MỨC ĐỘ BAO PHỦ TỪ VỰNG VÀ TIỀM NĂNG HỌC TỪ VỰNG NGẪU NHIÊN TRONG GIÁO TRÌNH TIẾNG ANH DÀNH CHO SINH VIÊN VIỆT NAM HỌC TIẾNG ANH NHƯ MỘT NGOẠI NGỮ

Lê Hoàng Thu Anh*, Nguyễn Hồng Giang*

Nghiên cứu này trình bày phân tích dựa trên ngữ liệu về mức độ bao phủ từ vựng trong các văn bản đọc của giáo trình Inside Reading dành cho sinh viên một trường đại học ở Việt Nam. Nghiên cứu khảo sát đặc điểm từ vựng của các văn bản đọc và đánh giá mức độ lặp lại của các từ chưa quen thuộc nhằm xác định liệu các văn bản này có hỗ trợ việc tiếp thu từ vựng ngẫu nhiên hay không. Các đoạn văn trong giáo trình được xử lý bằng công cụ Vocabprofilers trên trang Lextutor.ca, và được so sánh với kết quả từ bài kiểm tra Vocabulary Size Test. Tám mươi sáu sinh viên tham gia bài kiểm tra Vocabulary Size Test để xác định quy mô vốn từ vựng của người học. Kết quả cho thấy mức độ bao phủ từ vựng 95% đạt được với 4.000 từ phổ biến nhất và mức bao phủ tối ưu 98% đạt được với 6.000 từ. Với vốn từ vựng thụ động trung bình khoảng 8.000 từ của người tham gia, kết quả cho thấy các văn bản phù hợp với trình độ người học. Tuy nhiên, các nhóm từ mới hiếm khi xuất hiện lại. Mặc dù tài liệu đọc hỗ trợ hiểu bài, chúng ít tạo cơ hội củng cố từ vựng mới. Kết quả nghiên cứu đưa ra gợi ý cho người thiết kế giáo trình và giáo viên.

Từ khóa: *ngữ liệu, lặp lại từ vựng, học từ vựng ngẫu nhiên, độ bao phủ từ vựng, ngưỡng từ vựng.*

This study presents a corpus-based analysis of the vocabulary coverage in the reading texts of the "Inside Reading" series used by students at a Vietnamese university. It examines the lexical profile of the reading texts and evaluates the recurrence of unfamiliar vocabulary items to determine the extent to which these texts support incidental vocabulary acquisition. The reading passages were processed unsing the Vocabprofilers on the Lextutor.ca website and compared with the results of Vocabulary Size Test administered to 86 students to estimate their vocabulary size. The findings reveal that 95% lexical coverage was achieved with the first 4,000 most frequent words while the optimal 98% coverage was reached with 6,000 words. Given the participants' average receptive vocabulary size of approximately 8,000 words, the findings indicate that the texts match their proficiency level. However, unfamiliar word families rarely reoccur. While the reading materials support comprehension, they offer limited opportunities for the reinforcement of new vocabulary. These findings provide implications for textbook designers and language teachers.

Keywords: *corpus, vocabulary repetition, incidental vocabulary learning, vocabulary coverage, vocabulary threshold.*

* **ThS., Khoa tiếng Anh, Trường Đại học Hà Nội**

** **TS., Khoa tiếng Anh, Trường Đại học Hà Nội (tác giả liên hệ)**

Email: giangnh@hanu.edu.vn; anhlht@hanu.edu.vn

LEXICAL COVERAGE AND INCIDENTAL LEARNING POTENTIAL IN ENGLISH TEXTBOOKS FOR VIETNAMESE TERTIARY EFL LEARNERS

Introduction

Textbooks play a crucial role in shaping the lexical competence of second language (L2) learners, particularly in English as a Foreign Language (EFL) context, where exposure to authentic language use is limited. Among the key aspects of vocabulary learning, incidental vocabulary acquisition, where vocabulary is not the primary target of the activities (Hulstijn, 2003), has been widely recognized as a fundamental process in language development. This type of learning occurs when learners encounter unfamiliar words while focusing on meaning rather than memorization, making it essential for vocabulary expansion and long-term retention. Given that textbooks are one of the primary sources of linguistic input for EFL learners (Kodriyah et al., 2018), their design should facilitate incidental vocabulary learning by ensuring sufficient lexical coverage and repetition of new words. The selection and lexical composition of reading texts in textbooks are of paramount importance in determining the extent to which learners can expand their vocabulary knowledge (Le & Dinh, 2022). A well-structured textbook should contain texts that introduce new words in a meaningful context and provide enough repetitions to reinforce learning. Without adequate repetition, the reading passages might hinder vocabulary uptake (Bergström et al., 2022).

One way to examine the lexical appropriateness of textbooks is through corpus-based analysis, which allows for a systematic investigation of vocabulary frequency and repetition. Corpus tools can assess the lexical coverage of a text by determining how many words belong to high-frequency, mid-frequency, and low-frequency word lists, as well as whether new words are repeated enough to facilitate retention (Bergström et al., 2022).

