rEFLections
Vol 30, No 3, September - December 2023

Lexical Coverage Required for Minimal and Optimal Levels of Reading Comprehension in the English Tests of the Higher Education Institutions Examination

MUSTAFA YILDIZ
Foreign Languages Department, Samsun University, Türkiye
Author email: mustafa.yildiz@samsun.edu.tr

Article information

Article history:
Received: 19 Mar 2023
Revised: 10 Aug 2023
Accepted: 18 Sep 2023

Keywords:
Lexical coverage
Lexical thresholds
Minimal reading comprehension
Optimal reading comprehension
Word frequency bands

Abstract

The current research investigates the lexical coverage of reading passages required for achieving the minimal and optimal levels of reading comprehension in a foreign language. More specifically, the aim of the study is to identify the word frequency bands within the reading passages in the English tests of The Higher Education Institutions Examination that surpass the 95% and 98% lexical coverage thresholds needed for minimal and optimal reading comprehension. Fifty reading passages in the exams held in the last decade (2012-2021) were analyzed in the Compleat Web VP system to determine their lexical coverage. The results indicate that when the reading passages are analyzed separately, the vocabulary needed for minimal reading comprehension varies between 2,000 and 18,000 word families. Similarly, the vocabulary required to achieve an optimal level of reading comprehension ranges from 3,000 to 18,000 word families. In addition, upon analyzing reading passages grouped by year, it becomes evident that the word families necessary for achieving minimal and optimal levels of reading comprehension are found in the 4K and 8K word frequency bands, respectively. These findings suggest that when developing a curriculum for the entire high school period, emphasis should be placed on teaching the most frequent 8,000 word families to avoid inconsistency between taught and tested vocabulary. Furthermore, given the substantial number of word families to be addressed, it is important to establish opportunities for both intentional and incidental vocabulary teaching methods. Moreover, it is essential to acquaint learners with vocabulary learning strategies that enable them to acquire vocabulary indirectly.

INTRODUCTION

Vocabulary is a variable that affects all language skills, and there is a positive correlation between the extent of one's receptive vocabulary knowledge and a high level of reading comprehension rate. The higher the vocabulary knowledge is, the higher the reading comprehension rate a reader reaches in a text. The role of vocabulary in reading comprehension has received increased attention. Previous studies indicate a direct relationship between the

size of vocabulary and reading comprehension (Laufer & Ravenhorst-Kalovski, 2010; Nation, 2006; Qian, 2002). Laufer (2013) underlines that as learners' lexical threshold levels escalate, they are freed from being dependent on the assistance of a dictionary and reach the status of an unaided independent reader. Nation (2006) also asserts that 8,000-9,000 word families should be known to attain the optimal coverage of 98% of the running words in a text.

On the other hand, Laufer and Ravenhorst-Kalovski (2010) emphasize that minimal 95% lexical coverage can be achieved with the knowledge of 4,000-5,000 word families. These values can be briefly illustrated by explaining what these numbers signify. For example, an independent reader reaches 98% lexical coverage in a text thanks to the knowledge of 8,000-9,000 word families refers to a group of unknown 100 words in a text of 5,000 words, pointing to a ratio indicating the presence of 2 unknown words in every 100 words. However, another reader with 4,000-5,000 word families covering 95% of the running words in the same text encounters 250 unknown words, represented by a ratio of 5 unknown words in every 100 words. In these two examples, the first reader can deal with the text without needing help, while the second one would need the assistance of a dictionary or any other compensation strategy. To better visualize the word frequency bands, the lexical profiling analysis of the current paragraph indicates that while the words, such as 'abundance', 'attain', 'comprehension', 'dictionary', 'minimal', and 'threshold' appear among the most frequent 4,000 word families, the word 'lexical' stands out among the most frequent 8,000 word families. In addition, while the most frequent word families within the 3K word frequency band constitute 97% of the running words that make up the paragraph, a one-band increase in word frequency (4K) provides a 1.2 percent boost in lexical coverage, bringing it to 98.2%.

Laufer (1992) investigated the relationship between passive vocabulary knowledge and comprehension rate in an academic text. Learners' reading comprehension scores were categorized into five groups according to their passive vocabulary size, encompassing learners with vocabulary levels below 2,000, 2,000, 3,000, 4,000, and 5,000. The findings suggested that 3K word frequency band had the most students who scored a passing grade of 56 or higher, indicating a significant majority of participants at this level. Therefore, 3K word frequency band was expressed as the lexical threshold. According to the results of the regression analysis, every 1,000-word increase in passive vocabulary led to a 7-point increase in reading comprehension. Similarly, an increase of 1,000 words in their vocabulary led to an average reading score improvement of 10 points among students.

Highlighting the importance of establishing lexical threshold benchmarks for educators and curriculum developers when selecting vocabulary content in second language instruction, Laufer and Ravenhorst-Kalovski (2010) pointed out the ambiguity surrounding the terms 'adequate' and 'reasonable' reading comprehension, frequently mentioned in lexical threshold research, and their potential variability across different contexts. For example, Laufer (1989) set the adequate level of comprehension as 55, which was the passing grade at the University of Haifa. On the other hand, Hu and Nation (2000) determined 85.7 as an adequate level of comprehension, which was the average reading comprehension score of the students in the 100% coverage group and knew all of the lexical items in the reading text. The considerable difference in numerical data determined for reading comprehension in these two studies

indicated that a consensus could not be reached in previous studies. It seems that the adequate comprehension of a text varies according to teachers' expectations and the proficiency level of learners.

