From Textbook to Test: Vocabulary Frequency Analysis to AI-Generated Quizzes

Timothy Ang
Center for Global Education and Exchange
Toyo University

Katsuichiro ‘Ken’ Ohashi
Rikkyo University, Center for Foreign Language and Research (FLER)
(At time of research Toyo University, Center for Global Education and Exchange)

From Textbook to Test: Vocabulary Frequency Analysis to AI-Generated Quizzes

Timothy Ang
Toyo University, Center for Global Education and Exchange
Katsuichiro 'Ken' Ohashi
Rikkyo University, Center for Foreign Language and Research (FLER)
(At time of research Toyo University, Center for Global Education and Exchange)

Abstract

This paper outlines a two-stage approach to vocabulary selection and assessment for an International English Language Testing System (IELTS) preparation course within a Japanese university context. In stage 1, target words from twelve textbook passages were identified by profiling vocabulary against the frequency based lists New General Service List (NGSL) and New Academic Word List (NAWL) using LEXTUTOR. In stage 2, multiple choice quizzes were created using the identified target words from stage 1 and through use of a large language model (LLM). The study investigates both the effectiveness and limitations of using vocabulary profilers and LLMs. Our initial findings suggest that profiling vocabulary offers a theoretical foundation and framework to which AI-assisted methods can reduce relative teacher workload for vocabulary assessment.

Introduction

Vocabulary is fundamental for success in standardized English proficiency tests such as the IELTS. Language learners are aware of this, but due to academic demands of both English and non-English coursework, it is understandable that the learners encounter issues with study strategies such as time management and test preparation. The English language teacher can play an important role in reducing the language learner's time needed to study a language by guiding the learners' attention to words that the learners are most likely to encounter in multiple situations.

For the English language teacher in higher education that lacks theoretical knowledge in second language vocabulary studies, making decisions on what vocabulary they should guide their learners' attention to can be challenging. This is especially true when designing courses around materials that do not provide a list of target vocabulary. The words should be relevant to the context

that is being taught, while not being so content specific that the words will not be very useful in context outside of a given lesson. As such, for low to intermediate proficiency level learners, the words should be on the NGSL and NAWL for efficiently learning words that the learners are most likely to encounter (Browne, 2013). The teacher should also consider words that appear multiple times within the course since repeated exposure is known to be effective for vocabulary acquisition (Rott, 1999; Waring and Takaki, 2003).

Another issue for teachers, whether well versed in second language vocabulary theories or not, is that they may not be aware of openly available tools that can be used to create high-quality vocabulary lists and assessments. Vocabulary profiling technology such as LEXTUTOR *Complete Web-VP v.2.6* (Cobb), combined with the use of formulas on spreadsheet software such as Google Sheets or Microsoft Excel can be a viable solution for teachers to choose and assign target vocabulary for a given course over a semester.

High-quality assessment design in ESL contexts often faces challenges, particularly when instructors must develop additional materials without research-informed templates or textbook supplement material to draw from.

Research has shown that assessments can be prone to various deficiencies. These include potential fairness issues that arise when test tasks introduce factors unrelated to the skill being tested (Xi, 2010) or uneven skill profiles that reflect gaps between instructional focus and test demands (Sasaki, 2022). Gottlieb (2016) cautions that assessments may become linguistically inappropriate for English learners when test items contain complex structures (i.e., unnecessarily complicated instructions or overly dense question wording), thereby obscuring what is actually being measured. Our own departmental faculty experiences revealed issues with biased word selection, questions whose themes were distant or irrelevant in an IELTS context, and misaligned or inconsistent difficulty level of assessment questions.

This tenuous standardization in instructor-created quizzes reinforced the importance of the need for systematic design and planning to create assessments. This study reports on a dual-stage initiative conducted in the LEAP (Language Education for Academic Purposes) program at Toyo University, targeting Japanese university students whose English is at beginner to lower-intermediate levels, corresponding roughly to the Common European Framework of Reference for Languages (CEFR) A2–B1 (IELTS 3.5–5.0). Combining vocabulary profiling and LLMs remains underexplored, especially in test preparation contexts such as IELTS, where lexical range, frequency, and depth must align with exam expectations while maintaining balance with the wider aims of the course or program.

The first stage employed vocabulary profiling tools to construct targeted word lists from IELTS Trainer 2 (Cambridge University Press, 2019), the main textbook used in the course. The second stage involved generating quizzes using a LLM, with implementation supported by Google

Forms and Google Classroom.

Literature Review

Identifying Target Vocabulary

Past studies (Laufer, 1989) have indicated that the language learner would need to know at least 95% of the words in a given text to comprehend it. Van Zeeland and Schmit (2013) suggest that listening comprehension can be achieved with varying levels of success from 90% word coverage of the content. Taking this into consideration, it is reasonable to understand why frequency-based word lists have been compiled and used to identify words that are commonly used by native English speakers.

Among these vocabulary compilations, The General Service List (West,1953) and Academic Word List (Coxhead, 2000) have been used by many researchers and English language educators over the years to analyze texts, create textbooks and graded readers, create tests such as the Vocabulary Levels Test, Vocabulary Size Test (Nation and Beglar,2007), and Computer Adaptive Test of Size and Strength (Laufer,& Goldstein, 2004). Researchers have used word frequency bands of 1,000s to categorize and identify which words language learners of English should focus on acquiring.

It is said that the language learner would be able to identify around 80% of the words in any given text if they know the words in the first and second frequency bands, that is, words that are a member of the 2000 most frequently used word families (Waring & Nation, 1997). Coxhead's (2000) academic word list is a collection of 570 word families that appear in academic texts at a higher rate than in non-academic texts. Mastering these words would raise the word recognition rate close to 90% when reading academic texts, which is very beneficial for students in higher education (Coxhead, 2000).

Despite its high usage by researchers, the General Service List has been criticized for its text coverage rate for modern literature and using a corpus that can be considered limited in size compared to more modern ones (Browne, 2014). In 2013, Brown announced the New General Service List (NGSL) and New Academic Word List (NAWL) as an updated version of the GSL (West, 1953) and AWL (Coxhead, 2000). The NGSL and NAWL together provide close to a 90% coverage for abstracts in academic texts across 12 subject matter divisions of the American Educational Research Association (Hendry & Sheepy, 2018).

Vocabulary profilers have been used by researchers to critique EFL textbooks. Klinger (2004) examined six textbooks used at Japanese elementary and junior high schools and compared the vocabulary used in these books against the JACET8000 wordlist and found that 95% of the words used in these textbooks can be read by learning the first 3,000 words on the list. Nakayama (2022a,

2022b) looked at textbooks used at junior and senior high schools in Japan. He points out that roughly 95 % of the tokens used in the textbooks he examined were covered by the NGSL while the coverage is lower than 92% for senior high school textbooks. Despite its use in vocabulary research, documentation on the use of vocabulary profilers to identify vocabulary for instruction, especially studies that use NGSL and NAWL as the benchmark for vocabulary difficulty are difficult to locate.