While most studies focus on vocabulary coverage in K-12 EFL textbooks (Alsaif & Milton, 2012; Bergström et al., 2022; Nguyen, 2021; Rahmat & Coxhead, 2021; Sun & Dang, 2020), research on university-level materials, especially in Vietnam, remains limited. In addition, few studies examine whether textbooks are designed to promote effective incidental learning (Bergström et al., 2022; Nguyen, 2021). This study addresses this gap by conducting a corpus-based analysis of the reading texts in the 3 textbooks named "Inside Reading" (level B1, B2, C1). In the present study, we focused on analyzing reading passages. Exercises, focus vocabulary, and supplementary activities were not included in the analysis because they are often discontinuous, highly intentional, and tend to favor explicit

vocabulary learning over incidental learning. This approach has also been employed in some previous studies (see Le & Dinh, 2022; Sun & Dang, 2020), implying that this is a reasonable and accepted approach in the field of vocabulary research.

The study aims to investigate how vocabulary input in English language textbooks corresponds to the vocabulary size of Vietnamese tertiary EFL learners in an EFL context. Specifically, the study examines the relationship between Vietnamese EFL university learners' vocabulary size and the lexical demands of their English reading materials and evaluates the repetition rate of novel words to determine their potential contribution to incidental vocabulary learning. The following research questions (RQs) are addressed in this study: (1) What is the vocabulary size of the Vietnamese EFL tertiary learners? (2) What is the vocabulary profile of the English language textbooks used by these learners? (3) To what extent does the learners' vocabulary size provide adequate coverage of the vocabulary found in the textbooks? (4) To what extent do the textbooks support incidental vocabulary learning through word repetition? By providing insights into the vocabulary input of these textbooks, the study's ultimate goal is to inform educators and material developers about the effectiveness of reading texts in facilitating vocabulary acquisition in a Vietnamese EFL context.

1. Literature review

1.1. Lexical coverage and comprehension

Vocabulary is a strong, if not the most reliable, indicator of reading proficiency. Lexical coverage, also known as text or vocabulary coverage, refers to the proportion of words in a text that readers recognize (Laufer & Ravenhorst-Kalovski, 2010). For instance, if a text has 70% lexical coverage, readers understand 70% of its words. A higher percentage indicates improved comprehension. However, a major question is what lexical coverage can foster reading comprehension without assistance.

Many studies have been conducted to find the answer to the question above. Laufer (1989) was the very first to introduce the concept of lexical thresholds in reading, which refers to the percentage of words learners must know to achieve reading comprehension. If reading material has 100 running words and only five of those words are unfamiliar to the reader, the lexical coverage of the text would be 95%. Until now, the term "adequate comprehension" has not been defined clearly, and what is viewed as sufficient in one study may differ significantly in another (Nurmukhamedov & Webb, 2019). Early research by Laufer (1989) proposed that a 95% lexical coverage suffices for reading comprehension. However, a subsequent study by Hu and Nation (2000) suggested that aiming for 98% lexical

coverage is more effective. What remains consistent among lexical coverage studies is that the more words readers know in a text, the better they can comprehend it. Moreover, it is a major misconception to view that 100% lexical coverage can guarantee reading comprehension. Hu and Nation (2000) discovered that at 100% coverage, only two out of 17 participants in their study achieved perfect scores on their reading comprehension test. Laufer and Ravenhorst-Kalovski (2010) later argued that the level of lexical coverage required should align with the expected comprehension depth. They indicated that 95% coverage might serve a 'minimally acceptable' understanding, but achieving 98% coverage could lead to a more precise and comprehensive understanding of the material. These threshold figures have implications for vocabulary instruction and textbook design. If texts exceed the lexical threshold of learners, comprehension may suffer. In our study, 95% and 98% vocabulary coverage are used as the minimal and optimal thresholds that can foster comprehension and incidental vocabulary in reading.

1.2. Vocabulary profile

A vocabulary profile refers to the distribution of words in a text based on their frequency using a reference corpus that represents the language variety used in the target context. According to Milton (2009), the term "*word*" is commonly used, likely for the sake of convenience. However, in a more specialized linguistic context, it refers to distinct concepts such as: (1) types, (2) tokens, (3) lemmas, and (4) word families. (1) Tokens represent the total word count in a text, while (2) types count unique words. In "*The boy stood on the burning deck*", there are seven tokens but only six types, as "*the*" appears twice. (3) A lemma refers to a base word (headword) along with its inflected forms, all sharing the same part of speech (Nation, 2013). English has eight inflectional categories, such as plural nouns, past tense verbs, and comparatives. For instance, *"break," "breaks," "broke," "broken",* and *"breaking"* form a single lemma because they are inflectional variants of the verb *"break"*. Learning one form of lemma can often facilitate the recognition and understanding of its other inflected forms. (4) A word family is a broader unit. It includes the lemma and its derivational forms, which may belong to different parts of speech. For example, the word family of *"access"* includes not only its inflected forms (*"accesses," "accessing")* but also derived forms like *"accessible"* and *"accessibility"*. Word families present a larger learning burden than lemmas but offer greater vocabulary expansion once the base word is known. In this study, word family is used as a counting unit, as the participants are university students with sufficient understanding of prefixes and suffixes.

