Studies in Second Language Acquisition (2026), **48**, 757–783
doi:10.1017/S0272263126101818

RESEARCH ARTICLE

The effects of contextual diversity on incidental vocabulary learning

Gaia Oikawa and Takumi Uchihara

Graduate School of International Cultural Studies, Tohoku University, Sendai, Japan
Corresponding author: Gaia Oikawa; Email: oikawa.gaia.r5@dc.tohoku.ac.jp

(Received 12 June 2025; Revised 20 April 2026; Accepted 08 May 2026)

Abstract

This study investigated the effects of contextual diversity (CD) on second language incidental vocabulary learning. A total of 124 Japanese learners of English were allocated to a control group or 2 experimental groups, either a high contextual diversity (HCD) or a low contextual diversity (LCD) group. Participants in the HCD group encountered target words across three different texts that varied in genre and topic, while those in the LCD group read three different texts that shared the same genre and topic. Meaning recall and recognition tests were conducted at pretest, immediate posttest, and delayed posttest. Results showed that HCD outperformed LCD on meaning recognition at the delayed posttest. Moreover, learners with greater prior vocabulary knowledge tended to benefit more from contextually varied input, whereas such input may have adverse effects on learners with lower lexical proficiency. This study offers insights into the role of CD in incidental vocabulary acquisition and provides pedagogical implications for optimally incorporating input variability into L2 vocabulary instruction.

Keywords: contextual diversity; incidental vocabulary learning; input variability; reading; vocabulary knowledge

Introduction

Exposure to input through reading is a key source of second language (L2) vocabulary learning. Learners gradually acquire different aspects of word knowledge through repeated encounters with L2 words in context. However, vocabulary acquisition is not solely determined by the quantity of input; input quality also plays an equally critical role (Webb & Nation, 2017). In particular, repeated exposure to and use of words across varied contexts enhances input quality and promotes the enrichment and consolidation of lexical knowledge (Chandy et al., 2025; Joe, 1998; Webb & Nation, 2017).

Although input variability can take various forms, such as phonetic variation (e.g., talker variability; Uchihara et al., 2022) and morphological variation (e.g., derivations

© The Author(s), 2026. Published by Cambridge University Press. This is an Open Access article, distributed under the terms of the Creative Commons Attribution licence (http://creativecommons.org/licenses/by/4.0), which permits unrestricted re-use, distribution and reproduction, provided the original article is properly cited.

https://doi.org/10.1017/S0272263126101818 Published online by Cambridge University Press

and inflections; Reynolds, 2015), this study focuses on contextual diversity (CD), defined as the number of distinct contexts represented by different genres and topics in which a word appears (Adelman et al., 2006; Hulme et al., 2023; Rosa et al., 2022). The CD is considered an important source of variability as it is reported to facilitate the formation of form-meaning mappings in both first language (L1) (Johns et al., 2016; Mak et al., 2021; Rosa et al., 2017, 2022) and L2 contexts (Frances et al., 2020; Nguyen, 2023). Contextual variability is thought to enhance learners' inferencing and memory by activating broader background knowledge and promoting deeper semantic processing (Pagán & Nation, 2019; Pulido, 2007; Yamashita, 2022).

Despite the potential benefits of CD, much of the supporting evidence comes from studies on L1 children learning novel words. In L2 contexts, empirical research is still limited. Despite a few studies examining the CD effects on L2 vocabulary learning (e.g., Chandy et al., 2025), their operationalization of the contextual variability (i.e., a direct comparison between repeated exposures to the same sentence and different sentences) did not necessarily mirror authentic reading experiences learners have with texts from diverse genres and topics. Thus, it remains unclear whether, and to what extent, contextually varied input supports incidental vocabulary learning in more ecologically valid reading conditions.

To address this gap, the present study investigates the CD effects on incidental vocabulary learning among Japanese learners of English as a foreign language (EFL). Specifically, it examines whether reading multiple texts of varied genres and topics facilitates the incidental learning of L2 words. This research sheds light on how encountering new words in varied contextual discourses shapes L2 incidental vocabulary learning and offers pedagogical implications for optimizing vocabulary instruction through reading.

Literature review

Incidental vocabulary learning from reading

Given the conceptual ambiguity of incidental vocabulary learning from a psychological perspective (Webb, 2024), this study adopts a methodological definition: Vocabulary learning that occurs when learners are not explicitly aware of upcoming vocabulary tests and engage in meaning-focused reading tasks (Durrant et al., 2022; Hulstijn, 2003; Uchihara et al., 2019). Numerous studies have demonstrated the occurrence of incidental vocabulary learning in both L1 (Jenkins et al., 1984) and L2 (Horst et al., 1998; Pellicer-Sánchez et al., 2024; Teng, 2024a). A meta-analysis by Webb et al. (2023) reports that L2 learners acquire approximately 17% of the target words immediately after reading and retain 15% of the learned words. Although the number of encounters is positively correlated with vocabulary learning (r = .41; Uchihara et al., 2019), repetition alone accounts for only a modest portion of the variance (17%). Moreover, studies have reported inconsistent findings on the optimal number of repetitions, ranging from 6 (Rott, 1999) to 20 (Waring & Takaki, 2001). These findings suggest that incidental learning gains likely vary as a function of multiple variables beyond repetition alone. For example, Webb et al. (2023) indicated that moderating variables include a range of learner profiles (e.g., L2 proficiency and institutional level), material and activity features (e.g., text type, text audience, spacing, and mode of input), and methodological features (e.g., prior knowledge of target words and test format).

Incidental vocabulary learning from reading entails a dynamic interplay between bottom-up and top-down cognitive processes involved in text comprehension (Koda,

https://doi.org/10.1017/S0272263126101818 Published online by Cambridge University Press

2018; Smith et al., 2021; Yamashita, 2022). Through bottom-up processing, L2 readers construct a textbase by analyzing explicitly stated linguistic information. Top-down processing, in turn, enables them to integrate background knowledge into the product of bottom-up processing (i.e., textbase) to understand the content. During such complicated processing executed for reading comprehension, word learning can occur only on the condition that readers encountering unfamiliar words (a) recognize a gap in their lexical knowledge, (b) infer meaning from context, and (c) associate the L2 form with the inferred meaning and relevant contextual cues (de Bot et al., 1997). When any of the steps fail, incidental vocabulary learning is less likely to occur.

Contextual diversity and vocabulary learning

Defining CD remains an open question. One simple way of its operationalization is to manipulate the repetition of text input by comparing verbatim input (e.g., reading the same texts multiple times) with varied input (e.g., reading different texts in which the same target words appear multiple times) (Chandy et al., 2025; Durrant & Schmitt, 2010; Li et al., 2023; Liu & Todd, 2016). Liu and Todd (2016) found that the varied-input group reading seven different texts achieved significantly higher scores on the meaning recognition test than the verbatim group reading the same text seven times (but for mixed findings, for example, see Durrant & Schmitt, 2010). Using the eye-tracking technology, Chandy et al. (2025) demonstrated that contextually diverse contexts elicited more attention, indexed by slower total fixation duration and longer initial fixations, to unknown words while reading. Participants who read different sentences showed significantly longer processing time than those who read the same sentence repeatedly, indicating that reading contextually diverse texts may lead to deeper processing and stronger memory (for similar results with L1 learners, see Pagán & Nation, 2019). Although this line of research sheds light on the role of CD in vocabulary learning, it is important to note that the definition of CD adopted by earlier studies does not accurately reflect the degree of actual variability that learners encounter in real-life L2 input (e.g., extensive reading, Webb & Chang, 2015b).

The traditional approach (i.e., contrasting verbatim with varied conditions, as reviewed above) can be interpreted as the operationalization of CD at the linguistic or *local* level (e.g., target words are embedded in various contexts surrounded by different words). However, in the real-life reading environment, learners could repeatedly encounter words in multiple passages that are contextually diverse at the discursive, or *global* level. This global account of CD concerns variability at the discourse level, such as genres and topics. The two perspectives of CD align with the concepts of *verbal* and *nonverbal* contexts proposed by Engelbart and Theuerkauf (1999). Verbal context refers to grammar and semantics, whereas nonverbal contexts refer to situations, descriptions, subjects, and global clues such as general knowledge that are relevant to learning topics. The framework highlights the importance of considering not only linguistic variability occurring from lexico-grammatical features of a text but also non-linguistic variability at the global level (i.e., genres and topics).