Using Large Language Models to Create Tests

Large language models (LLMs) are artificial intelligence systems trained on large datasets to generate human-like language. In this study, we used ChatGPT (GPT-4o; OpenAI, 2024), a publicly available LLM released in 2022. GPT-4o is available in both free and paid tiers and was selected for its accessibility and capacity to generate structured assessment materials. LLMs are typically accessed through a graphical chatbot-style interface on a website or application, which allows users to interact with the model by entering prompts and receiving generated responses.

LLMs in relation to artificial intelligence can be a confusing connection as it is merely utilizing computing power on probabilistic outcomes of existing data. They do not possess understanding in a human sense but instead produce output by predicting likely word sequences based on statistical patterns in language. Therefore the quality and relevance of LLM-generated content depend heavily on two factors: the existing data the LLM has accumulated and user-provided instructions known as prompts. The careful design and testing is referred to as prompt engineering.

The rise and prevalence of large language models (LLMs), introduced new opportunities for teaching materials in education. Teaching materials refer to instructor-selected or created resources used to support learning, including quizzes, worksheets, lesson slides, and class activities. The underlying assumption is that LLMs can generate curriculum-relevant content, thereby supporting curriculum development. Early classroom applications suggest that while LLMs can assist in prompt creation, quiz writing, and feedback generation, they also pose practical and ethical challenges that require careful oversight (Yan et al., 2024). For this paper we focused on the practical uses and challenges.

Although the IELTS does not include a dedicated vocabulary section, vocabulary knowledge is an essential learning component to answering reading and listening multiple-choice tasks. Students who cannot recognize variations in vocabulary are prone to choosing the wrong option even when they understand the main idea of the reading passages, which questions are based on. Chen (2019) found that less proficient learners made higher rates of usage errors when dealing with paraphrases. Vocabulary quizzes can help this weakness by training learners to recognize synonyms, collocations, and paraphrases commonly found in IELTS questions.

It is worth noting that vocabulary assessments represent only one component of lexical retention. Stewart (2024) cautions that recognition-based vocabulary tasks, such as multiple-choice items, may not always provide a complete measure of productive word knowledge. However, in the context of IELTS preparation, recognition tasks remain pedagogically relevant. IELTS reading and listening tasks often require learners to discriminate between near-synonyms, attend to nuanced lexical distinctions, and identify and eliminate distractors under time pressure. In this sense, well-designed multiple-choice vocabulary quizzes may help students practice essential word discrimination skills required in IELTS.

Research Questions:
RQ1: How can vocabulary profilers be used to identify vocabulary that should be prioritized in an IELTS preparation course?
RQ2: How effective are large language models (LLMs) in vocabulary assessment preparation work in terms of productivity gains and test quality output?
RQ3: What are the limitations of using lexical profilers and LLMs for creating vocabulary quizzes?

Methodology

Stage 1: Vocabulary Selection

This first stage concerns the procedures that were implemented when selecting the target vocabulary for instruction for the IELTS preparation course. How LEXTUTOR, Microsoft Word and Excel was used to select the 180 target words from the 10,320 words that appeared in the 12 reading passages from *IELTS Trainer Academic* (Cambridge University Press, 2019) is described.

Phase 1: Preparing Data

As part of the course update, *IELTS Trainer Academic* (Cambridge University Press, 2019) was chosen as the new textbook for the level 1 courses at LEAP. This book is a test book containing six full practice tests. Each test contains three reading passages for a total of 18. IELTS reading tests are designed so passage 1 is the easiest and passage 3 is the most difficult within the test. Level 1 being the lower level within the LEAP program with a good number of students' English proficiencies being at the CERF A1-A2 level, a decision was made to use Passage 1 and 2 from each test as teaching material to be covered during the 15-week semester.

The 12 passages were prepared so the words could be run through a vocabulary profiler. The publisher was generous to provide a digital copy of the textbook that could be used by the teachers. Texts from each of the passages of this digital version were copied onto separate Microsoft Word documents. Microsoft Word was used to speed up the process of locating words misidentified by

the optical character recognition (OCR) technology. Each word that was marked with a red wavy line that indicates a spelling error was checked and corrected.

Phase 2: Identifying Frequency and Range

In order to identify which words from the textbook the teachers should bring to their students' attention, measures were taken to identify words of the appropriate difficulty and importance. Appropriate difficulty was considered as words that have a frequency higher than 1000 on the NGSL and words found on the NAWL. The 1000 most frequently used words were excluded since these words were considered as words all university students should know already. Importance was determined as words that appear in multiple texts within the 12 chosen passages so students will receive exposure to the target words multiple times over the semester to maximize opportunities for uptake.

The 12 passages were run through LEXTUTOR (Cobb) separately to be compared against the NGSL and NAWL. Words that were categorized as 2 (frequency of 1,001-2,000 on the NGSL), 3 (frequency of 2,001-2,800 on the NGSL) and A (New Academic Word List) were extracted onto spreadsheets on Microsoft Excel. The words were placed in individual cells with 10 words in a row for easy counting. Different color fonts were used for each frequency band for easy recognition of the word's frequency. Words from each passage were placed on separate tabs that were renamed to easily understand which passage the words were from (e.g. 'T1P1' for test 1, passage 1). Each of the 12 tabs contained roughly 90-120 candidate words.

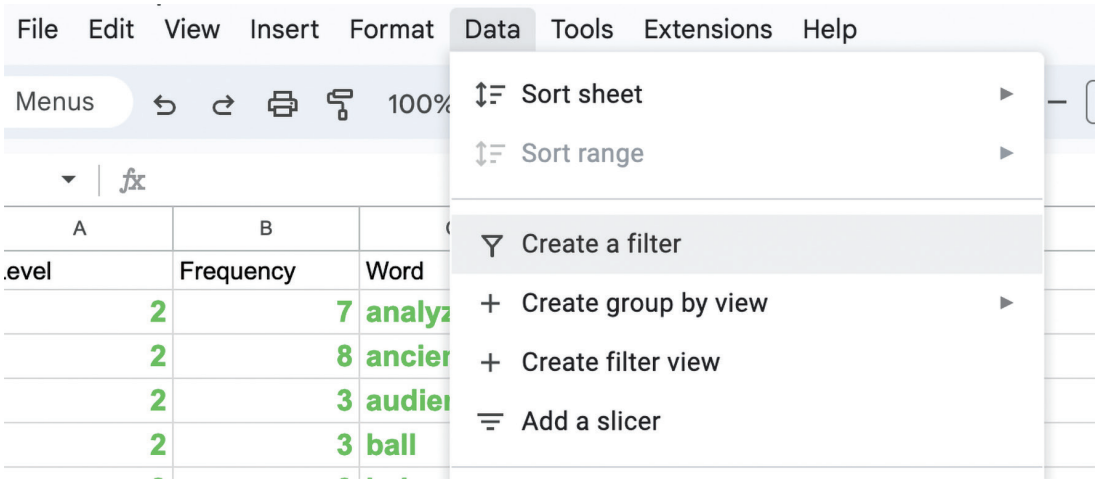
A combined list of all words from the 12 tabs were run through LEXTUTOR (Cobb) together to grasp the range of each word. If LEXTUTOR identified a word's frequency as 4 with this combined wordlist, this would translate as the word appearing in 4 different passages. The words were placed on a separate tab labeled Overall in the same Microsoft Excel file mentioned previously. All words were placed in a single column with the range listed in the cell directly to the right of the word. Each word was color coded to match the colors used on the 12 other tabs for individual passages. The accumulated frequency of each word was also added to this tab, on the same row as the word in a column labeled total frequency. This statistic was also acquired when running the combined file of all 12-passages, a corpus of 10,340 words, through LEXTUTOR.