Schmitt et al. (2011) also opposed the lexical threshold concept expressed by Laufer (1989, 1992) and Hu and Nation (2000). Schmitt et al. (2011) advocated that lexical threshold evokes the idea as if it led to a dramatic leap in reading comprehension after learners reach a certain lexical coverage point. Instead, Schmitt et al. (2011) mentioned that both the previous studies (Hu & Nation, 2000; Laufer, 1989) and their own research indicated a relatively linear relationship between the number of known words in a text and the level of reading comprehension of the same text. For example, Schmitt et al. (2011) found that each 1% increase in lexical coverage in a given text resulted in a 2.3% increase in reading comprehension rate.

Laufer and Ravenhorst-Kalovski (2010) sought the relationship between learners' vocabulary size, lexical coverage in a reading text, and their reading comprehension. The learners were divided into groups based on their scores from The Vocabulary Levels Test (Schmitt et al., 2001). Their reading comprehension was also gauged by a standardized reading test. Laufer and Ravenhorst-Kalovski (2010) alleged that if the most frequent 5,000 word families cover 94% of the text, it can be assumed that learners with the knowledge of 5,000 word families understand a similar percentage of the same text. For example, based on the results in Laufer and Ravenhorst-Kalovski (2010), it can be concluded that the learners with the knowledge of the most frequent first 1,000 word families cover and understand 78.58% of the reading comprehension tests. The results also indicated that a small increase in the number of low frequency words known to learners leads to a considerable enhancement in reading comprehension, even if it does not increase the lexical coverage rate. That is, small or larger increments in learners' lexical coverage under the 5K word frequency band do not differentiate in making a significant contribution to reading comprehension. However, a small increment (1.3%) in lexical coverage from 5K to 7K word frequency bands causes a huge increase (17%) in reading comprehension, indicating the importance of low frequency words, which appear as keywords in comprehending academic texts.

Although the interplay between lexical coverage and comprehension has mostly been investigated in reading texts, it has also been the subject matter of research in various text genres. Dang and Webb (2014) investigated 160 lectures and 39 seminars from 4 disciplines in British Academic Spoken English (BASE) corpus to reveal both the lexical coverage needed to attain 95% and 98% coverage and the lexical coverage of the Academic Word List in academic spoken English. The results indicated that it is necessary to know an average of 4,000 word families to reach beyond 95% coverage and 8,000 word families for 98% coverage in BASE Corpus, which consists of 4 different disciplines. In more detail, in the Social Sciences sub-corpus, the knowledge of the most frequent 5,000 word families is sufficient to exceed 98% lexical coverage, while in the Life and Medical Sciences sub-corpus, the number of word family required to exceed the same level is 13,000. Furthermore, although the use of words from the Academic Word List in 4 different disciplines varies between 3.82% and 5.21%, the percentage of use of these words in the entire BASE corpus is 4.41.

Tegge (2017) compared two pop song corpora, the Wellington Corpus of Popular Songs (WOP), which is formed of 408 popular English songs and the Wellington Corpus of Popular Songs in English Teaching (WOPET), which consists of 635 English songs recommended for classroom use in English language teaching, in order to reveal the size of vocabulary needed to reach 95% and 98% lexical coverage in pop songs. The findings indicated that the knowledge of the most frequent first 3,000 word families and 8,000 word families without the knowledge of proper nouns were required to reach the lexical coverage of 95% and 98%, respectively in WOP corpus. On the other hand, the songs in the WOPET corpus require the knowledge of 3,000 and 5,000 word families without the knowledge of proper nouns to cover 95% and 98% of the lyrics of the songs, respectively. It can be inferred that teacher-selected songs used in classrooms for pedagogical purposes are formed of primarily the most frequent 2,000 word families and lexically less demanding compared to the songs taking part in song charts.

The lexical coverage studies for the spoken language used in daily life have dramatically diverse results. Schonell et al. (1956, as cited in Adolphs & Schmitt, 2003) formed a database of nearly half a million words from spontaneous daily conversations and controlled interviews of Australian workers to investigate the lexical coverage in their spoken discourse. It was found that the most frequent 2,000 word families cover 99% of the spoken discourse of Australian workers. Another lexical coverage research that analyzed the spoken language was Adolphs and Schmitt (2003). Noting that Schonell et al. (1956) is relatively obsolete and the size of data having been analyzed is insufficient compared to today's mega-sized corpora resources, Adolphs and Schmitt (2003) investigated two spoken corpora: the CANCODE corpus and spoken data in British National Corpus. The results based on the CANCODE corpus revealed that the knowledge of 2,000 word families is insufficient to provide 95% lexical coverage. Further analyses focusing on the lexical coverage of individual words in the spoken component of the BNC and the CANCODE corpora indicated that while the knowledge of 3,000 individual word forms provides 95.13% lexical coverage in the BNC corpus, the CANCODE corpus requires the knowledge of 5,000 individual words to go beyond 95% lexical coverage (96.11%).