1.3. Categorizing words using frequency

Vocabulary frequency classification helps educators prioritize words for

instruction. According to Nation (2013), vocabulary can be classified into three categories based on frequency: (1) high-frequency words, (2) mid-frequency words, and (3) low-frequency words. High-frequency words, including the first and second most common 1,000-word families, account for approximately 80% of the vocabulary load in various written materials (Le & Dinh, 2022). Given their significance, teachers and students should prioritize these words, as mastering them enhances comprehension of a substantial portion of reading materials (Nation, 2013). Mid-frequency words, though less common, occur frequently enough to be a practical learning goal after high-frequency words. They consist of the next seven thousand word families, ranging from the third to the ninth 1,000. Finally, low-frequency words, such as *"illocutionary"*, are the least encountered by learners and include the 10th 1,000-word family and beyond. These are the words that rarely appear in daily life.

1.4. Incidental vocabulary learning

If a word in a text is unknown to a reader, incidental vocabulary learning may be feasible. According to Nation (2013), incidental learning is often viewed as something in contrast to intentional vocabulary learning, where words are deliberately acquired through conscious effort and study. However, they claim that this is an unfortunate point of view (as they are complementary activities. They further argued that a well-designed learning program should involve vocabulary learning opportunities from both meaning-focused input and direct learning of vocabulary items (Nation, 2013,).

Defining incidental vocabulary learning remains a challenge as the construct can be operationalized in different ways depending on the studies (Uchihara et al., 2019). A way to define this notion is that incidental vocabulary learning happens when learners are not informed of an upcoming vocabulary test (Hulstijn, 2003). In the presence of a follow-up test, intentional learning will occur instead because learners pay special attention to vocabulary. Another common definition of incidental vocabulary learning is a 'byproduct' of meaning-focused activities (Hulstijn, 2003; Uchihara et al., 2019). In other words, vocabulary is not the main target of the activities (Nation, 2013). In this study, we consider conceptualizing incidental learning as the unintentional acquisition of vocabulary through tasks primarily designed to promote comprehension.

A key context in which this type of learning occurs is extensive reading, especially when it incorporates both meaning-focused and form-focused dimensions (Huckin & Coady, 1999). Extensive reading supports incidental vocabulary learning through several interrelated factors. Firstly, readers must be able to recognize a great percentage of the surrounding words for successful word inference. 95% vocabulary coverage is

needed to attain general comprehension, while 98% is required to achieve full comprehension (Huckin & Coady, 1999). Secondly, unknown words presented within meaningful contexts support accurate guessing of their meanings (Huckin & Coady, 1999). Lastly, while the more often a new word is encountered, the more likely students are to retain the meaning of that word, there is no agreement on the exact number of exposures for incidental learning to occur. There are various factors to consider, such as word salience, its recognizability as a cognate, the learners' engagement, and the contextual richness (Huckin & Coady, 1999).

Promoting incidental vocabulary learning is essential because it helps learners reach the extensive word knowledge needed for comprehension—around 8,000–9,000 word families for reading and 4,000 for listening (Webb, 2019), which is impractical to achieve through intentional learning alone. Additionally, it enhances vocabulary depth by fostering an understanding of word structure, meaning, and collocation through extensive exposure (Webb, 2019).

In a study, Vidal (2011) found that reading is a superior source of incidental vocabulary learning than listening. Reading allows them to dwell upon words or reread if needed. On the other hand, incidental learning in learning is more challenging due to the transitory nature of listening and the obstacle of L2 speech segmentation (Uchihara et al., 2019).

Incidental vocabulary learning in reading can be affected by both reader-related (e.g. proficiency, vocabulary size, decoding ability) (Nation, 2013) and text-related factors (e.g. word repetition, quality of contextual clues) (Nguyen, 2021). In this study, we focus on word repetition as an important factor contributing to incidental vocabulary learning (Uchihara et al., 2019).

1.5. The role of repetition in incidental vocabulary learning

Repetition is one of the key determinants of incidental vocabulary learning, as repeated exposure strengthens long-term retention. This is particularly important in EFL settings, where opportunities to engage with the target language outside the classroom are scarce, and instructional time is limited. Uchihara et al. (2019) conducted a meta-analysis of correlational studies on the complex relationship between word repetition and second language incidental vocabulary learning. 45 effect sizes from 26 studies (N=1,918) reporting correlations between repetition and learning outcomes were synthesized. The overall analysis yielded a medium effect size (r = .34). Learner characteristics (age, vocabulary knowledge), treatment features (spaced learning, visual support, engagement, range of encounters), and methodological factors (nonword use, forewarning of comprehension tests, test format) are identified as significant sources of variability in repetition effects.

There has not been a consensus on the exact number of encounters needed for incidental vocabulary learning in reading to occur. The seminal study by Saragi et al. (1978) was conducted on a group of native English speakers required to read a novel called *A Clockwork Orange* by Anthony Burgess. This novel was written in English but contained some Russian slang words. The study found that the vocabulary uptake of Russian was more consistent at 10 or more encounters. However, no minimum number could guarantee successful learning. Investigating the effects of reading a graded reader and vocabulary learning, Waring and Takaki (2003) proposed a minimum of 20 encounters as the appropriate threshold to acquire new words successfully. Although target words were encountered more than 18 times, learners had only a 10-15% chance of recalling the word meaning after three months. Investigating vocabulary repetition in short texts and incidental lexical learning, Rott (1999) revealed that two encounters with unfamiliar words produced considerable gains, and six encounters produced much greater vocabulary uptake than two or four encounters. Therefore, the study proposed that six encounters are the minimum number to result in substantial learning gains.