Phase 3: Finalizing Word Lists Using Conditional Formatting

Candidate words on each tab were shortlisted using Microsoft Excel's conditional formatting function through a two-step process. The first step of isolating words with a high range score was conducted on the *Overall* tab. Words that appear on more than 3 passages were isolated using the filter function (Figure 1). All words that remained after filtering, along with the information for range and frequency, were copied to each of the 12 tabs at a location that did not interfere with the 90-120 candidate words on each tab.

Figure 1

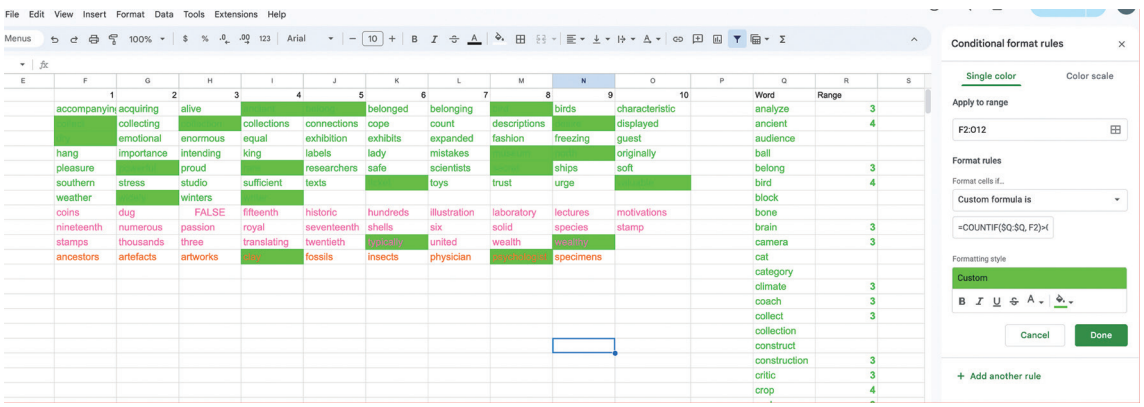
Screenshot of filter creating on Microsoft Excel



The second step was to locate the high frequency words that were identified in the previous step within the 90-120 candidate words on each tab by using a custom formula, "=COUNTIF($Q:$Q, F2)>1" as the conditional format rule. This formula identifies duplicate words within a set area. "QQ" specifies the column where the filtered words from the previous step are copied. "F2" is the first cell of the candidate words. As a result, approximately 20 words were highlighted for each passage (Figure 2). It is worthy to note that since words with a range score of 3 were used, each word was highlighted on more than three tabs.

Figure 2

Screenshot of highlighted candidate words after applying conditional formatting on Microsoft Excel



To standardize the difficulty of the 12 lists of 15-words, a decision was made to include 9 words from band 2, 3 words from band 3 and 3 words from band A, of the LEXTUTOR (Cobb) vocabulary profile results, on each list. Words with high range scores within each band were

prioritized in the selection process. Total frequency was the second criteria that was consulted in the selection process.

The shortlisted words on each tab were placed in a single table on the overall tab across 12 columns, one column of roughly 20 words for each passage. Microsoft Excel's conditional formatting formula "=COUNTIF(C2:C2,C2)=1" was used to identify the first appearance of each word. C2 represents the first cell of the shortlisted words.

As a general rule, each word was to be placed on a list at the earliest possible time in the semester. For example, the word *ancient* appears in four passages, test 1 passage 1, test 3 passage 2, test 4 passage 1, and test 5 passage 1. Out of the four passages, test 1 passage 1 is the passage scheduled to be used first within the semester, therefore the word *ancient* would be placed on the vocabulary list of 15 words for test 1 passage 1. This is to maximize the opportunities for students to encounter the target words consciously. An exception was made to this rule for some words that appeared in more than 5 passages. These were not placed on the list of first appearances to disperse the candidate words more evenly over the semester and still provide students with multiple opportunities of exposure to the word in different passages later in the semester.

Stage 2: Vocabulary Assessment

The second stage involved creating vocabulary assessments from the refined vocabulary list using both manual / human and LLM-assisted methods. Structured prompts were generated through ChatGPT (GPT-4). Twelve vocabulary quizzes were developed to correspond to the twelve textbook passages in *IELTS Trainer 2* (Cambridge University Press, 2019), with the first six created manually by one instructor and the remaining six produced with LLM assistance. This distribution was determined in advance to enable a systematic comparison between human-authored and LLM-assisted quiz construction.

Phase 1: Generating Quiz Questions Using LLM

The first step involved sourcing vocabulary from the master word list file compiled in Stage 1 (see Figure 3). This file served as the data source for creating 15-word subsets aligned with each passage from the course textbook. The subset list of vocabulary for Passage T1P1 was used and labelled as Quiz 1. These subset lists were then uploaded to Google Classroom separately so that students were expected to study all 15 words for each passage.