Jin et al. (2016) compared the vocabulary list in the curriculum to the lexical coverage of the reading texts in high school entrance English tests in order to reveal whether any vocabulary gap exists between teaching and testing reading. A total of 1,507 individual vocabulary items, which were converted to 1,357 word families, were investigated in 859 reading texts in high school entrance English tests in order to determine to what extent these word families provided lexical coverage in these reading texts. The results indicated that these items in the curriculum provide 92.82% lexical coverage on average in reading texts, suggesting that a vocabulary gap that may interfere with reading comprehension occurs in teaching and testing reading. Namely, the vocabulary taught at schools and the ones assessed in tests do not overlap, which may cause a lack of comprehension. In a similar vein, Webb and Paribakht (2015) sought the lexical coverage of a total of 87 reading, listening, and cloze texts in a standardized English proficiency test which was used for student admission purposes in order to determine the lexical demand for reaching minimal 95% and optimal 98% lexical coverage. The results indicated that the most lexically demanding passages were reading comprehension passages, which required the vocabulary size of 6,000 and 14,000 word families as well as proper nouns and marginal words to go beyond 95% and 98% lexical coverage, respectively. On the other hand, the

knowledge of 4,000 and 10,000 word families was sufficient to cover 95% and 98% of listening comprehension passages, respectively, which implied that listening comprehension passages were lexically less demanding than reading comprehension passages. As regards cloze texts, while the vocabulary size of 4,000 word families provided 95% lexical coverage, 6,000 word families covered 98% of the cloze texts, suggesting that the cloze texts were lexically the least demanding comprehension passages. It can be inferred that reading comprehension passages have the heaviest lexical burden while cloze texts have the least lexical burden.

A considerable amount of literature that is concerned with lexical coverage and its relationship with reading comprehension has been published. As stated in the above-mentioned research, many findings have demonstrated that the level of comprehension in receptive language skills increases as the lexical coverage rate increases. Nation (2013) summarizes that the probabilistic threshold stands at approximately 98%. At this level of lexical coverage, nearly all learners are afforded a reasonable opportunity to attain an 'adequate' level of reading comprehension. Nevertheless, when considering a standard of 'minimally acceptable' reading comprehension, it is probable that the probabilistic threshold would shift to approximately 95% lexical coverage.

In a similar vein, in the present study, the researcher aims at investigating the lexical coverage of the reading passages used in a student admission test for selecting students for universities. Because this exam is created by an official institution, it is not possible to access information regarding the reading comprehension levels of test-takers. However, the present study might help find the expected lexical coverage thresholds necessary for "minimally acceptable" and "adequate" levels (Nation, 2013, p. 206) of reading comprehension. More specifically, the present research explores the word frequency bands at which the reading passages in the English tests of The Higher Education Institutions Examination in Türkiye attain the 95% and 98% lexical coverage required for the minimal and optimal levels of reading comprehension, respectively. Therefore, the present study tries to find an answer to the following research question:

What is the lexical coverage of reading texts required for the minimal and optimal levels of reading comprehension in the English tests of a national student admission examination?

METHOD

Data

The reading passages in the English tests of The Higher Education Institutions Examination, held in the last decade (2012-2021) by the Student Selection and Placement Centre for the purpose of selecting students for the universities in Türkiye, constitute the data of the present research. Each of these examinations, which are held at the end of the academic year every year, includes five reading passages each, and a total of 50 reading passages were analyzed. The rationale behind not analyzing the remaining questions in the tests is rooted in the study's primary objective, which is to ascertain the lexical coverage required for both the minimal and optimal levels of reading comprehension within the reading texts. In the present research,

which aims to investigate the sufficient number of vocabulary necessary for reading comprehension, the analysis of the reading passages provides a rich content. Detailed information about the reading passages in each year's test is provided in Table 1.

Table 1
Detailed information on data

Exam	Number of Passages	Tokens	Types	Lexical Diversity	Lexical Density
2021 YDT	5	951	481	0.51	0.58
2020 YDT	5	975	512	0.53	0.58
2019 YDT	5	896	445	0.50	0.59
2018 YDT	5	877	458	0.52	0.57
2017 YDT	5	924	457	0.49	0.59
2016 YDT	5	950	485	0.51	0.56
2015 YDT	5	913	450	0.49	0.55
2014 YDT	5	831	424	0.51	0.58
2013 YDT	5	857	456	0.53	0.56
2012 YDT	5	844	421	0.50	0.55

Table 1 indicates the number of reading passages included in each of the exams held in the last ten years; the total number of tokens (including repetitive words) that make up those reading passages; the total number of types used when repetitive words are not considered; the lexical diversity indicating the ratio of types to tokens; and the lexical density reflecting the ratio of content words to tokens indicating how informative the reading passages are. According to Laufer and Nation (1995), lexical words are those that primarily carry information, making them a crucial factor in determining the density of a text. Accordingly, a text is deemed dense when it contains a high proportion of lexical words relative to the total number of words, encompassing both lexical and functional words. Table 1 indicates that lexical density of the reading passages varies between 0.55 and 0.59, showing that the reading passages contain more lexical words compared to functional words. Therefore, it can be inferred that they are lexically dense and informative. On the other hand, the analysis on the lexical diversity in the reading passages is in line with this inference. Table 1 shows that the lexical diversity of the reading passages varies between 0.49 and 0.53, indicating that the use of different word types in the reading texts is higher, which makes the texts more informative.

Data analysis

The reading passages were accessed from the official webpage of the Student Selection and Placement Centre in Türkiye. Each of these reading passages was analyzed by means of Compleat Web VP on www.lextutor.ca (Cobb, accessed 7 Sep 2021) which showed the number/percentage of the words each of the reading passages contained from the word frequency bands between 1K and 25K according to the BNC/COCA word frequency list. While the Compleat Web VP system on www.lextutor.ca (Cobb, accessed 7 Sep 2021) considered compound words as off-

list, it included proper nouns within the first 1,000 word families. However, the meanings of endocentric compound words are the sum of the meanings of their components, and the meaning is relatively easy to anticipate. Therefore, if both components that made up the compound word were in the first word frequency band, these words such as 'wildlife, lifestyle, and freshwater' were included in the first word frequency band instead of the off-list. On the other hand, proper nouns such as 'Amazon, British, and Himalayas' were already evaluated in the first word frequency band, but the ones that the system could not detect were manually included in the first word frequency band. In addition, some true cognates between English and Turkish such as 'clarinet, flute, supermarket, and laptop' were included into the first word frequency band as well. Also, since the exam was already prepared and administered by an official institution, no spelling errors were detected. After all this preliminary preparation stage, each of the 50 reading passages that formed the data was analyzed in the Compleat Web VP system to determine their lexical coverage.