Reviewing the literature reveals a wide variation in the number of encounters needed for lexical gains: 6 (Rott, 1999), 10 (Saragi et al., 1978), and 20 (Waring & Takaki, 2003). In our study, we consider 6 to be the minimum threshold. This is similar to the design of the study of Nguyen (2021). We followed this study as it is updated and used for the Vietnamese context.

1.6. Previous studies

Several studies have examined vocabulary profiles in EFL textbooks, primarily for K-12 students (e.g., Alsaif & Milton, 2012; Sun & Dang, 2020; Rahmat & Coxhead, 2021; Bergström et al., 2022). A recurring issue is that many textbooks exceed the lexical capacities of learners.

Alsaif and Milton (2012) investigated the vocabulary content of 22 textbooks from Year 6 to Year 12 in Saudi Arabian public schools. The result of this study aimed to help explain the students' low vocabulary scores when using these textbooks. RANGE program (Heatley et al., 2002) was used to explore the vocabulary profile of the textbooks. This program displays how words are distributed across two or more texts. The program *text_lex_compare* (http://www.lextutor.ca/text_lex_compare) was used to identify how much of the Minister of Education (MoE)'s word-list was included in the textbook vocabulary. The result revealed the shortcomings of the textbooks used in Saudi public schools. These textbooks cover only 84% of the most frequent 2,000 words and 55% of the most frequent 5,000 words, suggesting that they do not adequately support comprehension. Only 2.75% of the words on the MoEd word lists are represented in these books. These

textbooks do not present appropriate vocabulary to help learners achieve the required level of coverage for better comprehension, which explains the students' low vocabulary scores. In the context of China, however, Sun and Dang (2020) investigated the vocabulary in high school EFL textbooks. The study examined a 273,094-word corpus using the RANGE program (Heatley et al., 2002) and measured the vocabulary knowledge of 265 students using the Updated Vocabulary Levels Test (Webb et al., 2017). The findings suggested that achieving 95% and 98% coverage in Chinese high school textbooks required knowledge of 3,000 and 9,000 word families, respectively, while most students had only mastered the first 1,000 words. This means that the materials appear too demanding in terms of vocabulary.

Within the Indonesian educational setting, Rahmat and Coxhead (2021) analyzed the vocabulary profile in three EFL textbooks used by high-school students in Indonesia using RANGE program (Heatley et al., 2002). The findings indicated that the textbooks are challenging for students as much vocabulary is required. To be more specific, Indonesian EFL textbooks require knowledge of 5,000–6,000 words for 98% comprehension, but students typically know only 2,000. On a positive note, frequent word repetition for high-frequency words increases word retention chances. Focusing on Swedish intermediate EFL textbooks, Bergström et al. (2022) used Sketch Engine and Compleat Web VP (Cobb, n.d) to analyze the frequency distribution and word repetition. The results revealed that the texts offer appropriate input for vocabulary learning by exposing students to unknown and mid-frequency words, but insufficient recycling of new items limits vocabulary development, as familiar words are recycled instead. It is concluded that even though the input is suitable for the target students, they are not designed in a way that supports vocabulary development.

While all the papers reviewed above demonstrated a clear trend in the existing literature that the vocabulary in EFL textbooks are too demanding for their respective learners, research by Konstantakis and Alexiou (2012) was one of the few studies in which the lexical profile in the textbooks failed to meet the vocabulary learning targets. Konstantakis and Alexiou (2012) investigated five EFL textbooks for Greek primary students in the first two academic years, which were intended to improve students' proficiency to the CEFR A2 level. Using the RANGE program (Heatley et al., 2002), the researchers discovered that the vocabulary load failed to achieve the learning target as it heavily focused on the most frequent vocabulary. Analysis, based on the British National Corpus (BNC) 2,000 word list, reveals that the course books include between 74% and 85% of the most frequent words. Furthermore, the course books

display significant variability in vocabulary presentation, suggesting that learners transitioning between schools using different books may encounter difficulties. There is no explanation in any of the course books regarding the selection or variability of vocabulary items.

In the context of Vietnam, corpus-based analysis research on vocabulary in textbooks remains limited. Nguyen (2021) investigated the vocabulary profile of the main reading materials in the new high school textbook series (*Tieng Anh 10, 11, 12*) to discover whether these passages fostered reading comprehension and incidental vocabulary learning. Using Vocabprofilers of the Lextutor.ca website to conduct the frequency-based analysis and the Updated Vocabulary Levels Test (Webb et al., 2017) to measure the vocabulary knowledge of 422 high school students, Nguyen (2021) discovered that with a receptive knowledge of the first two 1,000-word lists, students could achieve 95% lexical coverage in only three out of 30 reading texts. In other words, students may struggle while reading as they encounter a high number of unfamiliar words. There were few vocabulary acquisition opportunities as the novel words rarely reappeared. Taking a closer look at both the written and spoken corpus of *Tieng Anh 10,* Le and Dinh (2022) conducted a frequency-based analysis of word families using the Vocabprofilers of the Lextutor.ca website with a corpus of 41,137 words. The results revealed that a vocabulary size of 3,000 and 5,000 word families was required to reach 95% and 98% coverage, respectively. This appears to be too demanding for Vietnamese high school students.