Figure 3

Screenshot of master word list file

	Quiz 1	Quiz 2	Quiz 3	Quiz 4	Quiz 5	Quiz 6	Quiz 7	Quiz 8	Quiz 9	Quiz 10	Quiz 11	Quiz 12
	T1P1	T1P2	T2P1	T2P2	T3P1	T3P2	T4P1	T4P2	T5P1	T5P2	T6P1	T6P2
1	ancient	audience	bone	brain	experiment	critic	climate	distance	block	coach	decade	crop
2	belong	camera	construction	cycle	global	eventually	directly	hardly	column	intelligence	hole	dry
3	collection	category	left	obvious	highly	frequently	upper	none	heat	implication	lake	engineering
4	museum	ideal	native	powerful	increasingly	professor	warm	outcome	practical	predict	consequence	fuel
5	secret	trend	river	rare	largely	scientific	burn	notion	proportion	objective	decline	output
6	writer	truth	stone	regular	ordinary	spread	farmer	suspect	component	analyze	observe	plenty
7	lady	platform	wood	survive	valuable	ticket	previously	apparently	construct	context	ought	tough
8	text	supporter	researcher	advice	widely	emotion	visitor	convince	engineer	enhance	threat	cheap
9	wealthy	digital	insight	device	emphasize	scientist	concrete	desire	expose	fan	sink	fruit
10	dig	festival	craft	recommend	reject	remarkable	layer	expert	heavy	football	agriculture	seed
11	stamp	historian	impact	motivation	technique	genuine	wooden	progress	sheet	speed	irrigation	vehicle
12	clay	maker	weave	alarm	flexible	laughter	everyday	found	apartment	league	drain	innovation
13	specimen	transform	tribe	rapid	stream	humor	roof	typically	steel	statistic	kilometer	extract
14	artifact	cinema	dye	hormone	influential	psychologist	pipe	honest	tower	algorithm	plug	fossil
15	artwork	depict	fiber	incredibly	limb	contradictory	outer	interestingly	architect	artificial	shallow	crude

From each 15-word list, ten target vocabulary items were manually selected as the correct answers for the quiz. Five of these came from the current passage's list, while the remaining five were drawn from earlier quizzes to reinforce retention. This design required students to recall previously studied vocabulary rather than relying only on the most recent items, which provided additional challenge.

The finalized target words, together with their answer choices, were then fed into an LLM prompt. For each quiz, the correct answers were manually selected from the relevant 15-word subset list (e.g., *decade, decline, sink, drain, shallow, statistic, expose, outer, spread, incredibly*). Distractor options were drawn from previously studied subset lists that had already been quizzed on, ensuring cumulative review. As the semester progressed, this distractor pool expanded naturally, allowing later quizzes to incorporate a broader range of prior vocabulary items.

The standardized prompt format used to generate all LLM-based quizzes is shown below. In the reproduced prompt, square brackets (e.g., [manually selected target vocabulary list], [manually selected distractor pool]) indicate where the researcher inserted the vocabulary selected in the procedures previously described.

LLM Prompt

1. Task Overview
   - Create a 10-question multiple-choice fill-in-the-blank quiz using the provided word lists
   - Each question must have one correct answer and three distractors.
2. Correct Answers
   - The 10 correct answers must be exactly these words: [manually selected vocabulary list]
   - Each correct answer must be used once only.
   - Correct answers must be inserted in a randomized order, not matching the sequence in which they are listed here, and must not appear consecutively or in a block (e.g., Q1–Q5). Ensure they are spread across the full set of 10 questions.

3. Distractors
    - All distractors must come from the provided word list: [manually selected distractor pool]
    - Distractors cannot repeat within the same quiz.
    - Distractors should normally match the part of speech of the correct answer, but up to three per quiz may deliberately differ if they remain grammatically plausible in the sentence.
    - Distractors must be grammatically correct and semantically possible in the sentence, but clearly wrong because they shift meaning along a different dimension (e.g., speed vs. significance, frequency vs. degree).
    - Distractors should be close in meaning, academic in tone, or thematically related, creating an IELTS-style challenge.
    - Avoid distractors that are clearly irrelevant (e.g., "ticket" in a science context) or overused examples.
    - Each quiz should contain at least one "productive trap" distractor (such as a*udience vs. critic* or *trend vs. progress*), but no more than three.

Special Parts of Speech Rules
- Adverbs: When the correct answer is an adverb, all distractors must also be adverbs from the provided list (e.g., *highly, frequently, widely, largely, increasingly*).
- Adjectives: When the correct answer is an adjective, distractors should also be adjectives (e.g., *ordinary, remarkable, rare*) and should test subtle differences in degree, intensity, or evaluation.
- Nouns: When the correct answer is a noun, distractors should preferably come from the same semantic category (e.g., trend vs. progress, audience vs. critic, truth vs. notion). However, distractors may also include adjectives or other descriptors (e.g., native, wealthy) if they are grammatically plausible in the sentence but represent a wrong dimension of meaning.
- Verbs: When the correct answer is a verb, distractors should be verbs with related but incorrect meanings (e.g., *belong, spread, survive*), making them plausible in structure but inaccurate in meaning.

4. Question Format
    - Each sentence must reflect IELTS-style content (themes from Reading, Listening, Writing, or Speaking).
    - Use authentic, natural English suitable for IELTS Band 6–7 level.
    - Each item must allow only one grammatically and semantically correct answer.
5. Output
    - Provide clean, unformatted plain text suitable for Google Forms.
    - Do not use bullet points or letter labels for the choices.

- For each question, present the stem followed by four answer options listed one per line.
- At the end, provide a clearly labeled Answer Key in this format:

Answer Key
1. correct answer
...
10. correct answer

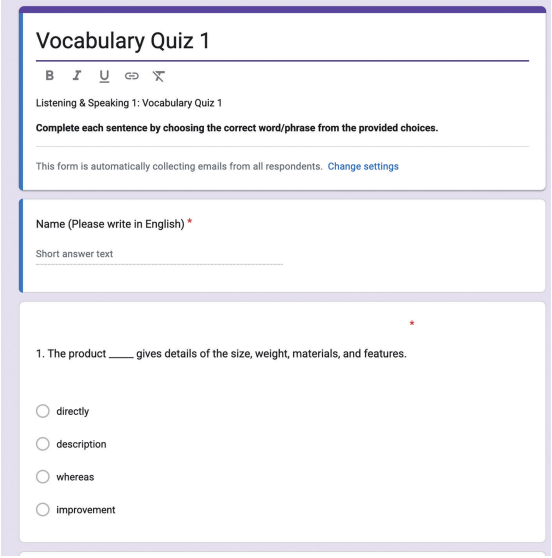
Phase 2: Google Forms Creation

The quiz questions generated in Phase 1 were transferred into Google Forms. During this phase, some answer choices were refined, with weaker distractors replaced by alternatives selected from the master word list. Once finalized, the multiple-choice fill-in-the-blank quizzes were published and shared via Google Drive for distribution to instructors, who then linked the quizzes to their individual Google Classroom pages (Figure 4), ensuring that students could access the material directly. This workflow also enabled a direct comparison of human and LLM-based quiz construction in terms of preparation time.

In the manually created vocabulary quizzes, decisions about distractor refinement were based on instructor judgment rather than formal statistical criteria. At the time of quiz construction, distractors were evaluated intuitively for grammatical fit, semantic plausibility, and perceived difficulty relative to the target vocabulary item. The length and complexity of the fill-in-the-blank stems were not standardized and were determined by instructor judgment during item construction.