FINDINGS

This section includes the lexical coverage analyses of each of 5 reading passages in the exams held in the last ten years. The shaded cells represent the word frequency bands in which 95% lexical coverage required for the minimal level of reading comprehension is exceeded, while the bold ones define the word frequency bands in which 98% lexical coverage required for the optimal level of reading comprehension is reached. The value in each cell refers to cumulative percentages of lexical items covered in each word frequency band. In addition, although the system provides specific values for each thousand-word frequency band between the first 1,000 and 25th 1,000 word families, the tables below exclude word frequency bands that lack cumulative alterations.

Table 2

Cumulative lexical coverage of each of five reading passages in 2021 test

Freq. Band	2021-1	2021-2	2021-3	2021-4	2021-5	Mean	SD
1K	84.3	69.8	72.4	71.8	75.6	74.78	5.72
2K	90.1	88.4	85.4	85.4	85.4	86.94	2.19
3K	93.8	97.9	93.2	91.2	93.9	94	2.44
4K	96.9	97.9	95.3	93.6	96.3	96	1,64
5K	97.9	**98.4**	96.3	97.5	97.5	97.52	0.78
6K	97.9	98.4	97.3	97.5	**98.1**	97.84	0.44
7K	**98.4**	98.4	**98.3**	97.5	98.1	**98.14**	0.38
8K	98.4	98.4	98.8	97.5	98.7	98.36	0.51
10K	98.4	98.4	98.8	97.5	99.3	98.48	0.66
11K	99.4	98.4	98.8	97.5	99.3	98.68	0.77
12K	99.9	98.4	98.8	97.5	99.3	98.78	0.91
Off-list	99.9	100	100	100	100	99.98	0.04

The analysis of the reading passages within the 2021 test reveals that possessing knowledge of the most frequent 3,000 word families results in minimal reading comprehension for just one reading passage, specifically 2021-2. Test-takers need to know the most frequent 4,000 word families to attain minimal level of reading comprehension in three of the reading passages while one of the reading passages requires the knowledge of 5,000 word families for minimal comprehension. Two reading passages require the knowledge of 7,000 word families to reach the optimal level of reading comprehension while two of them require the knowledge of 5,000 and 6,000 word families, respectively.

Table 3

Cumulative lexical coverage of each of five reading passages in 2020 test

Freq. Band	2020-1	2020-2	2020-3	2020-4	2020-5	Mean	SD
1K	69.9	79.5	75.8	72	82.9	76.02	5.32
2K	84.1	88.4	87.5	84.9	91.2	87.22	2.85
3K	91.2	93.1	95.1	91.4	94.3	93.02	1.73
4K	93.4	96.3	97.3	94.1	95.9	95.4	1.61
5K	93.4	96.8	**98.2**	97.3	97.5	96.64	1.88
6K	93.4	97.9	98.2	**98.9**	97.5	97.18	2.17
7K	93.9	**98.4**	99.5	99.4	97.5	97.74	2.30
8K	**98.8**	98.9	99.5	99.9	**98**	**99.02**	0.73
9K	98.8	99.4	99.5	99.9	98.5	99.22	0.56
10K	98.8	99.4	99.5	99.9	99.5	99.42	0.40
11K	98.8	99.4	99.5	99.9	100	99.52	0.48
17K	99.3	99.4	99.5	99.9	100	99.62	0.31
Off-list	99.85	100	100	100	100	99.97	0.07

The reading passages in the test in 2020 indicate that test-takers attain the lexical coverage for minimal reading comprehension between 3K and 8K word frequency bands. On the other hand, the optimal level of reading comprehension requires test-takers to know word families between 5K and 8K word frequency bands. The mean scores of the five reading tests indicate that while test-takers who have the knowledge of the most frequent 4,000 word families achieve the minimal reading comprehension level, those with the knowledge of the most frequent 8,000 word families attain the optimal level of reading comprehension.

Table 4

Cumulative lexical coverage of each of five reading passages in 2019 test

Freq. Band	2019-1	2019-2	2019-3	2019-4	2019-5	Mean	SD
1K	89	66.7	91	80.9	71.2	79.76	10.69
2K	96.1	84.1	94.6	88	86.5	89.86	5.23
3K	**99.4**	89.7	96.4	97.8	97.1	96.08	3.74
4K	99.4	90.2	97	**99.4**	**98.3**	96.86	3.85

Freq. Band	2019-1	2019-2	2019-3	2019-4	2019-5	Mean	SD
5K	99.4	90.7	**99.4**	99.4	99.5	97.68	3.90
6K	99.9	94.3	99.4	99.4	99.5	**98.5**	2.36
7K	99.9	94.3	100	99.4	99.5	98.62	2.43
8K	99.9	94.3	100	99.9	99.5	98.72	2.48
9K	99.9	94.8	100	99.9	99.5	98.82	2.26
10K	99.9	95.3	100	99.9	99.5	98.92	2.03
11K	99.9	96.3	100	99.9	100	99.22	1.63
13K	99.9	96.8	100	99.9	100	99.32	1.41
15K	99.9	97.3	100	99.9	100	99.42	1.19
Off-list	100	100	100	100	100	100	0.00