2. Methodology

2.1. Research setting and participants

The study was conducted in a public multi-disciplinary university in Hanoi. In this context, English is taught as a major language. A total of 86 students participated in this study. They were intermediate Bachelor of English Language (*Cử nhân Ngôn ngữ Anh*) students. They were taking a reading course to improve their English reading ability. Before university, they studied English for at least eight years in primary, secondary, and high school. All participants were native speakers of Vietnamese. They were either freshmen or sophomores. Convenience sampling was used to select participants who voluntarily participated in the study.

Three *'Inside Reading'* textbooks (B1, B2, C1) were chosen to be the material for the Reading course. Each book contains 10 units on different content areas.

2.2. Design of the study

Corpus linguistics is the study of large amounts of real language using computer-assisted methods. The data used in this field of study is called a "corpus", which is a computer-readable collection of text. Due to the improvement of technology over the

years, many vocabulary profiling tools have been developed, such as RANGE program (Heatley et al., 2002) and Vocabulary Profiler. They can count the number of times a word appears in a corpus or many corpora, classify it into a high, mid-, or low-frequency wordlist, and identify the cut-off points of 95% and 98% lexical coverage. These tools demonstrate strong validity and reliability, as they are built on research-driven algorithms with minimal reliance on subjective judgment during the process of using these tools.

This study adopted a quantitative, corpus-based research design to examine the relationship between Vietnamese EFL university learners' vocabulary size and the lexical demands of their English reading materials. To examine learners' vocabulary size (RQ1), vocabulary size tests were used. In our study, the Vocabulary Size Test by Nation and Beglar (2007) was employed. According to Nation and Beglar (2007), this is a reliable and accurate measure of a learner's receptive vocabulary size from the first to the 14th 1,000 word families of English. The test consists of 140 items (ten from each 1000-word level). Each correct response was interpreted as knowledge of 100 word families. Learners' total vocabulary sizes were calculated by multiplying the number of correct responses by 100.

The reason why this test was used for this study was that it covered up to the 14th 1,000 word families of English, while other existing vocabulary tests only covered up to the 5th 1,000 word families, such as the updated Vocabulary Levels Test (VLT) developed and validated by Webb et al. (2017). Using the Vocabulary Size Test by Nation and Beglar (2007) for university students in this context was the most suitable option because students majoring in English were likely to have a large vocabulary size. To examine the vocabulary profile of reading texts in *'Inside Reading'* textbooks, vocabulary profiling tools were used. Our study adopted the Vocab profilers of the Lextutor.ca website, as it could classify vocabulary based on frequency and identify the 95% and 98% vocabulary coverage to answer RQ2. LexTutor does not currently have the function of automatically recognizing fixed expressions or collocations. Instead, this software provides parameters such as tokens, types, word families, and frequency-based word lists.

Together, the data about learners' vocabulary size and vocabulary profile were combined to answer RQ3. These tools reveal the number of times a word appears in a corpus for RQ4 as well.

2.3. Data collection and analysis

For RQ1, the Vocabulary Size Test by Nation and Beglar (2007) was administered to 86 Vietnamese university students majoring in English. The result of the test was collected and analysed using Google Forms for convenience. To address RQ2, a frequency-based analysis was

performed using the *British National Corpus* (BNC)/*the Corpus of Contemporary American English* (COCA) lists from Lextutor.ca's Vocabprofilers. This list was chosen for the study as it is the most up-to-date list available and it was created using the largest natural language corpus to date. The analysis determined the lexical coverage required for reading comprehension and identified words that are considered unfamiliar to students. Reading texts were manually compiled to create a corpus. Compounds were separated into single lexical items. Proper nouns (e.g., *Sarah, Africa*) were eliminated from the corpus as they can be assumed to be known by readers already. There are 60 reading passages from '*Inside Reading B1-C1'*. The corpus contains 60,393 running words. Word family is the counting unit. To answer RQ3, learners' vocabulary sizes were compared with the lexical demands of the textbooks. And for the last question, the Vocabulary Size Test revealed the word families that were unknown to the students. Lextutor.ca's Vocabprofilers could provide evidence for the repetition of these word families.

3. Findings and Discussion

RQ1: What is the vocabulary size of the Vietnamese EFL tertiary learners?

Table 1 presents descriptive statistics for the participants' scores on the Vocabulary Size Test

Table 1. The scores of the participants in the Nation and Beglar's Vocabulary Size Test (2007)

Mean	79.74 / 140 points
Min	21/140 points
Max	139/140 points
Range	21 -139 points

The mean score was 79.74 out of 140 points, which corresponds to an estimated receptive vocabulary size of approximately 7,974 word families. The minimum score was 21 points (≈2,100 word families), while the maximum score was 139 points (≈13,900 word families). The range of scores extended from 21 to 139, indicating considerable variability in learners' vocabulary sizes. To conclude, on average, the receptive vocabulary knowledge of the participants in our study is approximately 8,000 words. This means that words from the 9th 1,000 word families and beyond are considered unknown words for students.