Genre and topic are similar but refer to different concepts (Biber & Conrad, 2009; Lee, 2001). Genre, typically conceptualized as the purpose of texts with different structural styles of texts, has been operationalized in L2 reading studies as a binary distinction between narrative and expository texts, while topic refers to the thematic content of a text (e.g., sports, health, or technology), rather than its structural style. In other words, genres define how information is presented, whereas topics denote what the text is about. Although the definition of genre remains debated, there are two major

https://doi.org/10.1017/S0272263126101818 Published online by Cambridge University Press

approaches to determining genre differences. One is a linguistic approach, which relies on the quantitative analysis of linguistic features such as co-occurrences of words (Biber, 1992; Biber & Conrad, 2009); the other is a rhetoric approach, which emphasizes readers' perceptual interpretation and the communicative purpose achieved by the text (Hyland, 2002, 2021; Lee, 2001; Suchomel, 2021). The former allows us to determine text genre more objectively, whereas the latter approach may better reflect the psychological reality underlying the perception of genre differences.

Regarding the topic, we based our classification on *general topical domains* rather than *specific topics* (Biber & Conrad, 2009). General topical domains refer to broader categories of topics such as science, society, religion, and sports, whereas specific topics are indicative of individual, more narrowly defined subjects or storylines within those domains. Since Graesser et al. (2011) demonstrated that indices of Coh-Metrix, which can help automatically analyze text characteristics, differed significantly between science and society topics, even within the same expository genre, we chose these two topical domains.

Previous research showed that genres and topics yield differences in text characteristics and affect vocabulary learning. Narrative texts, characterized by temporally sequenced events and human agents, tend to facilitate comprehension due to their familiar discourse structure and cohesive elements (Mar et al., 2021). Supporting this view, Webb et al.'s (2023) meta-analysis suggests that narrative texts yield greater vocabulary gains ($g = 1.43$) than expository texts ($g = 0.61$). However, most studies in this domain have predominantly employed narrative texts, and few have directly compared different genres, limiting our understanding of how genre-specific features affect incidental vocabulary learning (Arai & Takizawa, 2024). Moreover, previous comparisons have tended to dichotomize genres (narrative vs. expository) and inevitably overlooked variation within each genre that differs in thematic content and communicative style (e.g., topic differences). The influence of topic variations also needs to be considered as a potential variable contributing to L2 vocabulary learning (Cancino, 2023; Pulido, 2003, 2007). When topic familiarity is higher, learners can leverage background knowledge to compensate for their lack of understanding, improve their overall comprehension of the text, direct greater attentional resources for encoding, and eventually learn unknown words efficiently while reading.

In the literature of incidental vocabulary learning (Uchihara et al., 2019; Webb et al., 2023), linguistic features (e.g., vocabulary) are often controlled, for example, by ensuring that lexical coverage is appropriate for the target population. However, comparable attention has not been paid to non-linguistic variables (e.g., background knowledge), and it remains underexplored whether and how variability at the non-linguistic level enhances (or hinders) incidental vocabulary learning from reading. The present study operationalizes CD at a global level with variations in both genre and topic across different texts. Unlike previous studies that manipulated CD by varying linguistic (or local) contexts within a single topic (e.g., Durrant & Schmitt, 2010; Liu & Todd, 2016), we adopt a discourse-level approach (i.e., global CD), wherein target words appear in texts with different genres and topics. This allows us to examine the influence of CD while accounting for potential effects of non-linguistic contextual variability on L2 vocabulary learning.

Theoretical basis and empirical support for contextual diversity

Theoretical basis. Reading contextually diverse texts is theorized to facilitate the encoding of both linguistic and non-linguistic cues associated with newly encountered

https://doi.org/10.1017/S0272263126101818 Published online by Cambridge University Press

words. When reading, learners must retrieve lexical meanings while simultaneously activating relevant background knowledge (Yamashita, 2022). This dual activation of linguistic and non-linguistic memory engages deeper processing, integrating personal experience and world knowledge, which in turn strengthens memory consolidation for L2 vocabulary (Craik & Lockhart, 1972). Eye-tracking studies have shown that CD induced greater attention to unknown items (i.e., deeper processing). Chandy et al. (2025) demonstrated that contextually diverse texts induced slower total fixation durations and longer initial fixations compared to non-diverse texts, indicating that reading contextually diverse texts involves deeper cognitive engagement. Similarly, in L1 reading research, Pagán and Nation (2019) reported longer reading times and fixations during exposure to novel words in varied contexts, along with faster processing during subsequent reading. These findings suggest that contextually diverse input may increase attentional engagement, thereby promoting stronger memory formation.

Furthermore, Bolger et al.'s instance-based memory model of word learning (Bolger et al., 2008) accounts for the benefit of CD for word learning while considering the activation of wider background knowledge. This model suggests that when we learn words from reading, contextualized episodic memory of encounters with words incrementally accumulates, and the summation of memories shapes word meaning. In addition, higher CD (i.e., exposure to a wider range of different contexts) is expected to create richer memory traces that can resonate during subsequent encounters, thereby facilitating the learning of word meanings.

Empirical support. The theoretical benefits of CD have received partial support from existing studies, predominantly in L1 settings (Hulme et al., 2023; Johns et al., 2016; Joseph & Nation, 2018; Mak et al., 2021; Norman et al., 2023; Rosa et al., 2017, 2022). For instance, Rosa et al. (2017) found that Spanish-speaking children who read texts from three distinct genres (fables, science, and math) demonstrated significantly higher gains across four vocabulary measures compared to those who read the same genre. This pattern was replicated in a follow-up study with adolescents (Rosa et al., 2022). Similarly, Johns et al. (2016) found that L1 English university students recognized target words more quickly and accurately after reading passages drawn from varied topics (high contextual diversity [HCD]) versus a single topic (low contextual diversity [LCD]), as shown by lexical decision and semantic similarity tasks.

Conversely, some studies reported the opposite trend. Hulme et al. (2023) and Norman et al. (2023) found that learners in LCD conditions outperformed those in HCD conditions. One explanation is that narrower contexts, by maintaining consistent topics and vocabulary, may enhance familiarity and support inferencing, essentially simulating the conditions of narrow reading (Schmitt & Carter, 2000). In narrow reading, repeated exposure to similar content reduces lexical burden and facilitates a meaning-inference process through accumulated discourse knowledge. In contrast, HCD conditions may increase cognitive demand without the benefits of increasing content familiarity, especially for less proficient learners.

In the L2 domain, evidence for global CD effects remains limited. Frances et al. (2020) examined Spanish learners of English who encountered pseudowords across four levels of contextual variability. Vocabulary gains were significantly greater in the higher variability conditions across multiple outcome measures (form recall, recognition, picture matching). Nguyen (2023) similarly found that reading paired narrative and expository texts led to greater vocabulary gains than narrative-only input among Thai university students. These studies suggest that global CD, particularly across genres, can enhance L2 incidental vocabulary acquisition.

https://doi.org/10.1017/S0272263126101818 Published online by Cambridge University Press

However, methodological limitations in these studies should merit further exploration. First, in Frances et al. (2020), the amount of textual information available for each target word varied considerably because participants read different amounts of text for each target word. The authors manipulated CD by altering the number of texts in which each word appeared. Specifically, there were four conditions as a within-subject design: (a) 2 of the 8 target words occurred eight times within a single 100-word story (the lowest CD condition), (b) two words appeared four times in two texts, (c) two words appeared twice in two texts, and (d) two words appeared once in each of the eight different stories (the highest CD condition). Consequently, learners were exposed to only one text for two target words, but to eight different texts for the other two. Second, in Nguyen's (2023) study, while the length and number of texts were controlled, participants encountered each target item only once, where vocabulary learning might be severely underestimated. Also, the treatment involved text enhancement and output tasks to prompt text comprehension, in which learners' attention was likely directed to target words. These methodological variations, consequently, provide limited insights into how CD affects incidental vocabulary learning. The current study was designed to address these methodological limitations (i.e., issues with frequency of occurrence, lexical coverage, and text length) to determine the impact of CD on incidental vocabulary learning from reading.