The results of the test of 2019 show considerable variation in terms of the lexical coverage for the minimal level of reading comprehension. The first reading text surpasses the 95% lexical coverage threshold within the 2K word frequency band. In contrast, all the remaining reading texts, except the second one, attain the minimum level of lexical coverage necessary for basic reading comprehension within the 3K word frequency band. However, the second reading text demands that test-takers have knowledge of at least 10,000 word families to achieve the minimal level of reading comprehension. On the other hand, the vocabulary knowledge between 3K and 5K word frequency bands is essential for the optimal level of reading comprehension. For the other four reading passages, a grasp of the most common 5,000 word families achieves an optimal level of reading comprehension. Nevertheless, this knowledge fails even to meet the minimal level of reading comprehension in the second reading text.

Table 5

Cumulative lexical coverage of each of five reading passages in 2018 test

Freq. Band	2018-1	2018-2	2018-3	2018-4	2018-5	Mean	SD
1K	69.7	75.3	73.2	80.5	79.1	75.56	4.38
2K	91.4	87.9	87.2	87.6	87.7	88.36	1.72
3K	**98.8**	94.5	90.6	96	95.7	95.12	2.98
4K	99.9	**98.3**	92.8	96	96.8	96.76	2.67
5K	99.9	98.3	94.5	96.6	97.9	97.44	2.02
6K	99.9	98.3	95.1	97.9	**98.4**	97.92	1.75
7K	99.9	98.3	95.7	97.9	98.4	**98.04**	1.51
8K	99.9	98.3	96.3	97.9	99.5	98.38	1.43
10K	99.9	98.3	96.3	**98.5**	99.5	98.5	1.40
11K	99.9	98.3	97.4	98.5	99.5	98.72	1.00
12K	99.9	98.3	**98**	98.5	99.5	98.84	0.82
13K	99.9	98.3	98.6	98.5	99.5	98.96	0.70
16K	99.9	98.3	99.2	98.5	99.5	99.08	0.67
18K	99.9	98.3	99.8	99.1	99.5	99.32	0.65
Off-list	100	100	100	100	100	100	0.00

The 2018 exam exhibits disparities in vocabulary-based difficulty levels among its reading passages. To attain an optimal level of reading comprehension in the first two reading passages, one needs familiarity with 3,000 and 4,000 word families, respectively. In contrast, the third reading passage necessitates knowledge of 12,000 word families to achieve an equivalent level of reading comprehension. On the other hand, in the case of the third reading text, test-takers must be acquainted with 6,000 word families to attain even a minimal level of reading comprehension. However, in the preceding two reading texts, significantly fewer word families suffice for achieving an optimal level of reading comprehension.

Table 6

Cumulative lexical coverage of each of five reading passages in 2017 test

Freq. Band	2017-1	2017-2	2017-3	2017-4	2017-5	Mean	SD
1K	79.4	78.8	85	83.3	72.5	79.8	4.84
2K	87.4	90.2	91.7	93.2	88.9	90.28	2.28
3K	88	96.4	97.4	**98.4**	**98.3**	95.7	4.38
4K	89.1	96.9	97.9	98.4	100	96.46	4.26
5K	89.7	**99**	**98.4**	98.4	100	97.1	4.19
6K	90.3	99.5	98.4	99.4	100	97.52	4.08
7K	90.9	100	98.9	99.4	100	97.84	3.91
8K	92	100	98.9	99.4	100	**98.06**	3.42
9K	92.6	100	98.9	99.4	100	98.18	3.15
10K	93.2	100	98.9	99.4	100	98.3	2.89
18K	**98.3**	100	98.9	99.4	100	99.32	0.73
Off-list	100	100	100	100	100	100	0.00

The exam held in 2017 also contains inconsistencies in terms of difficulty based on vocabulary, just like the exam held in 2018. While the number of vocabulary required to achieve both minimal and optimal levels of reading comprehension in the first reading text is 18,000 word families, knowing the most frequent 3,000 word families in the fourth and fifth reading texts provides both minimal and optimal levels of reading comprehension. The second and third reading texts share a common trait in that they both surpass the thresholds for both minimal and optimal reading comprehension with vocabulary knowledge of 3,000 and 5,000 word families, respectively.

Table 7

Cumulative lexical coverage of each of five reading passages in 2016 test

Freq. Band	2016-1	2016-2	2016-3	2016-4	2016-5	Mean	SD
1K	82.1	76.7	80.7	79.2	79	79.54	2.02
2K	92.1	86.8	90.4	89.8	89.8	89.78	1.91
3K	96.3	91.2	96.7	96.6	94.9	95.14	2.32
4K	97.4	93.1	**100**	97.1	96.2	96.76	2.49

Freq. Band	2016-1	2016-2	2016-3	2016-4	2016-5	Mean	SD
5K	**98.5**	95	100	**98.1**	**99.4**	**98.2**	1.94
6K	99	95	100	99.1	100	98.62	2.08
7K	99.5	95	100	99.6	100	98.82	2.15
10K	100	**98.8**	100	99.6	100	99.68	0.52
11K	100	98.8	100	100	100	99.76	0.54
Off-list	100	100	100	100	100	100	0.00

In the 2016 exam, the vocabulary required for minimal reading comprehension ranges from 3,000 to 5,000 word families, whereas the vocabulary knowledge required for optimal reading comprehension spans from 4,000 to 10,000 word families, referring to a wider range of requirements. While an average familiarity with 5,000 word families suffices for achieving the optimal level of reading comprehension in 5 different reading passages, the second reading passage demands a significantly higher level of knowledge, specifically 10,000 word families, which exceeds the average requirement.