Figure 2 shows the distribution of test scores across 86 participants.

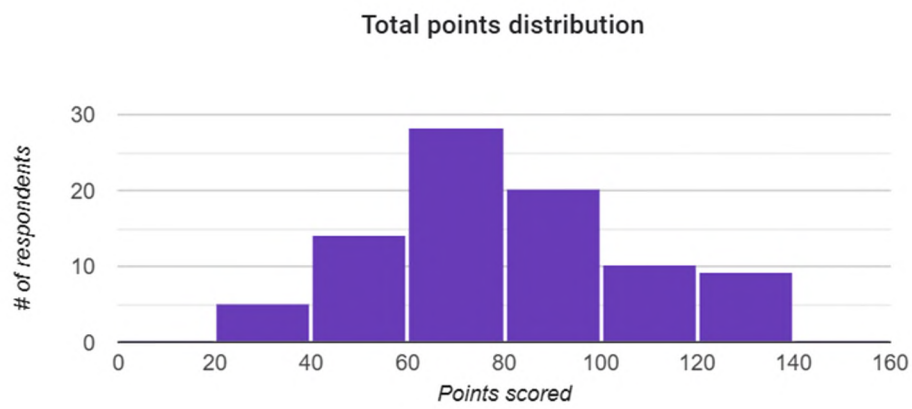

Figure 2. Total points distribution of Nation and Beglar's Vocabulary Size Test (2007)

The participants in this context had a much larger vocabulary size than the Vietnamese students in Nguyen (2021). The Vietnamese students in Nguyen (2021) had a receptive vocabulary of 2,000 most common word families. The explanation for this is that in Nguyen (2021), the participants were high school students selected from different geographical and socio-economic areas of Vietnam, while the participants in our study were university students majoring in English at a university specializing in foreign languages.

RQ2: What is the vocabulary profile of the English language textbooks used by these learners?

Table 2 presents the vocabulary coverage of 60 reading passages across three levels (B1–C1) of the 'Inside Reading' textbook series, based on word frequency levels derived from BNC/COCA lists. The data includes word families, types, and tokens across 25 frequency bands (K-1 to K-25).

Table 2. Vocabulary coverage of reading passages in three '*Inside Reading*' textbooks (B1-C1) (N=60)

Freq. Level	Families (%)	Types (%)	Tokens (%)	Cumul. token (%)
K-1:	911 (27.0)	2176 (35.80)	30961 (77.044)	**77.0**
K-2:	741 (21.9)	1475 (24.27)	4488 (11.168)	**88.2**
K-3:	694 (20.6)	1144 (18.82)	2545 (6.333)	**94.5**
K-4:	350 (10.4)	451 (7.42)	743 (1.849)	**96.4**
Coverage 95				
K-5:	192 (5.7)	234 (3.85)	444 (1.105)	**97.5**
K-6:	133 (3.9)	153 (2.52)	215 (0.535)	**98.0**

Freq. Level	Families (%)	Types (%)	Tokens (%)	Cumul. token (%)
Coverage 98				
K-7:	89 (2.6)	95 (1.56)	163 (0.406)	**98.4**
K-8:	69 (2.0)	74 (1.22)	128 (0.319)	**98.8**
K-9:	43 (1.3)	48 (0.79)	99 (0.246)	**99.0**
K-10:	38 (1.1)	47 (0.77)	109 (0.271)	**99.3**
K-11:	25 (0.7)	30 (0.49)	39 (0.097)	**99.4**
K-12:	18 (0.5)	18 (0.30)	21 (0.052)	**99.4**
K-13:	16 (0.5)	18 (0.30)	32 (0.080)	**99.5**
K-14:	11 (0.3)	13 (0.21)	22 (0.055)	**99.6**
K-15:	10 (0.3)	11 (0.18)	21 (0.052)	**99.6**
K-16:	12 (0.4)	14 (0.23)	24 (0.060)	**99.7**
K-17:	4 (0.1)	4 (0.07)	5 (0.012)	**99.7**
K-18:	1 (0.0)	1 (0.02)	3 (0.007)	**99.7**
K-19:	6 (0.2)	6 (0.10)	13 (0.032)	**99.7**
K-20:	7 (0.2)	7 (0.12)	10 (0.025)	**99.7**
K-21:	4 (0.1)	4 (0.07)	6 (0.015)	**99.8**
K-22:				**-**
K-23:	3 (0.1)	3 (0.05)	3 (0.007)	**99.8**
K-24:				**-**
K-25:				**-**
Off-List:	??	51 (0.84)	91 (0.23)	**100.00**
Total (unrounded)	3377+?	6078 (100)	40186 (100)	≈100.00

Note: K = 1,000

The reading materials consisted mostly of high-frequency words (K-1 and K-2) (48.9%) and mid-frequency words (K-3 to K-9) (46.5%). Low-frequency words (K-10 and beyond) accounted for 4.5%.