Prior vocabulary knowledge and text comprehension

Among reader-related factors, prior vocabulary knowledge (vocabulary size) has received the most extensive attention in research on incidental vocabulary learning (see Uchihara et al., 2019, for a review). Learners with a larger vocabulary are better equipped to process textual information fluently and allocate cognitive resources to infer meanings and activate background knowledge, enabling them to build coherent text representations and learn new words more efficiently (Kintsch & Welsch, 1991; Pulido, 2007; Yamashita, 2022). In contrast, learners with limited vocabulary may struggle with surface-level processing (i.e., retrieving word meanings) (Webb & Chang, 2015a), leaving little attentional capacity for higher-level processing such as inference-making and vocabulary learning (Robinson, 2003). Consistent with this view, empirical studies have shown that learners with larger vocabulary knowledge, as measured by instruments like the Vocabulary Levels Test, tend to acquire more L2 single words (Teng, 2024b; Webb & Chang, 2015a) as well as multiword sequences (Vilkaitė, 2017; Vu & Peters, 2022, 2023) from reading than those with limited vocabulary knowledge. Scholars conducting corpus-based studies have attempted to establish the threshold of lexical coverage necessary for effective reading comprehension, with estimates ranging from 95% (Laufer, 1989) to 98% or higher (Nation, 2006; Pellicer-Sánchez et al., 2024). Zhang and Zhang (2022) conducted a meta-analysis and demonstrated that vocabulary knowledge significantly impacted L2 reading comprehension ($r = .57$, $p < .01$). Recent research suggests that the relationship between lexical coverage and reading comprehension tends to be linear and that there may not be a single threshold for adequate reading comprehension (Webb, 2021; Webb et al., 2025). The relationship may also vary depending on relevant factors such as topic familiarity (Schmitt et al., 2011) and genre (Kremmel et al., 2023). Although mixed on the cut-off points guaranteeing optimal comprehension, these findings indicate that higher lexical coverage is associated with greater comprehension and increased opportunities for word learning.

https://doi.org/10.1017/S0272263126101818 Published online by Cambridge University Press

The importance of prior vocabulary knowledge may be even more critical when learners are exposed to contextually diverse input. According to the Cognitive Load Theory (Sweller, 1988; Sweller et al., 2019), learners with limited vocabulary knowledge are hypothesized to face increased extraneous cognitive load when reading contextually diverse texts. With insufficient lexical support, these learners may struggle to process surface-level linguistic information and, as a result, have limited capacity to activate or apply background knowledge. In contrast, learners with larger vocabulary sizes can process texts more easily and direct more attention to unfamiliar items, making contextual inferencing more feasible. Also, in the HCD condition in particular, where readers were exposed to texts from different genres and topics, the required cognitive load can be expected to be higher. This reasoning suggests that CD benefits L2 vocabulary learning if learners possess adequate prior vocabulary knowledge (i.e., a desirable level, Bjork & Kroll, 2015; Suzuki et al., 2019). Accordingly, this study aims to reveal, if any, the moderating role of prior vocabulary knowledge (or vocabulary size) in incidental word learning under varying levels of CD.

The current study

As reviewed above, prior research has suggested that CD facilitates vocabulary learning. However, existing studies have focused predominantly on L1 learners with limited evidence available for L2 populations. Furthermore, several methodological limitations regarding the control of frequency of exposure to target items, lexical coverage, and use of output tasks in earlier L2 studies made it difficult to draw firm conclusions about the effects of CD on L2 incidental vocabulary acquisition. Addressing these gaps is essential for understanding how contextual variability, as a key dimension of input quality (Webb & Nation, 2017), influences word learning in L2 contexts.

To this end, the present study adopted a pretest–posttest–delayed posttest design to investigate the effects of CD on L2 incidental vocabulary learning with Japanese EFL learners. Participants in the two experimental groups—HCD and LCD—encountered 10 target words 6 times while reading 3 different texts in English. Following the global perspective of CD (Engelbart & Theuerkauf, 1999), we manipulated CD as a function of varying genres (narrative vs. expository) and topics (science vs. society). Specifically, the HCD group read three texts differing in both genre and topic (i.e., narrative, science-themed expository, and society-themed expository), while the LCD group read three texts representing a single genre and topic (e.g., three narratives, three science texts, or three society texts). Vocabulary learning was assessed using meaning recall and meaning recognition tests. To examine how prior vocabulary knowledge modulates the CD effects on vocabulary learning, we measured participants' vocabulary breadth using the updated Vocabulary Levels Test (uVLT; Webb et al., 2017). Accordingly, this study was guided by the following research questions:

1. To what extent does exposure to L2 words in texts with higher CD lead to greater lexical gains compared to in texts with lower CD?
2. To what extent does prior vocabulary knowledge modulate the effects of CD on vocabulary learning from reading?

Method

Reading materials, dataset, and analysis code are accessible through OSF (https://osf.io/5xzj6/overview).

https://doi.org/10.1017/S0272263126101818 Published online by Cambridge University Press

Participants

A total of 164 Japanese undergraduate students (aged 18–21) enrolled in 4 English classes participated in this study. Because participants were assigned to classes based on TOEFL scores obtained upon university entry, they were considered to be at pre-intermediate to intermediate levels. Forty participants were excluded from the final analysis due to one or more of the following reasons: (a) absence from treatment sessions, (b) substantial prior knowledge of the target words as indicated by pretest scores (i.e., exceeding 50% accuracy on either the meaning recall or recognition tests), or (c) self-reported engagement in the external learning of the target words outside the treatment sessions. Consequently, the data from 124 participants were retained for statistical analysis. Participants attended a 90-minute English class once a week, taught by the same instructor across all groups. One class was assigned to a test-only control group, while the remaining three classes formed experimental groups (see Procedure and Table 4).

Vocabulary tests

We administered meaning recall and meaning recognition tests before the treatment (pretest), immediately after the treatment (immediate posttest), and one week after the treatment (delayed posttest) to measure vocabulary gain and retention. The meaning recall test preceded the meaning recognition test to prevent learners from seeing the meanings of the target items (Nation & Webb, 2011). Since the learning activity in the present study primarily focused on meaning (i.e., learners attended to the semantic content of the reading materials), we concentrated on the retrieval of the word meaning cued by the L2 written form. These tests help capture the learning of form-meaning mapping with varying degrees of precision from partial to fuller knowledge (Durrant et al., 2022; Kremmel & Schmitt, 2016). In the meaning recall test, participants translated L2 target words into L1 equivalents as accurately as possible while encouraged to provide the word meanings represented in the reading materials (e.g., *baboon*: "animals living in forests and sometimes come down to cities"). In the meaning recognition test, participants selected the correct meaning of a target word from four options. Each question included one correct answer and three distractors with the same part of speech as the correct answer, and their meanings were not too far from the correct meaning. For example, for *encumber*, the three distractors were *threaten*, *remove*, and *observe*. Two intermediate- to advanced-level Japanese EFL learners who did not participate in this experiment reviewed the distractors to confirm that the distractors could not easily be eliminated based on mere differences in semantics and orthography from the correct answers. Other issues pointed out through the review process were rectified (e.g., two choices semantically indistinct were removed).

In addition, we employed the updated Vocabulary Levels Test (uVLT; Webb et al., 2017) in the first week of the procedure to measure prior vocabulary knowledge (see Table 1). According to the suggested threshold of mastery of each frequency level (i.e., 29/30 for 1K to 3K levels, 24/30 for 4K and 5K levels, Webb et al., 2017), 59.7% of the participants were considered to master the most frequent 1,000 word families, 26.6% of the participants mastered the most frequent 2,000 words or more, and 4.03% mastered the most frequent 3,000 words or more. Preliminary analysis confirmed no significant difference between the HCD and LCD (see Online Appendix S1).

https://doi.org/10.1017/S0272263126101818 Published online by Cambridge University Press

Table 1. Mean and standard deviation (in parentheses) of uVLT for five frequency levels

	1K	2K	3K	4K	5K
HCD (n = 50)	29.4 (0.93)	28.4 (1.37)	26.2 (2.38)	21.6 (3.35)	17.0 (4.40)
LCD (n = 42)	29.3 (1.00)	28.5 (1.61)	26.1 (2.18)	21.5 (2.64)	17.0 (5.05)
Control (n = 32)	29.3 (0.97)	27.8 (1.89)	24.6 (3.20)	19.0 (3.96)	13.8 (4.16)

Note. The theoretical maximum score per frequency level = 30.

Reading materials

We developed nine written texts representing two genres (Narrative and Expository) and two topics within expository texts (Science and Society), producing three discourse units of reading texts: (a) Narrative, (b) Expository (Science), and (c) Expository (Society). Participants in the HCD condition read one text each from (a), (b), and (c), whereas those in the LCD condition read three texts from one discourse unit (see Table 4). We selected nine texts from the British Council and EPFL news (https://news.epfl.ch/), with three for each discourse unit (see Table 2).