Table 8

Cumulative lexical coverage of each of five reading passages in 2015 test

Freq. Band	2015-1	2015-2	2015-3	2015-4	2015-5	Mean	SD
1K	81.4	80.7	70.1	83.7	76.6	78.5	5.35
2K	91.9	93.4	89.5	94.9	87.5	91.44	2.97
3K	97.1	**99.5**	95.1	97.1	97.8	97.32	1.58
4K	**98.5**	100	97.9	97.7	**98.9**	**98.6**	0.92
5K	98.5	100	**98.6**	97.7	99.4	98.84	0.88
6K	99.5	100	98.6	**98.3**	99.4	99.16	0.69
7K	99.5	100	98.6	98.9	99.4	99.28	0.54
9K	100	100	99.3	99.5	99.4	99.64	0.34
10K	100	100	100	99.5	99.4	99.78	0.30
23K	100	100	100	99.5	99.9	99.88	0.22
Off-list	100	100	100	100	100	100	0.00

In the 2015 test, achieving minimal reading comprehension in 5 different reading passages necessitates an average vocabulary knowledge of 3,000 word families, whereas achieving optimal reading comprehension requires a vocabulary knowledge of 4,000 word families. It can be inferred that the second reading text is less demanding in terms of vocabulary because it achieves both minimal and optimal levels of reading comprehension within the same frequency band, which consists of 3,000 word families.

Table 9

Cumulative lexical coverage of each of five reading passages in 2014 test

Freq. Band	2014-1	2014-2	2014-3	2014-4	2014-5	Mean	SD
1K	78.7	71.7	61.1	77.5	82	74.2	8.21
2K	93.6	90	81.2	86.4	89.9	88.22	4.68
3K	**99.5**	95.8	90.2	95.9	96.4	95.56	3.36
4K	100	97.9	95.1	**98.9**	**98.6**	**98.1**	1.84
5K	100	**99.5**	96.5	98.9	99.3	98.84	1.37
6K	100	100	97.2	100	99.3	99.3	1.21
10K	100	100	97.9	100	99.3	99.44	0.91
13K	100	100	**99.3**	100	99.3	99.72	0.38
15K	100	100	100	100	99.3	99.86	0.31
Off-list	100	100	100	100	100	100	0.00

Similar to the 2015 exam, the 2014 exam also requires an average vocabulary knowledge of 3,000 and 4,000 word families for minimal and optimal levels of reading comprehension, respectively. However, there is an inconsistency in that the first reading text necessitates knowledge of the most frequent 3,000 word families to achieve the optimal level of reading comprehension, whereas the third reading text demands a significantly higher count of 13,000 word families for the same level of comprehension.

Table 10

Cumulative lexical coverage of each of five reading passages in 2013 test

Freq. Band	2013-1	2013-2	2013-3	2013-4	2013-5	Mean	SD
1K	76.8	67.2	75.2	75.2	80.5	74.98	4.86
2K	89.8	87	89.7	89.6	91.9	89.6	1.74
3K	**98.8**	94.9	96.4	95.5	95.7	96.26	1.52
4K	98.8	97.2	**98.2**	96.2	**98.9**	97.86	1.15
5K	98.8	97.8	98.8	96.2	98.9	**98.1**	1.15
6K	99.4	**98.9**	98.8	96.2	99.4	98.54	1.34
7K	100	99.5	99.4	96.9	99.4	99.04	1.22
8K	100	100	99.4	96.9	99.9	99.24	1.33
Off-list	100	100	100	100	100	100	0.00

In the 2013 exam, the first reading text stands out as being lexically less demanding and necessarily easier to comprehend. In addition, an average vocabulary knowledge of 3,000 and 5,000 word families is adequate for achieving minimal and optimal levels of reading comprehension in this year's reading texts.

Table 11

Cumulative lexical coverage of each of five reading passages in 2012 test

Freq. Band	2012-1	2012-2	2012-3	2012-4	2012-5	Mean	SD
1K	78.7	91.4	81	73.9	86.3	82.26	6.78
2K	95.2	96.5	95.7	89.8	95	94.44	2.66
3K	**98.2**	**98**	**98.4**	97.8	**99.3**	**98.34**	0.58
4K	99.4	100	98.9	**98.5**	99.9	99.34	0.64
5K	99.4	100	99.4	98.5	99.9	99.44	0.59
6K	99.4	100	99.4	99.2	99.9	99.58	0.35
Off-list	100	100	100	100	100	100	0.00

The 2012 test stands out as the least lexically demanding, as test-takers can attain both minimal and optimal levels of reading comprehension within the same frequency band, specifically the most frequent 3,000 word families. For all texts except the fourth one, achieving minimal and optimal reading comprehension necessitates vocabulary knowledge of the most frequent 2,000 and 3,000 word families, respectively.