The first 1,000 most frequent word families (K-1) contribute the largest proportion, accounting for 27.0% of all word families found in the texts. The next two frequency bands, K-2 and K-3, contribute an additional 21.9% and 20.6%, respectively. Together, the top 3,000 word families make up approximately 69.5% of all word families present. From K-4 to K-

6, the contribution decreases progressively, with K-4 at 10.4%, K-5 at 5.7%, and K-6 at 3.9%. Notably, by K-6, 89.9% of all families in the texts are covered. Beyond K-6, the contribution of each subsequent 1,000-word band becomes minimal, often below 3%. For instance, K-7 and K-8 account for only 2.6% and 2.0%, respectively. The lower-frequency bands (K-10 to K-20) contribute less than 1% each, suggesting the presence of relatively rare or specialized vocabulary.

To achieve the minimal threshold of 95%, knowledge of the first 4,000 most common words is needed. As for the optimal threshold for reading comprehension of 98%, students need to know at least 6,000 most common words.

RQ3: To what extent does the learners' vocabulary size provide adequate coverage of the vocabulary found in the textbooks?

With the receptive knowledge of 8,000 words, the target learners can certainly comprehend the reading materials in the textbooks. These findings are incongruent with other studies (e.g. Le & Dinh, 2022; Rahmat & Coxhead, 2021; Sun & Dang, 2020), which reported that the input vocabulary was too difficult for the current level of the intended learners. This contrast can be attributed to the difference in the research participants of our study and other papers. The participants of this current study were university students majoring in English; therefore, they were more likely to have a large vocabulary size. Meanwhile, for other studies, such as Rahmat and Coxhead (2021), the participants were regular high school students. However, it remains unclear how challenging the reading materials in *'Inside Reading'* are for the students. As the reading texts can foster comprehension, it may be possible for incidental vocabulary learning to happen.

RQ4: To what extent do the textbooks support incidental vocabulary learning through word repetition?

Table 3 presents the frequency of novel word families found in the target reading texts, along with some examples of those that appear at least six times within or across the texts.

Table 3. Frequency of unknown word families (N = 198)

Frequency	**1-5**	**>6**
Number of word families	189	9
Examples	cartography meteorological decelerate	Afar Prodigy Coaster

As learners in our study had a receptive knowledge of 8,000 words, word families from K-9 and beyond were considered unknown word families, which may be learned incidentally.

To illustrate, the word family *'prodigy'* belongs to the List 9 in the BNC-COCA

corpus, which was supposed to be novel to students in this context. It appears 18 times in the *'Inside Reading 3'*. Therefore, this word can likely be learned incidentally, as repetition of a word can capture learners' attention, which serves as a foundation of incidental vocabulary learning. To be more specific, this word appears 17 times in Unit 8, Reading 1, and one time in Reading 2 in *Inside reading 3.*

The occurrence of novel word families meeting this threshold was remarkably low, only 9 out of 198 (approximately 4.5%). This indicates that the repetition of novel word families falls significantly below the levels recommended by previous research for effective incidental vocabulary acquisition. One possible explanation is that the number of novel words in the corpus is quite small (5.8%). Therefore, there are few chances for incidental vocabulary learning to happen. This finding is similar to the result of the studies conducted by Bergström et al., (2022) and Nguyen, (2021). Low levels of vocabulary repetition are common in EFL materials worldwide, regardless of age group (Bergström et al., 2022). One possible explanation for this is that the 'Inside reading' textbooks provided learners with a wide range of different topics. For example, in 'Inside reading 3', unit 3 was about art and design, unit 4 are about public health and unit 5 is about film studies. Each unit consisted of 2 reading materials. Therefore, it is unlikely that topic specific vocabulary can be repeated enough in a textbook to foster incidental vocabulary learning.

4. Conclusion

In conclusion, the findings of this study offer insights into the lexical demands of the *Inside Reading* textbook series and their alignment with the vocabulary knowledge of Vietnamese EFL university learners. First, the participants demonstrated a mean receptive vocabulary size of approximately 8,000 word families, which is notably larger than those reported in previous studies involving high school students. Second, the vocabulary profile of the textbooks revealed that the reading texts are composed largely of high- and mid-frequency words, with 95% coverage achieved at the 4,000 word level and 98% at the 6,000 word level. Third, given their vocabulary size, the learners are well-equipped to comprehend the reading materials, as they significantly exceed the lexical thresholds required for adequate comprehension. However, the opportunities for incidental vocabulary learning are limited. Only 9 out of 198 unknown word families (4.5%) appeared six times or more, indicating insufficient repetition to support incidental acquisition. Overall, while the textbooks are suitable in terms of readability, their design does not optimally support vocabulary growth through incidental learning.

Based on these findings, the following suggestions are proposed. As for learners in this context, these textbooks should

continue to be used as they are suitable for the students. As for teachers, they can ensure additional repetitions of important vocabulary items by revisiting words in class discussions or assigning extra tasks. As for textbook designers, they should take into account the vocabulary frequency when writing reading texts for EFL students. Online vocabulary profilers can be used to analyze their texts from a vocabulary standpoint. Additionally, designing additional tasks or supplemental materials to create more opportunities for word recycling is recommended.

