teacher*, *62*(1), 69-73. https://doi.org/10.1598/RT.62.1.8

Kim, M., Crossley, S. A., & Kyle, K. (2018). Lexical sophistication as a multidimensional phenomenon: Relations to second language lexical proficiency, development, and writing quality. *The Modern Language Journal*, *102*(1), 120-141. https://doi.org/10.1111/modl.12447

Kyle, K. (2016). *Measuring syntactic development in L2 writing: Fine grained indices of syntactic complexity and usage-based indices of syntactic sophistication* [Doctoral dissertation, Georgia State University]. https://scholarworks.gsu.edu/alesl_diss/35

Kyle, K., & Crossley, S. A. (2015). Automatically assessing lexical sophistication: Indices, tools, findings, and application. *TESOL Quarterly*, *49*(4), 757-786. https://doi.org/10.1002/tesq.194

Kyle, K., Crossley, S. A., & Jarvis, S. (2020). Assessing the validity of lexical diversity indices using direct judgements. *Language Assessment Quarterly*. Advance online publication. https://doi.org/10.1080/15434303.2020.1844205

Kyle, K., Crossley, S., & Berger, C. (2018). The tool for the automatic analysis of lexical sophistication (TAALES): version 2.0. *Behavior research methods*, *50*(3), 1030-1046. https://doi.org/10.3758/s13428-017-0924-4

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

Laufer, B. (1994). The lexical profile of second language writing: Does it change over time? *RELC Journal*, *25*(2), 21-33. https://doi.org/10.1177%2F003368829402500202

Laufer, B., & Nation, P. (1995). Vocabulary size and use: Lexical richness in L2 written production. *Applied Linguistics*, *16*(3), 307-322. https://doi.org/10.1093/applin/16.3.307

Louwerse, M. M. (2004). A concise model of cohesion in text and coherence in comprehension. *Revista Signos*, *37*(56), 41-58. https://doi.org//10.4067/S0718-09342004005600004

Lu, X. (2010). Automatic analysis of syntactic complexity in second language writing. *International Journal of Corpus Linguistics*, *15*(4), 474–496. http://doi.org/10.1075/ijcl.15.4.02lu

Lu, X. (2011). A corpus-based evaluation of syntactic complexity measures as indices of college-level ESL writers' language development. *TESOL Quarterly*, *45*(1), 36–62. https://doi.org/10.5054/tq.2011.240859

McCarthy, P. M., & Jarvis, S. (2010). MTLD, vocd-D, and HD-D: A validation study of sophisticated approaches to lexical diversity assessment. *Behavior Research Methods*, *42*(2), 381-392. https://doi.org/10.3758/BRM.42.2.381

Mesmer, H. A., Cunningham, J. W., & Hiebert, E. H. (2012). Toward a theoretical model of text complexity for the early grades: Learning from the past, anticipating the future. *Reading Research Quarterly*, *47*(3), 235-258. https://doi.org/10.1002/rrq.019

Miller, G. A., Beckwith, R., Fellbaum, C., Gross, D., & Miller, K. J. (1990). Introduction to WordNet: An on-line lexical database. *International Journal of Lexicography, 3*(4), 235-244. https://doi.org/10.1093/ijl/3.4.235

Ministry of Education. (2003). *普通高中英语课程标准(2003 版) [General high school English curriculum standards (2003 ed.)]*. People's Education Press. http://www.moe.gov.cn/srcsite/A26/s8001/200303/W020200401347863199102.pdf

Ministry of Education. (2017). *普通高中英语课程标准(2017 版) [General high school English curriculum standards (2017 ed.)]*. People's Education Press.

http://www.jwc.ecnu.edu.cn/_upload/article/files/f7/28/dc6ae6dc46faa43b343da2b24d7a/6f6c020f-37d9-4ca2-845a-a75064a8d01f.pdf

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

Nation, I. P., & Waring, R. (1997). Vocabulary size, text coverage and word lists. In N. Schmitt & M. McCarthy (Eds.), *Vocabulary: Description, acquisition and pedagogy* (pp. 6-19). Cambridge University Press.