Figure 4

Screenshot of Google Form integration



Phase 3: Administering Quiz

A one to two-week gap was implemented between the pre-quiz exposure (announcement of quiz and reminder to students to study) and the actual quiz to minimize short-term recall effects and ensure that performance reflected meaningful vocabulary acquisition. Quizzes were administered every other lesson across the 15-week semester consisting of 30 lessons (two per week). This approach aligns with principles of retrieval practice and spaced repetition, both of which are recognized to enhance long-term vocabulary acquisition by requiring learners to actively recall words across multiple intervals (Roediger & Butler, 2011).

Each quiz had an allotted time of 5–10 minutes. Quizzes were auto-scored through the Google Forms platform, and results were made immediately available to students, allowing them to reflect on their performance. Each vocabulary quiz was worth 1%, for a total of 10% of the overall class grade.

Data Collection

The dataset consisted of responses from two IELTS Reading and Writing courses in the LEAP program at Toyo University during the Spring 2025 semester, with 20 and 18 registered students, respectively. Across both classes, the number of submissions per quiz ranged from 16 to 20, depending on attendance. Quiz 12 had no responses at the time of analysis due to scheduling constraints and was therefore excluded from the dataset.

To address RQ2 (How effective large language models are in vocabulary assessment preparation work in terms of time saved and quality of output), effectiveness was operationalized along two dimensions: (a) productivity and (b) test-item quality.

To measure productivity, the instructor recorded the time required to complete key stages of quiz production for both manually created and LLM-generated quizzes. For manual / human creation, time segments included item construction (drafting fill-in-the-blank questions and selecting correct answers, distractor selection), and entry and formatting of items in Google Forms. For LLM-generated quizzes, timing included prompt submission and model response (automatic production of quiz items and answer options by the LLM), post-editing of output (correcting deviations, refining distractors, and adjusting phrasing), and entry into Google Forms.

It should be noted that substantial upfront time (approximately 30–45 minutes) was required to design (designing task instructions and constraints for quiz generation) and refine the initial prompt used for quiz generation. However, this prompt engineering constituted a one-time setup and was therefore excluded from the per-quiz timing, as the same prompt structure was reused across all subsequent quizzes.

To evaluate test-item quality, we examined student performance outcomes and distractor quality. Student response data were exported from Google Forms for each quiz, including total

scores per item (performance outcomes) and item-level response frequencies for all multiple-choice options. Distractor analysis was conducted for all 110 questions. A distractor was classified as 'functioning' if it attracted at least 5% of student responses, a threshold commonly used in item analysis to distinguish meaningful distractors from infrequently selected options. These metrics were computed separately for manually created quizzes (Quizzes 1–6) and LLM-generated quizzes (Quizzes 7–11) in order to compare test quality across conditions.

All productivity and test-item quality analyses were conducted using Microsoft Excel for data organization, calculation of descriptive statistics, and aggregation of response-level data. Limitations related to the use of lexical profilers and LLMs for quiz creation (RQ3) were identified through qualitative examination of item-generation errors, prompt deviations, and constraints observed during the productivity and test-quality analyses.

Results

Productivity

As shown in Figure 5, for manual / human created quizzes, total item-development time ranged from approximately 11–14 minutes per 10-item quiz. For the instructor, item construction and Google Forms entry occurred largely simultaneously, as sentence stems, correct answers, and distractors were composed and typed directly into the Google Forms. Approximately 8–10 minutes was spent on item construction (the cognitive task of formulating sentence stems and selecting correct answers / distractors) and 3–4 minutes was given to manual entry into Google Forms, including typing fill-in-the-blank sentences and entering multiple-choice options. It should be noted that the test creation times recorded were an average of the author's own and variations may occur depending on the instructor.

Figure 5

Table of Manual vs. LLM-assisted Vocabulary Quiz Creation

Stage	Manual / Human	LLM-assisted
Item construction	8–10 min	~0 (LLM does it)
Prompt submission + model response	-	~0.5 min
Post-editing	-	2–4 min
Entering into Google Forms	3–4 min	2–4 min (copy-paste)
Total preparation time per quiz	11–14 min	5-8 min

Note. For manually created quizzes, item construction and Google Forms entry were performed concurrently by the instructor; time estimates are presented as conceptual components rather than strictly sequential stages.

For LLM-generated quizzes, total per-quiz preparation time after prompt development ranged

from approximately 5–8 minutes. Prompt submission and model response time using GPT-4o averaged approximately 20–40 seconds per quiz, during which the system generated the full set of ten questions and answer options. On average, approximately 2–4 minutes per quiz were spent on post-editing to correct prompt deviations (instances where the generated output did not fully align with the intended format or constraints), refine distractors (to improve plausibility and alignment with the target vocabulary), or adjust phrasing (to enhance clarity and appropriateness for the learner level), while the remaining time was used for Google Forms entry. As with the manual condition, all items still had to be transferred into Google Forms; however, this step was faster due to copy-and-paste insertion of complete question sets generated by the LLM.

A comparative analysis of the two workflows revealed that LLM-assisted quiz construction yielded a 45–50% reduction in total preparation time relative to manual methods. This efficiency gain was primarily driven by the automation of initial item drafting and distractor suggestion.

Test-Item Quality

Student Performance Outcomes

The score distributions for both conditions showed similar results (Figure 6). Both sets of quizzes exhibited scores on the upper end of the scale, with the majority of students achieving high marks, clustering between 8 and 10.

Figure 6

Item Difficulty and Quiz Results

Condition	Quiz Range	Total Student Responses (N)	Mean Score / Facility (p)	Standard Deviation (SD)
Manually Created	Quizzes 1–6	211	0.825	1.454
LLM-Generated	Quizzes 7–11	146	0.830	1.507

The transition from manually authored assessments (Quizzes 1–6) to LLM-generated quizzes (Quizzes 7–11) resulted in fairly stable outcomes across the dataset. Statistics for the human-authored quizzes (N = 211) revealed a mean score facility (p) of 0.825 with a standard deviation (SD) of 1.454, indicating a generally competent performance of vocabulary mastery among students.

Following the implementation of the LLM-assisted workflow, the LLM-generated quizzes (N = 146) demonstrated nearly identical performance characteristics, with a mean facility of 0.830 and an SD of 1.507. Despite a different number of total responses, score distributions remained

consistent across conditions, with results clustering toward the upper end of the scale.

Distractor Analysis

Distractor quality was evaluated using a standard applied criterion. A distractor was classified as *functioning* if it attracted at least 5% of student responses. The metrics in Figure 7 summarize distractor efficiency across manually created quizzes (Quizzes 1–6) and LLM-generated quizzes (Quizzes 7–11).