Table 12

The number of reading passages reaching minimal and optimal levels of reading comprehension in various word frequency bands

Freq. Band	95%	98%
2K	5	-
3K	29	11
4K	9	11
5K	3	10
6K	1	5
7K	-	3
8K	1	2
10K	1	2
12K	-	1
13K	-	1
18K	1	1
Off-list	-	3
Total	50	50

Table 12 indicates that the reading passages require lexical coverage between 2K and 18K word frequency bands for the minimal level of reading comprehension. The optimal level of reading comprehension is attained with the lexical coverage between 3K and 18K word frequency bands. None of the reading texts achieve an optimal level of reading comprehension within the 2K word frequency band. Additionally, having lexical coverage within the most frequent 3K word frequency band leads to minimal reading comprehension in 68% of all the reading

texts. A one-band increase in vocabulary knowledge, specifically acquiring familiarity with the most frequent 4,000 word families, leads to minimal reading comprehension in 86% of the total number of reading texts. On the other hand, mastery of the most frequent 3,000 word families enables optimal reading comprehension in 22% of the reading passages. However, a two-band increase in vocabulary knowledge, i.e., gaining proficiency in the most frequent 5,000 word families, leads to optimal reading comprehension in 64% of all the reading passages.

DISCUSSION

Individual reading passages in the present data considerably vary in their lexical demands, which cause variation in text difficulty. Looking at the overall results, it can be mentioned that it is really difficult to make reliable inferences about the expected number of vocabulary known to test-takers for the minimal and optimal levels of reading comprehension because of the lexical coverage of the words that make up the reading passages used in the tests over the past decade. Among the tests, there does not appear to be any consistency in terms of the word frequency bands that need to be known for both minimal and optimal levels of reading comprehension. This lack of consistency, apparent when evaluating the 50 reading passages individually, is also observed in the average lexical coverage results, where all 5 reading passages from each test are collectively assessed. For example, upon individual evaluation of each reading passage, it becomes apparent that minimal reading comprehension can be attained with the knowledge of both the second and eighteenth 1,000 word families. Likewise, to reach an optimal level of reading comprehension, it is crucial to have a grasp of both the third and eighteenth 1,000 word families. On the other hand, when considering the average lexical coverage results, where all 5 reading passages from each test are collectively examined, it can be inferred that possessing vocabulary knowledge of both the third and fourth 1,000 word families is essential to surpassing the minimal level of reading comprehension. However, to achieve an optimal level of reading comprehension, the required word knowledge extends from 3K to 8K word frequency bands, encompassing every band within this interval. Therefore, although it is difficult to make an inference regarding the vocabulary required for the minimal and optimal levels of reading comprehension due to the inconsistency in the lexical coverage of the words in the reading passages, it might be reasonable to suggest teaching or knowing the most frequent 4,000 and 8,000 word families for the minimal and optimal levels of reading comprehension, respectively. These former values constitute the upper limit of the vocabulary knowledge required for both levels of understanding. Therefore, it can be risky to suggest teaching or learning vocabulary less than 4,000 and 8,000 word families for the minimal and optimal levels of reading comprehension, respectively.

The lexical thresholds suggested above for attaining minimal (4,000 word families) and optimal (8,000 word families) levels of reading comprehension in the reading passages in the English tests of The Higher Education Institutions Examination in Türkiye are similar to the results obtained in previous studies. While Laufer and Ravenhorst-Kalovski (2010) underline the significance of the knowledge of 4,000-5,000 word families for providing minimal 95% lexical coverage, Nation (2006) emphasizes the prominence of 8,000-9,000 word families to be able to attain the optimal lexical coverage of 98% of the running words in a text.

Although analyzing the spoken language data, Dang and Webb (2014) also reveal the importance of the knowledge of 4,000 and 8,000 word families for going beyond the 95% and 98% lexical coverage, respectively. These overall results of the spoken data based on British Academic Spoken English (BASE) Corpus show similarities with the findings attained at the end of the present data. However, the further results, in which Dang and Webb (2014) compare the data from the Social Sciences and the Life and Medical sub-corpora, indicate that different genres may show variation in lexical coverage. Dang and Webb (2014) find that while the Social Sciences sub-corpus requires the vocabulary knowledge of the most frequent 5,000 word families to exceed 98% lexical coverage in a text, the threshold level for the same lexical coverage in a text in the Life and Medical Sciences sub-corpus is 13,000 word families. This data indicates that the results are text-dependent and may differ. Similarly, in Tegge's (2017) comparison of the Wellington Corpus of Popular Songs (WOP) and the Wellington Corpus of Popular Songs in English Teaching (WOPET), it is observed that both corpora achieve over 95% lexical coverage within the 3K word frequency band. However, the corpus designed for educational purposes proves to be less lexically demanding when compared to the original Wellington Corpus of Popular Songs (WOP). This is evident as the educational corpus requires knowledge of 5,000 word families to yield 98% lexical coverage, whereas the original corpus surpasses the same threshold with 8,000 word families.

In a study on another exam held in Türkiye, Unaldi and Bardakci (2014) analyze the lexical coverage of the reading passages available in The Interuniversity Foreign Language Examination (UDS) held between 2006 and 2012. The results of the lexical frequency profiling analysis in these exams, which are held separately in the fields of Hard Sciences, Social Sciences and Health Sciences, show that the lexical coverage rate of the most frequent first 2,000 words is 79.11%, 81.49%, and 76.84%, respectively. When comparing these findings to the results of the present study, it becomes apparent that The Higher Education Institutions Examinations are less demanding in terms of vocabulary difficulty. In the Interuniversity Foreign Language Examination (UDS), which is specifically tailored for various scientific disciplines, the use of domain-specific academic terminology contributes to a lexical coverage rate of approximately 85% (Unaldi & Bardakci, 2014). This suggests that even achieving proficiency in the most frequent first 2,000 words and the Academic Word List (Coxhead, 2000) is insufficient for UDS test-takers to attain the same level of lexical coverage as those participating in the student admission exam in the current study with equivalent vocabulary knowledge.