To determine the differences at the discourse levels (i.e., genre and topic variations), we first followed the New Rhetoric Approach to account for the readers' perceptual interpretation and the text's communicative purpose (Hyland, 2002, 2021; Lee, 2001; Suchomel, 2021). This emphasis was critical for confirming that readers could perceive genre distinctions, ensuring these perceptions were accurately reflected in the experimental materials. As for the topic, we based our classification on general topical domains (Biber & Conrad, 2009) with a particular focus on science and society topics (for the evidence showing linguistic differences, see Graesser et al., 2011). We subdivided expository texts into two topical categories: Science, which included texts

Table 2. Information on nine texts.

Title	Genre and topic	Length	Lexical coverage with the most frequent 2,000 words	Type/token ratio	Comprehension test scores (*SD*)	Familiarity (*SD*)
A walk in the forest	Narrative	950	922 (97.8%)	0.32	5.55 (0.56)	2.85 (1.12)
Bad habits	Narrative	1023	985 (96.0%)	0.35	4.39 (1.09)	2.48 (0.81)
Mystery train	Narrative	1035	995 (96.2%)	0.35	5.19 (0.98)	2.03 (0.95)
Animals in the city	Expository (Society)	1004	955 (95.0%)	0.37	5.39 (0.8)	2.42 (1.06)
Street art	Expository (Society)	999	952 (95.1%)	0.39	3.86 (1.03)	2.97 (1.09)
Youtuber	Expository (Society)	983	945 (95.9%)	0.38	4.76 (0.79)	2.34 (0.67)
Dogs and avocados	Expository (Science)	977	929 (95.0%)	0.39	4.9 (1.01)	2.73 (1.05)
Dreams	Expository (Science)	1006	964 (95.8%)	0.34	4.2 (1.1)	3.24 (0.95)
Hand sensor	Expository (Science)	985	936 (95.0%)	0.38	3.73 (0.83)	3.9 (0.67)

Note. The comprehension score was the mean of the comprehension test scores per text (max = 6) across all participants. Topic familiarity was rated by the participants with a five-point Likert scale.

https://doi.org/10.1017/S0272263126101818 Published online by Cambridge University Press

describing natural phenomena and technology (e.g., biology, medicine, engineering), and Society, which included texts addressing social and cultural issues (e.g., lifestyle, art, media).

Accordingly, the selection and categorization of texts were initially based on the evaluation by relevant stakeholders specific to our target context (i.e., EFL contexts in Japan). The first author, a Japanese EFL teacher, searched for candidate texts by checking titles and reading the full passages. When the content was not overly specific and could be classified into one of the three text categories (narrative, science-themed expository, or society-themed expository), the text was retained. Another experienced EFL instructor verified the classification and the appropriateness of the content for their undergraduate students. Two Japanese EFL learners (one advanced and one intermediate), who did not participate in the experiment, also confirmed that the classification was legitimate, denoting perceptually distinctive features across three different types of texts. To confirm whether the selected materials represent distinctive linguistic properties across different genres and topics (i.e., for the linguistic features approach, see Biber & Conrad, 2009), we conducted corpus-based analyses of the target texts using Coh-Metrix indices (Graesser et al., 2004, 2011). The results showed that narrative texts exhibited higher Narrativity (i.e., $M = 1.58$ for narrative texts, whereas $M = 0.12$ for both topics of expository texts), confirming that our perceptual classification was supported by quantitative analysis of the linguistic characteristics (for the full results, see Online Appendix S2).

After the text selection, we seeded 10 target words so that each word appeared twice per text. Because participants in the experimental groups read three texts, they encountered each target word six times in total. The six encounters are regarded as a minimal threshold of repetition leading to substantial vocabulary learning (e.g., Rott, 1999; Uchihara et al., 2019). To minimize the primacy and recency effects on learning, target words were not placed at the beginning or end of sentences. The results of a pilot study with different participants of similar proficiency levels indicated that most of our participants mastered the most frequent 2,000 words. Thus, to ensure that over 95% of the words in each text were within the most frequent 2,000 words (Laufer, 1989), we paraphrased or replaced low-frequency words with high-frequency synonyms. The 95% threshold was determined based on both methodological and empirical considerations. Methodologically, seeding 20 occurrences of the target words in a 1,000-word text resulted in a maximum achievable coverage of approximately 98%, even before other necessary contextual modifications were made. Empirically, previous studies on lexical coverage have suggested that repeated encounters with unknown words, together with opportunities to infer their meanings from context, can potentially increase lexical coverage as learners read a text, thereby facilitating text comprehension (Laufer, 2020; for reviews, see Webb, 2021). The frequency profile analysis using the Compleat Lexical Tutor (Cobb, n.d., ver. 8.5, accessed on 1 January 2024) indicated that all adapted texts contained 95% of words within the most frequent 2,000 words level, plus other words considered familiar to our participants (proper nouns and loanwords). After the text modification, two native English speakers reviewed all texts to ensure grammatical accuracy and naturalness. Based on their feedback, additional adjustments were made to rectify any unnatural phrasing. The classroom instructor reviewed all texts and confirmed that the level of the texts was appropriate for the students. Table 2 shows that the proportion of low-frequency words, type-token ratios, and word counts are comparable across all texts.

https://doi.org/10.1017/S0272263126101818 Published online by Cambridge University Press

Table 3. Target words, Japanese translations, and frequency level

Target words	Japanese translations	Frequency level
[N] baboon	ヒヒ	10K
[N] lore	知識	9K
[N] luster	つや	8K
[N] paucity	不足	10K
[N] plateau	高原	6K
[N] surveillance	見張り	5K
[V] encumber	妨げる	10K
[V] insinuate	ほのめかす	9K
[V] relish	楽しむ	5K
[V] soothe	やわらげる	5K

Note. [N] indicates nouns, and [V] refers to verbs.

Target words

We selected 10 target words comprising 6 nouns and 4 verbs (see Table 3). The ratio of nouns to verbs (6:4) reflected the distribution in natural language use (Webb, 2007). Given that some target words carried multiple meanings, we ensured that only one meaning was used consistently across all texts to avoid semantic ambiguity. Using Nation's (2012, 2016) BNC/COCA 25,000-word lists, we sampled target words from the most frequent 5,000 (5K) to 11,000 (11K) words. We determined this range by referring to the results of the uVLT administered to learners of comparable proficiency in the previous semester and through consultation with the instructor of the target class (i.e., higher than 5K level).[1] The results of uVLT confirmed the frequency range falling outside our participants' lexical knowledge (see Table 1). One instructor of academic English courses and two Japanese graduate students, advanced users of L2 English, reviewed the selected words and confirmed that the words were unlikely to be familiar to the participants.

Text comprehension and topic familiarity

To ensure that learners stay focused on the meaning of the texts while engaging in the reading activity, we set six True or False questions for each text. We developed the questions from the parts that did not involve understanding of the target items to avoid practice effects resulting from engagement with the words. After the creation of the questions, we piloted the difficulty with two Japanese EFL learners who did not participate in this study. We confirmed no significant difference in comprehension scores between the HCD ($M = 4.76$, $SD = 1.16$) and LCD ($M = 4.61$, $SD = 1.03$) with a Welch's t-test, $t(88.48) = 0.805$, $p = .423$.[2] Since the questions were included for the purpose of drawing learners' attention to the content as part of the methodological

[1]One anonymous reviewer raised the potential issue caused by the inclusion of relatively higher frequency words (i.e., 5K items) in our target vocabulary. To address this concern, we conducted a follow-up analysis with the frequency of the target items included in the statistical model and confirmed that it did not significantly affect vocabulary gains (see Online Appendix S3).