Figure 7

Distractor Efficiency Comparison

Metric	Quizzes 1–6 (Human)	Quizzes 7–11 (LLM)
Mean Functioning Distractors per Item	1.46	1.64
Items with 3 Functioning Distractors	7 / 60 (11.6%)	10 / 50 (20.0%)
Total Non-Functioning Distractors (NFDs)	91	68
Mean NFDs per Quiz (10 items)	15.17	13.60

Although the mean number of functioning distractors per item increased from 1.46 to 1.64 in the LLM-generated quizzes, this represents a modest change in absolute terms (approximately 1–2 additional functioning distractors per 10-item quiz). Importantly, this increase did not correspond to a meaningful change in overall item difficulty, as reflected in the stable facility values across conditions. Across all items, manually created quizzes contained a higher total number of non-functioning distractors (NFDs; n = 91) than LLM-generated quizzes (n = 68), indicating fewer unused distractor options in the LLM condition. In proportional terms, non-functioning distractors accounted for 50.6% (91/180) of distractors in the human-authored quizzes and 45.3% (68/150) in the LLM-generated quizzes, representing a modest reduction of approximately five percentage points in the LLM condition.

While both conditions contained a substantial number of distractors that attracted zero selections, the LLM-generated quizzes showed better distractor efficiency, with a higher mean number of functioning distractors per item (1.64 vs. 1.46) and a greater proportion of items where all three distractors functioned (20.0% vs. 11.9%). In practical terms, the LLM condition reduced the average number of non-functioning distractors per quiz (13.6 vs. 15.2), suggesting more consistent plausibility across the full set of items.

Discussion

Stage 1

RQ1: How can vocabulary profilers be used to identify vocabulary that should be prioritized in an IELTS preparation course?

The authors have described how LEXTUTOR (Cobb), a freely accessible vocabulary profiler, can be a valid tool for homing in on vocabulary for instruction when creating target vocabulary lists for an IELTS preparation course when used in tandem with spreadsheet software such as Microsoft Excel. For this case study, LEXTUTOR made it possible for the teacher to not only understand the word difficulty level of reading passages objectively but also identify high frequency words that appear multiple times over the course of a semester. Klinger (2024) suggests students should be exposed to words that are worthwhile to learn more than three times to be learned effectively. This was incorporated as one of the criteria when screening the textbook reading passage corpus down to 1.7% of its entirety.

The IELTS is a language proficiency test which does not focus on a specific content topic. This makes learning high frequency words in the English language an obvious necessity to achieve a high score on the test. Especially for students with an English proficiency of CEFR A2-B1. For students with a higher proficiency, including words that are identified as off list of the NGSL and NAWL is also possible.

Stage 2

RQ2: How effective are large language models (LLMs) in vocabulary assessment preparation work in terms of productivity gains and test quality output?

Productivity

The use of large language models (LLMs) to generate vocabulary quizzes demonstrated clear productivity gains in the assessment preparation process. A time-on-task comparison showed that LLM assistance substantially reduced human item-construction workload, particularly in the generation of sentence stems and distractor sets. While manual quiz creation required instructors to design, type, and format each item individually, LLM-assisted workflows shifted much of this burden to automated content generation followed by human review. Although post-editing was still necessary, particularly in relation to distractor refinement and adherence to prompt constraints, these adjustments functioned primarily as targeted calibration rather than substantive correction.

Although initial prompt engineering required a one-time investment of time, this cost was amortized across subsequent quizzes and did not affect per-quiz efficiency. As model response accuracy and prompt adherence continue to improve, it is likely that post-editing time will decrease

further, leading to additional efficiency gains in AI-assisted assessment development.

These time savings suggest that LLMs may allow instructors to reallocate effort away from clerical tasks, though how this reclaimed resource is ultimately used remains dependent on instructor practice.

Test Quality Output

Productivity represents only one component of effective assessment design; efficiency gains are meaningful only if resulting quizzes maintain appropriate difficulty levels and avoid introducing unintended complexity through distractor design.

The quantitative analysis of Quizzes 1–10 indicates that the transition from manual item generation to LLM-assisted generation resulted in no meaningful change in student performance. Mean item facility remained nearly identical across conditions ($p \approx .81$), and standard deviation values showed comparable score distributions. These findings suggest that LLM-assisted quiz construction replicated the intended difficulty level of the instructor-authored assessments without increasing score variability.

The data from Quizzes 1–6 (human-authored) illustrate challenges in distractor performance across items. The human-authored quizzes (Quizzes 1–6) contained a higher number of non-functioning distractors, reflecting less consistent distractor performance across items. In some cases, non-functioning distractors appeared thematically or contextually misaligned with the semantic constraints of the stem. For example, in the item "The book had a huge _____ on his life," options such as *insight*, *secret*, and *left* were grammatically plausible but failed to satisfy the collocational and causal meaning required by the sentence, leaving *impact* as the only semantically appropriate choice. While human intuition can produce effective items, the LLM-assisted condition demonstrated slightly more consistent distractor functioning across the assessment set, as reflected in the lower number of non-functioning distractors.

Taken together, these results indicate that the reduction in preparation time did not come at the cost of assessment quality. The LLM-assisted workflow maintained overall difficulty and performance characteristics while improving distractor consistency. Beyond the present dataset, the scalability of LLM-assisted workflows warrants consideration. While the current study involved approximately 180 target vocabulary items, instructional contexts involving substantially larger lexical sets would require proportionally greater manual item-construction effort. Under such conditions, the relative efficiency advantage of LLM-based generation may become more pronounced, particularly in large-scale or cumulative vocabulary programs.

While prior research has documented substantial item-writing flaws in some LLM-generated assessments, the present findings suggest that outcomes may be context-dependent. When compared with manually authored quizzes in this study, the LLM-generated items demonstrated

slightly greater consistency in distractor functioning, including a modest proportional reduction in non-functioning distractors. Moreover, many distractors classified as non-functioning were grammatically plausible yet unselected by students, indicating that non-functioning status may reflect cohort response behavior and overall performance level rather than inherent irrelevance alone.

RQ3: What are the limitations of using lexical profilers and LLMs for creating vocabulary quizzes?

Limitations During Stage 1

The limitations observed in this stage can be grouped into two broad categories: word misrecognition by software and lack of candidate words fully meeting criteria. Word misrecognition by the different software used added procedures to the workflow that were not originally anticipated. The phenomenon was a good reminder to not put too much trust in technology and to manually double check if the output makes sense. Lack of candidate words that fulfill the set criteria for creating 12 wordlists of 15 words each from a corpus of 10,340 can be a challenge for any teacher. Details of the limitations during phase 1 and how they were coped with are described.