This study is not without limitations. Firstly, it highlights the importance of applying corpus-based methods to analyze reading materials for language learning. However, the research relies solely on a corpus-based analysis of reading texts. While this approach provides valuable quantitative data, it does not capture qualitative aspects such as contextual richness, which may be important for incidental vocabulary learning. Secondly, the participants of this research are Vietnamese university students majoring in English with a relatively high vocabulary size. The findings cannot be generalized to all EFL contexts in Vietnam as it does not reflect the performance of learners at different English proficiency levels (e.g., non-English majors). Finally, this analysis does not take word lists and exercises into consideration. Future researchers are encouraged to conduct more corpus-based analyses of teaching materials. Moreover, further research on how teachers utilize textbooks for teaching vocabulary is essential to understanding the role of textbooks in students' language learning.

REFERENCES

1. Alsaif, A., & Milton, J. (2012). Vocabulary input from school textbooks as a potential contributor to the small vocabulary uptake gained by English as a foreign language learner in Saudi Arabia. *The Language Learning Journal*, *40*(1), 21–33. https://doi.org/10.1080/09571736.2012.658221

2. Bergström, D., Norberg, C., & Nordlund, M. (2022). Do textbooks support incidental vocabulary learning? – A corpus-based study of Swedish intermediate EFL materials. *Education Inquiry*, *16*(1), 69–87 https://doi.org/10.1080/20004508.2022.2163050

3. Huckin, T., & Coady, J. (1999). Incidental vocabulary acquisition in a second language: A review. *Studies in Second Language Acquisition*, *21*(2), 181–193. https://doi.org/10.1017/S0272263199002028

4. Hulstijn, J. H. (2003). Incidental and intentional learning. In C. J. Doughty & M. H. Long (Eds.), *The Handbook of Second Language Acquisition* (1st ed., pp. 349–381). Wiley. https://doi.org/10.1002/9780470756492.ch12

5. Kodriyah, L., Dayu, A. T., & Hakim, A. R. (2018). ELT textbooks pros and cons in EFL environment: How teachers should make a decision. *Intensive Journal*, *1*(1), 1-8. https://doi.org/10.31602/intensive.v1i1.1239

6. Konstantakis, N., & Alexiou, T. (2012). Vocabulary in Greek young learners' English as a foreign language coursebooks. *The Language Learning Journal*, *40*(1), 35–45. https://doi.org/10.1080/09571736.2012.658222

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

8. Le, N. T. M., & Dinh, H. T. (2022). Vocabulary coverage in a high school Vietnamese EFL textbook: A corpus-based preliminary investigation. *Vietnam Journal of Education*, *6*(2), 102-113. https://doi.org/10.52296/vje.2022.187

9. Milton, J. (2009). *Measuring second language vocabulary acquisition*. Multilingual Matters. https://doi.org/10.21832/9781847692092

10. Nation, I., & Beglar. (2007). A vocabulary size test. *The Language Teacher, 31*(7), 9–13.

11. Nguyen, C.-D. (2021). Lexical features of reading passages in English-language textbooks for Vietnamese high-school students: Do they foster both content and vocabulary gain? *RELC Journal*, *52*(3), 509–522. https://doi.org/10.1177/0033688219895045

12. Nurmukhamedov, U., & Webb, S. (2019). Lexical coverage and profiling. *Language Teaching*, *52*(02), 188–200. https://doi.org/10.1017/S0261444819000028

13. Rahmat, Y. N., & Coxhead, A. (2021). Investigating vocabulary coverage and load in an Indonesian EFL textbook series. *Indonesian Journal of Applied Linguistics*, *10*(3), 804-814. https://doi.org/10.17509/ijal.v10i3.31768

14. Rott, S. (1999). The effect of exposure frequency on intermediate language learners' incidental vocabulary acquisition and retention through reading. *Studies in Second Language Acquisition*, *21*(4), 589–619. https://doi.org/10.1017/S0272263199004039

15. Saragi, T., Nation, I. S. P., & Meister, G. F. (1978). Vocabulary learning and reading. *System*, *6*(2), 72–78. https://doi.org/10.1016/0346-251X(78)90027-1

16. Sun, Y., & Dang, T. N. Y. (2020). Vocabulary in high-school EFL textbooks: Texts and learner knowledge. *System*, *93*, 1–40. https://doi.org/10.1016/j.system.2020.102279

17. Uchihara, T., Webb, S., & Yanagisawa, A. (2019). The effects of repetition on incidental vocabulary learning: A meta-analysis of correlational studies. *Language Learning*, *69*(3), 559–599. https://doi.org/10.1111/lang.12343

18. Vidal, K. (2011). A comparison of the effects of reading and listening on incidental vocabulary acquisition. *Language Learning*, *61*(1), 219–258. https://doi.org/10.1111/j.1467-9922.2010.00593.x

19. Waring, R., & Takaki, M. (2003). At what rate do learners learn and retain new vocabulary from reading a graded reader? *University of Hawaii National Foreign Language Resource Center Center for Language & Technology*, *15*(2), 130–163.

20. Webb, S. (2019). Incidental vocabulary learning. In S. Webb (Ed), *The Routledge Handbook of Vocabulary Studies* (1st ed., pp. 225–239). Routledge.

(Ngày nhận bài: 23/5/2025; ngày duyệt đăng: 28/7/2025)