[2]We kept the analysis of the comprehension scores minimal (i.e., we reported only descriptive statistics and examined between-group differences to verify the comparable difficulty of the reading materials). See Online Appendix S4 for the full results on internal consistency (Cronbach's alpha) and intraclass correlation coefficients (ICCs) based on variance components of the random effects.

https://doi.org/10.1017/S0272263126101818 Published online by Cambridge University Press

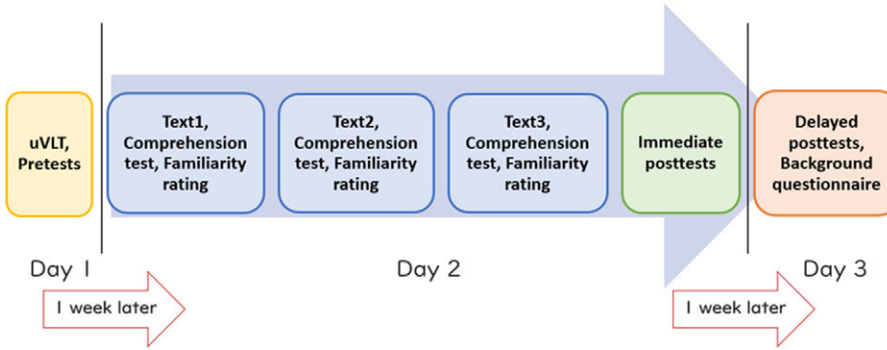


Figure 1. The procedure for the experimental groups over three weeks.

operationalization of incidental vocabulary learning (Uchihara et al., 2019), we did not use the comprehension scores for further statistical analysis. Topic familiarity was measured via a five-point Likert scale. The participants were required to rate their familiarity with the topics after reading each text. There was no significant difference between the HCD ($M = 2.74$, $SD = 1.09$) and LCD ($M = 2.79$, $SD = 1.05$) groups in topic familiarity, $t(85.93) = -0.27$, $p = .787$.

Procedure

Data collection lasted for three weeks (for the schedule, see Figure 1). In Week 1, participants consented to participate and then completed two pretests (i.e., meaning recall and recognition, in this order) followed by the updated VLT. This test order is recommended by Nation and Webb (2011), as it prevents test takers from figuring out the answers from the options provided in the recognition test, which in turn may risk inflating the scores of the recall test administered subsequently. Note that presentation of target items was randomized across participants as well as across three test timings to prevent potential learning by taking the same tests multiple times.

One week later (Week 2), we conducted cluster random sampling with convenience sampling (Mackey & Gass, 2022). First, one class was allocated to a test-only control group, and three classes were assigned to experimental groups. Next, to further control the effects of text presentation order and other potential variables pertaining to text-related features (e.g., topic familiarity and text difficulty), three subgroups were created within each experimental condition and counterbalanced (see Table 4). Within each of the three experimental classes, participants were individually and randomly assigned to one of the subsets in either the HCD or LCD condition. These subgroups read different

Table 4. The subgroups in the experimental groups

Set-H1	Set-H2	Set-H3	Set-L1	Set-L2	Set-L3
Narrative 1	Society 2	Science 3	Narrative 1	Science 1	Society 1
Science 1	Narrative 2	Society 3	Narrative 2	Science 2	Society 2
Society 1	Science 2	Narrative 3	Narrative 3	Science 3	Society 3

Note. Sets H1–3 were used for the high contextual diversity (HCD) condition; Sets L1–3 were used for the low contextual diversity (LCD) condition.

https://doi.org/10.1017/S0272263126101818 Published online by Cambridge University Press

texts in varying orders. During the learning session, participants read three texts in their assigned condition and answered six true/false binary comprehension questions and completed ratings of topic familiarity after reading each text. After completing the treatment, participants took unannounced immediate posttests.

In Week 3, participants took unannounced delayed posttests and completed a questionnaire. The questionnaire asked about their demographic information, English learning experience, daily use of English, and whether they had looked up the meanings of target words at any stage.

Because this study focused on incidental vocabulary learning, participants were not informed about the true purpose of the study until the experiment was completed. After the questionnaire was completed, we debriefed about the study's purpose and procedures in detail, and participants were given opportunities to ask questions. They were also assured that their decision to participate in this research would not affect their course grades. Moreover, participants were informed that they could withdraw their consent to provide data at any time.

Scoring and analysis

For the meaning recall test, we first adopted both strict and lenient scoring methods. In the strict scoring, only responses that matched the exact meanings of the target words (e.g., L1 translations) or described the word meanings represented in the reading materials were marked as correct (e.g., for the word *baboon*: "animals living in forests and sometimes come down to cities"). In the lenient scoring, we additionally gave credit for responses denoting partial information about the target words (e.g., *baboon*: "animal"). The lenient scoring resulted in nearly all answers being marked as correct; therefore, we used the scores based on the strict scoring for statistical analyses. The first author marked all answers, and a Japanese English teacher scored 10 % of the entire responses independently (Mackey & Gass, 2022). Cohen's kappa coefficient was 0.83, indicating sufficient inter-rater consistency (Landis & Koch, 1977). For the meaning recognition test, one point was assigned for each correct answer.

Cronbach's alpha was calculated to check reliability for the meaning recall and meaning recognition tests at immediate and delayed posttests: meaning recall (immediate posttest: α = .80; delayed posttest: α = .69), meaning recognition (immediate posttest: α = .49; delayed posttest: α = .56). Although alphas for meaning recall test scores can be considered within the acceptable range (i.e., closer to or higher than α = .70), those for meaning recognition tests appeared relatively low. To further evaluate the reliability of the meaning recognition test and to estimate the number of additional items required to attain acceptable reliability, we conducted two supplemental analyses. First, we examined a test-retest consistency based on the test scores from the test-only control group (see Schmitt et al., 2020, for the suggestion). This analysis confirmed that there were (a) moderate to strong correlations between test scores (r = .54 to .78, ps < .001), (b) no significant differences in the mean test scores (ps > .05), and homogeneity of variance (Levene's test, p > .05) across three test timings (see Online Appendix S5 for the results in greater detail). These results suggested the sufficient internal consistency of the meaning recognition test, at least for the research purpose (i.e., an experimental study with not a high-stakes purpose). We interpreted the relatively small alpha mainly because of the relatively small number of target items (k = 10). See Online Appendix S5 for the full

https://doi.org/10.1017/S0272263126101818 Published online by Cambridge University Press

results of follow-up analysis to verify that the reduced alpha was mainly attributed to the limited number of items rather than the issue with the reliability of the test.

Data analysis

Prior to the main inferential analyses, we conducted outlier inspection (defined as data points exceeding ± 2.5 *SD*s). Missing data and responses from participants who reported looking up target word meanings outside of the classroom were also excluded from the analysis (n = 40). Therefore, data from the remaining 124 participants were analyzed.

To answer two research questions, we used R (version 4.5.2) to conduct inferential statistical analyses. To confirm the extent to which vocabulary gain is attributable to treatment rather than test effects, and more importantly, to examine whether CD leads to vocabulary learning (RQ1), we conducted generalized linear mixed effects modeling (GLMM) to compare the difference in vocabulary learning among three groups (two experimental and one control). Since the dependent variables (meaning recall and meaning recognition scores) were coded as binary outcomes (0 = incorrect, 1 = correct), we utilized a logistic regression model. Predictors included Group (0 = Control vs. 1 = HCD vs. 2 = LCD), uVLT (z-score), and Time (0 = Pretest, 1 = Immediate posttest, 2 = Delayed posttest). The uVLT scores were included as a covariate to control for the initial difference in baseline vocabulary knowledge (see Supporting Appendix S1). Topic familiarity was not included as the data were not available for the control group. We included a two-way interaction between the Group and Time because previous studies showed the influence of CD may vary depending on time (Chandy et al., 2025). Random intercepts for participants and items were included to account for individual and item variability. This analysis allowed us to assess the learning through the treatment session (HCD vs. LCD vs. Control) as well as differences among the groups (for additional analysis with a three-way interaction among Group, Time, and uVLT, see Supplemental Material S9).

Second, to further scrutinize the influence of CD (i.e., HCD-LCD contrast specifically) while controlling for potential variables such as topic familiarity and prior knowledge of target items, and the interaction between CD and prior vocabulary knowledge (RQ2), we constructed separate GLMM models for each immediate and delayed posttest. Predictors included CD (0 = LCD vs. 1 = HCD), uVLT (z-score), Pretest score (0 = incorrect, 1 = correct), and topic familiarity (z-score) with an interaction term between CD and uVLT. Including pretest scores (as recommended by Maie et al., 2024) and topic familiarity as covariates enabled us to investigate the effects of CD while accounting for prior vocabulary knowledge and familiarity with the reading topics. To avoid building unduly complicated models and to reveal when the CD effects emerged, we built separate models for each time point and the meaning recall and recognition test. Cohen's d was calculated from the odds ratio using the formula below (Chinn, 2000):

$$d = \log(OR) \times \frac{\sqrt{3}}{\pi} \approx \log(OR) \times 0.5513$$

We evaluated models by checking the variance inflation factor (VIF), overdispersion, and residual Q-Q plots with the *DHARMa* package (see Session Information through OSF). All adjusted VIF values were below 9 (see Online Supporting Appendix S6), suggesting an acceptable range of multicollinearity (Hair et al., 2019), and the dispersion

https://doi.org/10.1017/S0272263126101818 Published online by Cambridge University Press

parameters were well below 1, indicating no overdispersion (Bolker et al., 2009). Q–Q plots showed acceptable distributional patterns of residuals, and scatterplot matrices of continuous predictors indicated approximately linear relationships among variables, no extreme outliers, and no strong pairwise correlations ($r > .70$; see Online Appendix S8).