Word Misrecognition

Word misrecognition is a phenomenon that can occur when using OCR software to recognize the text on documents that were digitized with a scanner. It can also occur when copying and pasting texts from one format to another. For example, copying texts from a PDF to Microsoft Word or other online platforms. In the case of this study, word misrecognition occurred when; copying text from a PDF to LEXTUTOR, copying text to Microsoft Word, and using OCR on a scanned document.

In the case of copying the reading passages from a PDF to LEXTUTOR directly, unnecessary spaces appeared in the middle of words such as “mention”. This created an unusually high count of the word “ion” in the LEXTUTOR output, which would be very unlikely since the 12 passages covered a variety of topics including history, psychology, and engineering. As a method to check if the error was due to directly copying and pasting texts from a PDF to the webpage and checking for any other type of mistakes in the text, the reading passages were copied to a Microsoft Word document for checking.

When the documents were transferred to Microsoft Word, four types of errors were identified: letters morphing into Chinese characters, letters missing, addition of unnecessary spaces mid-word, and letter shuffling. On 6 out of the 12 passages four sets of Chinese characters were found. The letter combination “fi” was represented as 昀椀 94 times, “fl” as 昀氀 7 times, and “ff” as 昀昀 4 times. On four passages some letters of words were not recognized leading to words with missing

letters. The letter combination “Th” was missing from the words “this” and “the” 3 times, “tt” in “cutting” once, “fl” in words such as “flow” 9 times, and hyphens in hyphenated words such as “short-distance” twice. Unnecessary spaces mid-word such as “plat form” were located 15 times in two passages. In one passage the letters were scrambled for 345 words out of 945 making it impossible to read. For this unreadable passage, the textbook pages containing this reading were scanned.

A CZUR Shine digital scanner was used to digitize the passage from the physical textbook. Unfortunately for reasons unknown the OCR software that is included in the scanner driver could not convert the scanned documents into a searchable PDF or a Microsoft Word document successfully. The Microsoft Word document version of the problematic passage was cleaned using ChatGPT and manually compared against the original passage in the textbook to assure there were no mistakes then run through LEXTUTOR.

Lack of Candidate Words Fully Meeting Criteria

Due to the nature of the shortlisted words, lists that were scheduled to be used later in the semester did not have enough candidate words to be placed on the final vocabulary list of 15-words. For these lists, a second criterion was used to decide which words to add. Words that appear in less than two passages but appear multiple times in a single passage were considered to be added to the list for that passage. Frequency of appearance in a single passage was determined by running all words from each passage separately through LEXTUTOR (Cobb) a second time. When there were not enough words to complete the lists of 15 using the two previously mentioned criteria, an executive decision was made to include words that made an appearance in only one passage and belonged in bands 2, 3, or A in the LEXTUTOR vocabulary profile output. A total of 8 words fit this description.

In relation to the nature of shortlisted words, not all vocabulary lists were able to fully follow the word difficulty standard of 9 words from band 2, 3 words from band 3, and 3 words from band A of the LEXTUTOR output.

Limitations During Stage 2

LLM-Generated Output

The limitations observed primarily concerned inconsistencies in model-generated output. While LLMs substantially accelerated quiz creation, adherence to prompt constraints was not always reliable and required human review. Even when instructed to use IELTS Band 6–7-level English, some generated items exhibited phrasing that was overly formal or textbook-like rather than reflective of authentic usage. In several instances, the model failed to follow explicit prompt

constraints, such as the requirement that all distractors be drawn from a provided vocabulary list; for example, items occasionally used distractors that also appeared in the pool of eligible correct answers (e.g., *trend* or *artifact*), necessitating post-generation editing.

Despite improved distractor functioning overall, some challenges were observed in distractor construction within the LLM-generated items. In some instances, although distractors were intended to be plausible but incorrect, some options overlapped too closely with the correct answer, creating ambiguity, while others were clearly implausible or unrelated, reducing item quality and discriminatory power. These inconsistencies underscore the need for post-editing and instructor judgment when deploying LLM-generated assessment materials.

These limitations align with findings reported in prior research on LLM-generated assessment items. Camarata et al. (2025) reported that nearly half of ChatGPT-generated medical multiple-choice items contained item-writing flaws, with over one-fifth exhibiting factual or conceptual errors. Similarly, Bitew et al. (2023) found that 25% of stems and 33% of distractors in LLM-generated vocabulary items were inappropriate, while Putri et al. (2025) reported that 41.27% of generated distractors were irrelevant.

Sample Size and Analytic Approach

Several limitations should be considered when interpreting the findings of stage 2. First, class sizes were relatively small, with between 16 and 24 responses per quiz, which may limit the generalizability of the results to larger or more diverse learner populations. In addition, counts of non-functioning distractors (NFDs) are sensitive to sample size, as distractors may remain unselected due to chance rather than poor design, particularly in high-performing cohorts.

An additional limitation concerns the operationalization of item difficulty in the present study. Difficulty was inferred primarily from student response behavior and distractor performance, without accounting for other dimensions that may contribute to perceived item difficulty, such as stem length, syntactic complexity, or overall reading load. As a result, the analysis reflects difficulty at the level of response options rather than the full cognitive demands of each item.

The analysis relied primarily on descriptive statistics rather than inferential testing, reflecting the process-based and exploratory nature of the study. While this approach was appropriate given the bounded timing estimates and repeated-measures structure of the data, it limits the strength of statistical claims that can be made regarding observed differences between conditions.

Instructional Context and Task Scope

The study was conducted within a single instructional context and involved quizzes created and reviewed by one instructor, which may constrain the applicability of the findings to other teaching settings or levels of instructor expertise. In the LLM-assisted condition, all generated quiz

items were post-edited by the instructor, meaning that the observed outcomes reflect a human–LLM collaborative workflow rather than fully automated item generation.

Furthermore, the 5% threshold used to classify functioning distractors represents a discretionary analytic decision rather than a fixed standard, and alternative thresholds could produce slightly different classifications. Finally, the analysis focused exclusively on vocabulary quizzes, and the findings may not extend to other assessment formats or language skill domains.

Conclusion

This preliminary study demonstrates that vocabulary profilers and LLMs can be used in combination to support IELTS test preparation. The use of lexical profilers such as LEXTUTOR with NGSL/NAWL ensured level-appropriate vocabulary selection, while LLM assistance substantially reduced the time required for quiz construction. Together, this approach offers an efficient work flow to vocabulary assessment design within a test-preparation context.

The observed limitations of LLM-generated items did not translate into reduced assessment quality when compared with manually authored quizzes. Item-analysis metrics indicated comparable difficulty levels and slightly improved distractor functioning in the LLM-assisted condition. Within this dataset, model-generated quizzes were at least as stable as instructor-created items.