National Education Examinations Authority. (2019a). *2019 年普通高等学校招生全国统一考试大纲 (总纲) [General college admissions unified national examination outline 2019 (General)]*. http://gaokao.neea.edu.cn/html1/report/19012/5989-1.htm

National Education Examinations Authority. (2019b). *2019 年普通高等学校招生全国统一考试大纲 (英语) [General college admissions unified national examination outline 2019 (English subject)]*. http://gaokao.neea.edu.cn/html1/report/19012/5951-1.htm

Ortega, L. (2003). Syntactic complexity measures and their relationship to L2 proficiency: A research synthesis of college-level L2 writing. *Applied Linguistics*, *24*(4), 492–518. https://doi.org/10.1093/applin/24.4.492

Pellicer-Sánchez, A., & Schmitt, N. (2010). Incidental vocabulary acquisition from an authentic novel: Do Things Fall Apart? *Reading in a Foreign Language, 22*(1), 31-55.

Qi, L. (2005). Stakeholders’ conflicting aims undermine the washback function of a high-stakes test. *Language Testing, 22*(2), 142-173. https://doi.org/10.1191%2F0265532205lt300oa

Qi, X. (2007). Is testing an efficient agent for pedagogical change? Examining the intended washback of the writing task in a high-stakes English test in China. *Assessment in Education*, *14*(1), 51-74. https://doi.org/10.1080/09695940701272856

Read, J. (2000). *Assessing vocabulary*. Cambridge University Press.

Roohani, A., Dayeri, K., & Farhang-Ju, M. (2017). Role of background knowledge in Iranian EFL learners' reading comprehension test performance. *International Journal of Language Testing*, *7*(1), 28-39.

Shakibaei, G., Namaziandost, E., & Shahamat, F. (2019). The effect of using authentic texts on Iranian EFL learners' incidental vocabulary learning: The case of English newspaper. *International Journal of Linguistics, Literature and Translation, 2*(5), 422-432. http://doi.org/10.32996/ijllt.2019.2.5.47

Shiotsu, T., & Weir, C. J. (2007). The relative significance of syntactic knowledge and vocabulary breadth in the prediction of reading comprehension test performance. *Language Testing*, *24*(1), 99-128. https://doi.org/10.1177%2F0265532207071513

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

Wang, R. (2018). 2008-2017 年江苏省高考英语阅读理解文本的词汇研究 [A study on the vocabulary of English reading comprehension texts in Jiangsu college entrance examination, 2008-2017]. Educational Measurement and Evaluation, 1, 19-25. http://10.16518/j.cnki.emae.2018.01.004

Webb, S., & Paribakht, T. S. (2015). What is the relationship between the lexical profile of test items and performance on a standardized English proficiency test? *English for Specific Purposes*, *38*, 34-43. https://doi.org/10.1016/j.esp.2014.11.001

Wolfe-Quintero, K., Inagaki, S., & Kim, H. Y. (1998). *Second language development in writing: Measures of fluency, accuracy, and complexity.* Hawaii: University of Hawaii.

Wong, L. H., & Looi, C. K. (2010). Vocabulary learning by mobile-assisted authentic content creation and social meaning-making: Two case studies. *Journal of Computer Assisted Learning*, *26*(5), 421-433. https://doi.org/10.1111/j.1365-2729.2010.00357.x

Wright, T. S., & Cervetti, G. N. (2017). A systematic review of the research on vocabulary instruction that impacts text comprehension. *Reading Research Quarterly*, *52*(2), 203-226. https://doi.org/10.1002/rrq.163

Yan, C. (2015). 'We can't change much unless the exams change': Teachers' dilemmas in the curriculum reform in China. *Improving Schools, 18*(1), 5-19. https://doi.org/10.1177%2F1365480214553744

Zenker, F., & Kyle, K. (2021). Investigating minimum text lengths for lexical diversity indices. *Assessing Writing*, *47,* Advance online publication. https://doi.org/10.1016/j.asw.2020.100505