Results

RQ1: Group comparisons (HCD vs. LCD vs. control)

For the meaning recall test, LCD yielded slightly larger relative gains than HCD at the immediate posttest (15.5% for LCD vs. 14.4% for HCD). In contrast, at the delayed posttest, HCD showed greater relative gains than LCD (26.0% for HCD vs. 21.0% for LCD) (see Table 5). These results suggest that while LCD may facilitate initial encoding, HCD may be more beneficial for longer-term retention. The control group showed little learning at the immediate posttest (relative gain with the pretest score as the baseline = 1.9%), yet it demonstrated a relatively larger gain at the delayed posttest (9.5%), suggesting potential test effects. Given the potential test effect and the limited number of target words, we also compared absolute (raw) gain scores (posttest minus pretest) between each experimental group and the control group. Participants in the HCD condition gained 1.13 words and 1.65 words at the delayed posttest, whereas those in the LCD condition learned 1.24 and 1.07 words for the immediate and delayed posttest, respectively.

The GLMM results revealed no significant main effect of Group, indicating that overall meaning recall accuracy of each HCD and LCD was not significantly different from that of the control group (see Table 6 and Figure 2). The main effect of Time was significant for the delayed posttest, indicating higher odds of a correct response at the delayed posttest than at the pretest across the three groups. Interaction between Group and Time was also significant for each combination except LCD × Delayed posttest. The post hoc test (for the full results, see Online Supporting Appendix S7) showed that at the immediate posttest, both experimental groups achieved significantly higher odds of a correct response than the control group ($ps = .001$; $d = 0.82$ for HCD, $d = 0.80$ for LCD). At the delayed posttest, both the HCD and LCD groups showed significantly higher odds of correct response than the control group ($p < .001$, $d = 0.70$ for HCD; $p = .046$, $d = 0.46$ for LCD). As for the comparison between HCD and LCD, no significant differences were observed across any testing time ($ps > .05$).

Table 5. Means, standard deviations (in brackets), and 95% confidence intervals for meaning recall

	Pretest	Immediate posttest	Delayed posttest	Relative gain (%) (pre-immediate)	Relative gain (%) (pre-delayed)
Group					
HCD ($n = 50$)	0.84	2.16	3.24	14.4	26.0
	(0.87)	(1.62)	(2.10)	(16.2)	(32.4)
	CI [0.59, 1.09]	CI [1.70, 2.62]	CI [2.64, 3.84]	CI [9.84, 19.0]	CI [19.7, 32.4]
LCD ($n = 42$)	0.69	2.12	2.64	15.5	21.0
	(0.75)	(1.77)	(1.79)	(17.7)	(26.5)
	CI [0.46, 0.92]	CI [1.57, 2.67]	CI [2.08, 3.20]	CI [9.95, 21.0]	CI [15.4, 26.5]
Control ($n = 32$)	0.53	0.72	1.41	1.91	9.5
	(0.95)	(1.08)	(1.52)	(7.64)	(12.1)
	CI [0.19, 0.87]	CI [0.33, 1.11]	CI [0.86, 1.95]	CI [−0.85, 4.66]	CI [5.13, 13.9]

Note. Max = 10.

https://doi.org/10.1017/S0272263126101818 Published online by Cambridge University Press

Table 6. GLMM for the meaning recall test (Control vs. HCD vs. LCD)

Predictor	β	*OR* [95% CI]	*z*	*p*	Cohen's *d* [95% CI]
Intercept	−3.93	0.02 [0.01, 0.06]	−7.27	<.001	−2.17 [−2.75, −1.58]
Group: HCD	0.39	1.47 [0.64, 3.37]	0.91	.364	0.21 [−0.25, 0.67]
Group: LCD	0.11	1.11 [0.47, 2.65]	0.24	.809	0.06 [−0.42, 0.54]
Time: Immediate	0.42	1.53 [0.74, 3.14]	1.15	.249	0.23 [−0.16, 0.63]
Time: Delayed	1.47	4.37 [2.26, 8.45]	4.38	<.001	0.81 [0.45, 1.18]
uVLT	0.56	1.75 [1.39, 2.22]	4.68	<.001	0.31 [0.18, 0.44]
HCD × immediate	1.10	3.01 [1.30, 7.00]	2.56	.010	0.61 [0.14, 1.07]
LCD × immediate	1.35	3.85 [1.59, 9.28]	3.00	.003	0.74 [0.26, 1.23]
HCD × delayed	0.88	2.41 [1.10, 5.28]	2.21	.027	0.49 [0.05, 0.92]
LCD × delayed	0.73	2.07 [0.91, 4.73]	1.74	.082	0.40 [−0.05, 0.86]

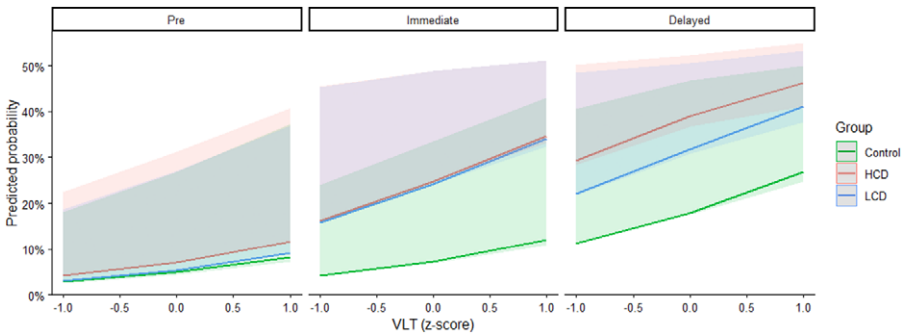


Figure 2. Regression lines and 95% confidence intervals by group and time for the meaning recall test.

For the meaning recognition test, LCD yielded larger relative gains than HCD at the immediate posttest (27.7% for LCD vs. 23.7% for HCD). In contrast, at the delayed posttest, HCD demonstrated greater relative gains than LCD (26.0% for HCD vs. 18.0% for LCD) (see Table 7). Absolute gains showed that participants in the HCD group learned 1.46 words at the immediate posttest and 1.52 words at the delayed posttest. Participants in the LCD condition attained 1.89 words at the immediate posttest and 1.28 words at the

Table 7. Means, standard deviations (in brackets), and 95% confidence intervals for meaning recognition

	Pretest	Immediate posttest	Delayed posttest	Relative gain (%) (pre-immediate)	Relative gain (%) (pre-delayed)
Group					
HCD	3.80	5.32	5.48	23.7	26.0
(*n* = 50)	(1.64)	(2.14)	(2.16)	(34.0)	(33.2)
	CI [3.39, 4.21]	CI [4.71, 5.93]	CI [4.87, 6.09]	CI [14.0, 33.3]	CI [16.5, 35.4]
LCD	3.24	5.19	4.55	27.7	18.0
(*n* = 42)	(1.16)	(1.71)	(1.67)	(27.8)	(25.6)
	CI [2.88, 3.60]	CI [4.66, 5.72]	CI [4.03, 5.07]	CI [19.1, 36.4]	CI [10.0, 26.0]
Control	3.72	3.78	3.69	−3.62	−7.19
(*n* = 32)	(1.71)	(1.84)	(1.65)	(38.2)	(40.2)
	CI [3.10, 4.33]	CI [3.12, 4.45]	CI [3.09, 4.28]	CI [−17.4, 10.1]	CI [−21.7, 7.30]

Note. Max = 10.

https://doi.org/10.1017/S0272263126101818 Published online by Cambridge University Press

Table 8. GLMM for the meaning recognition test (Control vs. HCD vs. LCD)