While prior studies have documented substantial rates of item-writing flaws in LLM-generated assessments, the present findings underscore the importance of contextual variables, including prompt structure and analytic thresholds such as the 5% criterion used to classify functioning distractors. In this study, many distractors deemed non-functioning were grammatically plausible yet unselected, suggesting that error rates in LLM-assisted assessment design should be interpreted with methodological caution rather than treated as inherent instability. A pragmatic approach rather than categorical acceptance or rejection may therefore be more productive for instructional contexts.

Finally, the findings help clarify the pedagogical role of LLMs in assessment design: not as replacements for instructor expertise, but as structured generative partners whose value lies in accelerating draft production while preserving human supervision. The results therefore suggest that LLMs operate most effectively as collaborative tools rather than fully autonomous assessment generators, preserving the need for human editorial oversight.

Implications and Future Research

Future research on curated vocabulary lists should examine their integration with additional language skills, particularly writing and speaking, as well as their impact across longer instructional periods and more diverse learner populations. Expanding the scope with these skills would allow

for a more comprehensive evaluation of how vocabulary-focused interventions contribute to broader language development within IELTS-oriented curricula. In addition, learner surveys and reflective instruments could be employed to investigate student perceptions of curated vocabulary lists, including their perceived usefulness in developing awareness of meaning, usage, and paraphrasing, and whether such knowledge transfers to authentic IELTS reading and listening tasks.

Unlike prior studies that systematically coded item-writing flaws or calculated formal error rates in LLM-generated assessments, the present study focused primarily on performance-based indicators such as facility values and distractor functioning. Future investigations could therefore incorporate structured coding of item-writing errors and formal error-rate analysis alongside multidimensional measures of item difficulty, including stem length and overall reading load, in order to better capture the full cognitive demands of assessment items.

LLMs may also be explored in the post-assessment processes, such as generating individualized feedback, identifying common error patterns, and tailoring follow-up exercises based on student responses. Lastly, future studies should also examine educator receptivity to LLM generated assessments and post-assessment processes, as the pedagogical value of these tools ultimately depends on how teachers and administrators engage, interact, and form policy with these technologies available.

References

Bitew, S. K., Deleu, J., Develder, C., & Demeester, T. (2023, July 30). *Distractor generation for multiple-choice questions with predictive prompting and large language models*. arXiv. https://doi.org/10.48550/arXiv.2307.16338

Browne, C. (2013), The New General Service List: Celebrating 60 years of vocabulary learning, *The Language Teacher*, 37:4, 13-15.

Browne, C. (2014). A new general service list: The better mousetrap we've been looking for?. *Vocabulary learning and Instruction*, 3(2), 1-10.

Camarata, T., McCoy, L., Rosenberg, R. L., Temprine Grellinger, K. R., Brettschneider, K., & Berman, J. (2025). LLM-generated multiple choice practice quizzes for pre-clinical medical students: Prevalence of item writing flaws. *Advances in Physiology Education*. https://doi.org/10.1152/advan.00106.2024

Chen, Q. (2019). Phrasal paraphrase learning: The role of memory, analogy, and input frequency. *English Language Teaching*, *12*(5), 68–80. https://files.eric.ed.gov/fulltext/EJ1215106.pdf

Cobb, T. *Complete Web-VP v.2.6* [computer program]. Accessed 11 Jan 2025 at https://www.lextutor.ca/vp/comp/.

Coxhead, A. (2000). A new academic word list. *TESOL Quarterly, 34(2)*, 213-238.

Gottlieb, M. (2016). *Assessing English language learners: Bridges to educational equity: Connecting academic language proficiency to student achievement* (2nd ed.). Thousand Oaks, CA: Corwin.

Hendry, C., & Sheepy, E. (2018). How much vocabulary is needed for comprehension of research publications in education?. *Future-proof CALL: language learning as exploration and encounters*, 94.

Klinger, R. (2024). Vocabulary frequency and dispersion in Japanese junior high school EFL textbooks. *Vocabulary Learning and Instruction*, *13*(2), 1-18.

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., & Goldstein, Z. (2004). Testing vocabulary knowledge: Size, strength, and computer adaptiveness. *Language learning*, *54(3)*, 399-436.

Nakayama, S. (2022a). A close examination of vocabulary in Japanese EFL textbooks. *Reflections and New Perspectives*, 209-216.

Nakayama, S. (2022b). Vocabulary in Japanese EFL textbooks: A bidirectional coverage analysis. In *The IAFOR International Conference on Education–Hawaii 2022 Official Conference Proceedings* (pp. 157-169).

Nation, P., & Beglar, D. (2007). A vocabulary size test. *The Language Teacher*, *31(7)*, 9-13.

Putri, A., Fauziati, S., & Rizqi, A. A. (2025). Syntagmatic distractor generation for multiple-choice language tests: A large language model-based approach. *Proceedings of the 4th International Conference on Electronics Representation and Algorithm (ICERA)*. https://doi.org/10.1109/ICERA66156.2025.11087340

Roediger, H. L., III, & Butler, A. C. (2011). The critical role of retrieval practice in long-term retention. *Trends in Cognitive Sciences*, *15*(1), 20–27. https://doi.org/10.1016/j.tics.2010.09.003

Rott, S. (1999). The effect of exposure frequency on intermediate language learners' incidental vocabulary acquisition through reading. *Studies in Second Language Acquisition 21*, 589-619.

Sasaki, M. (2022). Skill profiles of Japanese EFL learners: Evidence from large-scale assessments. *Language Testing in Asia*, *12*(1), 1–23. https://doi.org/10.1186/s40468-022-00203-3

Stewart, J. (2024). Establishing meaning recall and meaning recognition vocabulary knowledge as distinct psychometric constructs in relation to reading proficiency. *Language Testing*. https://doi.org/10.1177/02655322231162853

Van Zeeland, H., & Schmitt, N. (2013). Lexical coverage in L1 and L2 listening comprehension: The same or different from reading comprehension?. *Applied linguistics*, *34*(4), 457-479.

Waring, R. and Nation, I.S.P. (1997) Vocabulary size, text coverage, and word lists. In *Vocabulary: Description, Acquisition and Pedagogy* N. Schmitt and M. McCarthy (eds.). Cambridge University Press, Cambridge: 6-19.

Waring, R., & Takaki, M. (2003). At what rate do learners learn and retain new vocabulary from reading a graded reader? *Reading in a Foreign Language*,*15(2)*, 130-163.

West, M. (1953). A general service list of English words. London: Longman, Green & Co.

Xi, X. (2010). How do we go about investigating test fairness? L*anguage Testing*, *27*(2), 147–170. https://doi.org/10.1177/0265532209349465

Yan, L., Sha, L., Zhao, L., Li, Y., Martinez-Maldonado, R., Chen, G., Li, X., Jin, Y., & Gašević, D. (2024). *Practical and ethical challenges of large language models in education: A systematic scoping review*. arXiv. https://arxiv.org/abs/2303.13379