CONCLUSION

The present study is conducted to reveal the lexical coverage of the reading passages in the English tests of The Higher Education Institutions Examination, held in the last ten years (2012-2021) by the Student Selection and Placement Centre for the purpose of selecting students for the universities in Türkiye. By determining how lexically demanding each of 50 reading passages is, it is aimed to make an inference for the vocabulary size test-takers need to attain for achieving minimal and optimal levels of reading comprehension. The findings reveal that the lexical coverage values for each of the reading passages vary widely. Namely, while test-takers achieve the minimal level of lexical coverage with 2,000 word families in some

of the reading passages, there is also a particular reading passage that demands a substantial vocabulary knowledge of 18,000 word families for yielding the same minimal level of lexical coverage. Similarly, the vocabulary knowledge required for the optimal level of lexical coverage ranges between 3,000 and 18,000 word families. However, rather than assessing each reading passage individually, based on the collective analysis of reading passages used over the past decade, it is more reasonable to propose targeted vocabulary knowledge thresholds of 4,000 and 8,000 word families as the benchmarks for the minimal and optimal levels of lexical coverage, respectively.

These findings have some pedagogical implications. To mitigate issues arising from insufficient vocabulary in the reading passages that evaluate test-takers' reading comprehension skills in the English tests of The Higher Education Institutions Examination, it is advisable to incorporate the most frequent 8,000 word families into the curriculum, spanning from early education through high school. Furthermore, the most frequent 8,000 word families are linguistically challenging and too numerous to be exclusively taught through intentional vocabulary instruction methods. Consequently, it is essential to establish opportunities for incidental vocabulary acquisition and to guide students, particularly, in adopting individual vocabulary learning strategies. In addition, achieving both minimal and optimal levels of reading comprehension in two different reading passages necessitates knowledge of the most frequent 18,000 word families, presenting a significant challenge for test-takers. Considering the word frequency bands encompassing words utilized in all the reading passages, it would be more prudent for test designers to maintain consistency in their word choices throughout the test.

The overall findings suggest that it is really difficult to make reliable inferences about the expected number of vocabulary known to test-takers for the minimal and optimal levels of reading comprehension because of the lexical coverage of the words that make up the reading passages. There is a lack of consistency among the tests in terms of the frequency bands of the words, which can spoil the validity of the tests. Therefore, test designers should strive to improve and standardize the lexical coverage of words in reading passages. Clear criteria should be established for selecting words appropriate for the desired levels of vocabulary knowledge and comprehension, which could include categorizing words according to frequency, relevance, and difficulty, thus providing a more accurate assessment of test takers' vocabulary knowledge.

THE AUTHOR

Mustafa Yildiz has earned his doctorate in English Language Teaching and is currently working as a lecturer in the Foreign Languages Department at Samsun University, Türkiye.
mustafa.yildiz@samsun.edu.tr

REFERENCES

Adolphs, S., & Schmitt, N. (2003). Lexical coverage of spoken discourse. *Applied Linguistics, 24*(4), 425–438. https://doi.org/10.1093/applin/24.4.425

Cobb, T. *Compleat Web VP* (v.2.5). https://www.lextutor.ca/vp/comp/

Coxhead, A. (2000). A new academic word list. *TESOL Quarterly, 34*, 213-238. https://doi.org/10.2307/3587951

Dang, T. N. Y., & Webb, S. (2014). The lexical profile of academic spoken English. *English for Specific Purposes, 33,* 66-76. https://doi.org/10.1016/j.esp.2013.08.001

Hu, M., & Nation, I. S. P. (2000). Unknown vocabulary density and reading comprehension. *Reading in a Foreign Language, 13*, 403−430.

Jin, T., Li, Y., & Li, B. (2016). Vocabulary coverage of reading tests: Gaps between teaching and testing. *TESOL Quarterly, 50*(4), 955–964. http://www.jstor.org/stable/44984727

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. (1992). How much lexis is necessary for reading comprehension?. In P. J. L. Arnaud & H. Béjoint (Eds.), *Vocabulary and applied linguistics*. Palgrave Macmillan. https://doi.org/10.1007/978-1-349-12396-4_12

Laufer, B. (2013). Lexical thresholds for reading comprehension: What they are and how they can be used for teaching purposes. *TESOL Quarterly, 47*(4), 867-872. http://dx.doi.org/10.1002/tesq.140

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

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

Nation, I. S. P. (2006). How large a vocabulary is needed for reading and listening? *The Canadian Modern Language Review, 63*, 59–82.

Nation, I. S. P. (2013). *Learning vocabulary in another language* (2nd ed). Cambridge University Press. https://doi.org/10.1017/CBO9781139858656

Qian, D. D. (2002). Investigating the relationship between vocabulary knowledge and academic reading performance: An assessment perspective. *Language Learning, 52*, 513–536.

Schmitt, N., Schmitt, D., & Clapham, C. (2001). Developing and exploring the behavior of two new versions of the Vocabulary Levels Test. *Language Testing, 18*(1), 55–88.

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

Schonell, F. J., Meddleton, I. G., & Shaw, B. A. (1956). *A study of the oral vocabulary of adults*. University of Queensland Press.

Tegge, F. (2017). The lexical coverage of popular songs in English language teaching. *System, 67*, 87-98.

Unaldi, İ., & Bardakci, M. (2014). Vocabulary profiles of English language proficiency exams in Turkey: The case of ÜDS. *Turkish Studies, 9*(3), 1523-1534.

Webb, S., & Paribakht, T. (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