Predictor	β	*OR* [95% CI]	*z*	*p*	Cohen's *d* [95% CI]
Intercept	−0.45	0.64 [0.32, 1.26]	−1.30	.194	−0.25 [−0.62, 0.13]
Group: HCD	−0.25	0.78 [0.52, 1.18]	−1.18	.240	−0.14 [−0.36, 0.09]
Group: LCD	−0.53	0.59 [0.38, 0.90]	−2.43	.015	−0.29 [−0.53, −0.06]
Time: Immediate	0.03	1.03 [0.72, 1.48]	0.19	.853	0.02 [−0.18, 0.22]
Time: Delayed	−0.02	0.98 [0.69, 1.41]	−0.09	.926	−0.01 [−0.21, 0.19]
uVLT	0.36	1.44 [1.27, 1.63]	5.63	<.001	0.20 [0.13, 0.27]
HCD × immediate	0.75	2.12 [1.34, 3.36]	3.22	.001	0.42 [0.16, 0.67]
LCD × immediate	0.98	2.67 [1.66, 4.29]	4.04	<.001	0.54 [0.28, 0.80]
HCD × delayed	0.89	2.43 [1.53, 3.84]	3.78	<.001	0.49 [0.24, 0.74]
LCD × delayed	0.71	2.03 [1.26, 3.27]	2.91	.004	0.39 [0.13, 0.65]

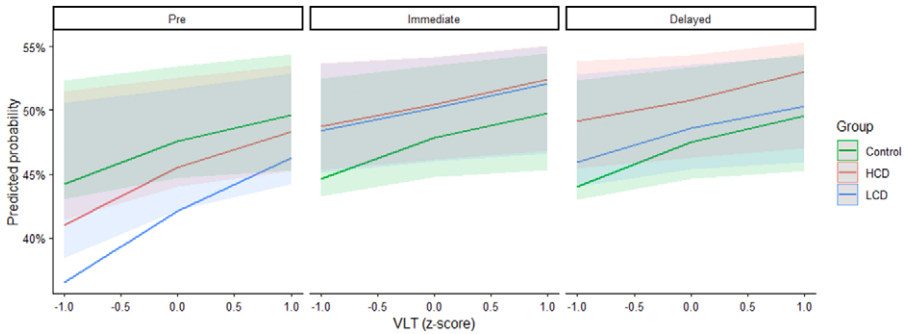


Figure 3. Regression lines and 95% confidence intervals by group and time for the meaning recognition test.

delayed posttest. These results suggest a similar trend that while LCD may be beneficial for initial encoding immediately after reading, HCD may facilitate longer-term retention.

The results of GLMM showed the main effect of Group was significant for LCD, suggesting the LCD condition tended to achieve higher accuracy than the control group. The main effect of Time was not significant. Interaction between Group × Time was significant for all combinations, indicating both HCD and LCD achieved significantly higher recognition accuracy in the immediate posttest and the delayed posttest (see Table 8 and Figure 3). Post-hoc pairwise comparisons (see Online Supporting Appendix S7) showed HCD showed significantly higher odds of a correct response, suggesting better recognition accuracy than the control group at the immediate posttest ($p = .04$, $d = 0.28$), whereas no significant difference was observed between LCD and the control group ($p = .091$). At the delayed posttest, HCD outperformed the control group with higher odds of a correct response ($p = .006$, $d = 0.35$), while LCD did not, indicating that only HCD led to significantly higher recognition accuracy than the control group (see Online Appendix S7). Between HCD and LCD, the former group outperformed the latter at the delayed posttest ($p = .033$, $d = 0.26$), suggesting that higher CD resulted in greater retention of target items at the recognition level.

https://doi.org/10.1017/S0272263126101818 Published online by Cambridge University Press

Table 9. GLMM for the meaning recall test (immediate posttest)

Predictor	β	*OR* [95% CI]	*Cohen's d* [95% CI]	*z*	*p*
Intercept	–2.12	0.12 [0.06, 0.25]	–1.17 [–1.56, –0.77]	–5.80	<.001
CD: HCD	0.01	1.01 [0.57, 1.80]	0.01 [–0.31, 0.32]	0.03	.975
uVLT	0.53	1.69 [1.06, 2.68]	0.29 [0.03, 0.54]	2.22	.026
Familiarity	0.08	1.09 [0.83, 1.43]	0.05 [–0.10, 0.20]	0.60	.549
Pretest	2.86	17.46 [7.11, 43.01]	1.58 [1.08, 2.07]	6.23	<.001
HCD × uVLT	–0.01	0.99 [0.52, 1.87]	–0.01 [–0.36, 0.35]	–0.04	.967

Table 10. GLMM for the meaning recall test (delayed posttest)

Predictor	β	*OR* [95% CI]	Cohen's *d* [95% CI]	*z*	*p*
Intercept	–1.65	0.19 [0.09, 0.40]	–0.91 [–1.32, –0.51]	–4.40	<.001
CD: HCD	0.38	1.46 [0.86, 2.49]	0.21 [–0.08, 0.50]	1.41	.159
uVLT	0.24	1.27 [0.83, 1.95]	0.13 [–0.10, 0.37]	1.09	.274
Familiarity	–0.13	0.88 [0.68, 1.13]	–0.07 [–0.21, 0.07]	–1.03	.305
Pretest	2.45	11.60 [4.86, 27.60]	1.35 [0.87, 1.83]	5.52	<.001
HCD × uVLT	0.17	1.18 [0.66, 2.14]	0.09 [–0.23, 0.42]	0.56	.575

RQ2: Prior vocabulary knowledge as a moderator of CD effects

For the meaning recall test, no significant effect of CD emerged at either the immediate or delayed posttest, indicating CD did not affect meaning recall accuracy significantly (see Tables 9 and 10). Vocabulary knowledge (uVLT) was a significant predictor in the immediate posttest (Immediate: $\beta = 0.53$, $p = .026$, $d = 0.29$), suggesting that learners with larger vocabulary knowledge achieved higher odds of correctly recognizing the target meaning. None of the CD × uVLT interactions reached significance, suggesting

Table 11. GLMM for the meaning recognition test (immediate posttest)

Predictor	β	*OR* [95% CI]	Cohen's *d* [95% CI]	*z*	*p*
Intercept	–0.35	0.71 [0.42, 1.19]	–0.19 [–0.48, 0.09]	–1.31	.190
CD: HCD	–0.16	0.85 [0.58, 1.25]	–0.09 [–0.31, 0.13]	–0.82	.411
uVLT	0.03	1.03 [0.76, 1.41]	0.02 [–0.15, 0.19]	0.20	.844
Familiarity	–0.08	0.92 [0.76, 1.12]	–0.04 [–0.15, 0.06]	–0.82	.415
Pretest	1.58	4.84 [3.34, 7.03]	0.87 [0.67, 1.08]	8.31	<.001
HCD × uVLT	0.62	1.85 [1.19, 2.89]	0.34 [0.09, 0.59]	2.71	.007

Table 12. GLMM for the meaning recognition test (delayed posttest)

Predictor	β	*OR* [95% CI]	Cohen's *d* [95% CI]	*z*	*p*
Intercept	–0.91	0.40 [0.22, 0.75]	–0.50 [–0.84, –0.16]	–2.87	.004
CD: HCD	0.42	1.51 [0.97, 2.37]	0.23 [–0.02, 0.48]	1.82	.069
uVLT	0.41	1.51 [1.05, 2.16]	0.23 [0.03, 0.43]	2.21	.027
Familiarity	–0.05	0.95 [0.76, 1.19]	–0.03 [–0.15, 0.09]	–0.44	.659
Pretest	2.04	7.65 [5.09, 11.50]	1.12 [0.90, 1.35]	9.79	<.001
HCD × uVLT	0.09	1.09 [0.66, 1.81]	0.05 [–0.23, 0.33]	0.34	.736

https://doi.org/10.1017/S0272263126101818 Published online by Cambridge University Press

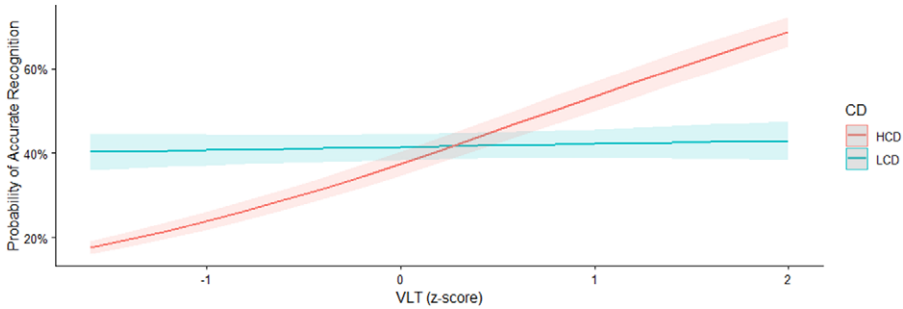


Figure 4. Interaction between CD and uVLT for the meaning recognition at the immediate posttest.

that the influence of CD did not differ depending on learners' vocabulary knowledge across each test timing at the recall level.

For the meaning recognition test, no CD effect was found at the immediate posttest, and at the delayed posttest, suggesting CD did not result in significantly different accuracy in the meaning recognition test (see Tables 11 and 12). Vocabulary knowledge (uVLT) also showed significant main effects at the delayed posttest ($\beta = 0.41$, $p = .027$, $d = 1.51$), indicating the importance of prior vocabulary knowledge for the retention of target items at the recognition level.

As for the interaction between CD × uVLT, it is notable that the interaction was significant at the immediate posttest ($\beta = 0.62$, $p = .007$, $d = 0.34$; see Figure 4). This interaction indicates that the effect of CD varied according to learners' vocabulary size: learners with larger vocabulary knowledge were more likely to achieve higher recognition accuracy under the HCD condition, whereas those with smaller vocabulary knowledge tended to benefit more from the LCD condition, especially in the immediate posttest. At the delayed posttest, this interaction was no longer significant ($\beta = 0.09$, $p = .736$), suggesting that the moderating influence of vocabulary knowledge on CD effects was prominent in the early stages of learning.

Discussion

This study investigated how CD affects incidental vocabulary learning from reading while considering the moderating influence of prior vocabulary knowledge. In terms of relative gains, experimental groups achieved more than 10% higher in meaning recall and more than 20% higher in meaning recognition than those in the test-only control group, suggesting that reading texts with carefully controlled lexical coverage (i.e., six encounters with target items in three texts with higher than 95% lexical coverage in this study) may facilitate incidental vocabulary learning to some extent. The results with absolute gains also indicated that although both experimental groups showed learning, the amount is not so large (approximately 1 to 2 words out of the 10 target words), which is aligned with previous meta-analysis (17% for immediate posttest and 15% for the delayed posttest, Webb et al., 2017). In addition, although the GLMM results suggested the possible superiority of HCD over LCD, especially in the delayed meaning recognition test, the absolute difference was only 0.24 (1.52 in HCD vs. 1.28 in LCD).

Regarding RQ1 on the effects of CD, although HCD and LCD led to similar learning gains, results suggested the potential advantage of HCD, especially in the meaning recognition test. In the first model, which included the test-only control group as a

https://doi.org/10.1017/S0272263126101818 Published online by Cambridge University Press

reference group, HCD achieved significantly higher accuracy than the control group on the meaning recognition test at both the immediate (p = .040, d = 0.28) and delayed posttests (p = .006, d = 0.35), whereas LCD did not (p = .091 for the immediate posttest and p = .688 for the delayed posttest). Although the effect size was small, these findings provide indirect evidence supporting the benefit of higher contextual variability for incidental vocabulary learning. Furthermore, post hoc analysis for the meaning recognition test in the first model revealed a significantly higher score at the delayed posttest for HCD compared to LCD (p = .033, d = 0.26, 95% CI [0.06, 0.45]). However, the results of the second model also showed a tendency that HCD outperformed LCD on the delayed meaning recognition test, again with a small effect (p = .069, d = 0.23). Taken together, although the effect sizes were small, the results suggest that increased contextual variability may facilitate meaning recognition, particularly if such effects accumulate over the course of a long-term, extensive reading program. At the same time, this interpretation should be treated cautiously because the effect sizes were modest and the second model suggested that the effect of CD varied depending on learners' baseline vocabulary knowledge. Below, we discuss cognitive mechanisms underlying the potential advantage of contextually varied input for L2 vocabulary learning.

First, greater CD may be more likely to result in deeper cognitive processing (Craik & Lockhart, 1972; Craik & Tulving, 1975). Exposure to contextually diverse texts likely promoted greater noticing of and attention to unfamiliar words during reading (Chandy et al., 2025; Pagán & Nation, 2019), engaging deeper lexical processing. Compared to the LCD group, the HCD group may have paid more attention to unknown words with deeper processing induced by exposure to greater contextual variability at the discourse level. The level-of-processing account was also drawn to explain the benefit of varied input over verbatim repetition, as argued by Liu and Todd (2016), that learners' reading different texts helps maintain a higher level of attentional engagement, whereas repeated reading lowers the perceived need to understand unknown words for achieving overall comprehension.

Although different in focus (i.e., global vs. local CD), the present results align with the argument by Chandy et al. (2025) that, despite small immediate learning, greater CD appeared to foster more stable long-term retention with a sustained greater attention to target items. In their study, although reading a single text repeatedly (lower CD) led to greater learning gains in the immediate meaning recognition test compared to reading three different texts (higher CD), it also led to greater loss of the learning. This suggests that lower CD may be more conducive to initial word encoding due to the availability of coherent contextual clues; however, the resulting processing may be shallower than that of higher CD. Greater contextual variability requires learners to engage more deeply with unknown words because broader contextual clues are introduced and encoded at each encounter with the same L2 word. The deeper processing induced by higher CD may account for the potential advantage of HCD for retention observed in this study.

Furthermore, the present results support the instance-based memory model of word learning (Bolger et al., 2008), which suggests that CD fosters more consolidated word memories with incremental resonant memory of different contexts. Although participants read different texts in both experimental conditions, greater variability of genres and topics in which a target word appeared for the HCD condition may provide more memory traces for the resonance process, which might have strengthened word memory. Taken together, these findings indicate that although reading thematically and discursively related texts (i.e., lower CD) may support text comprehension, greater

https://doi.org/10.1017/S0272263126101818 Published online by Cambridge University Press

CD fosters deeper, more elaborative processing with richer episodic memory, which in turn supports durable lexical retention.

As for RQ2, the moderating effect of prior vocabulary knowledge on CD effects was observed in meaning recognition at the immediate posttest. The significant CD × uVLT interaction indicated that learners with larger vocabulary knowledge benefited more in the HCD condition (i.e., above a score of approximately 125 on the uVLT in this study), whereas those with smaller vocabulary knowledge benefited less. In other words, especially for the encoding of the target words, sufficient prior vocabulary knowledge is a prerequisite for benefiting from higher CD, and lower CD may favor lexically less proficient learners. Due to the limited cognitive resources (Sweller, 1988; Sweller et al., 2019), the processing demands of HCD input may exceed the cognitive capacity for learners with lower lexical proficiency and therefore constrain learning. Because learners in the HCD condition read three different texts that differed in genres and topics, those with limited vocabulary knowledge may have struggled to integrate diverse linguistic and discourse-level information efficiently. The lack of vocabulary knowledge may have reduced the efficiency of bottom-up processing, depleted attentional resources for inefficient lexical retrieval, and left little cognitive resource necessary for making successful inferences about the meaning of unknown words (Kintsch & Welsch, 1991; Sweller et al., 2019). In contrast, for lexically proficient learners, reading contextually diverse texts may provide richer semantic and discourse-level cues, which can activate broader background knowledge and facilitate deeper semantic elaboration and memory encoding (Craik & Lockhart, 1972; Yamashita, 2022).

From a different theoretical viewpoint, the benefit of CD can also be explained by the framework of desirable difficulty (Bjork & Kroll, 2015; Suzuki et al., 2019). Specifically, a greater extent of prior vocabulary knowledge may mitigate the perceived difficulty of the task (i.e., reading for comprehension); consequently, the level of challenge presented in the HCD condition may be more beneficial for learners with high lexical proficiency. Notably, unlike previous studies manipulating CD through lexical or sentence-level variation (e.g., Liu & Todd, 2016), we adopted the global approach to operationalizing the CD condition (i.e., genre and topic levels); therefore, our manipulation of CD likely introduced an additional layer of cognitive demands at the discourse level, beyond lexico-grammatical processing. Especially for the HCD condition, both genre and topic varied across texts, which may have induced a greater cognitive demand. Learners with a larger vocabulary knowledge may have perceived this condition to be challenging yet cognitively stimulating, thus rendering the task difficulty desirable, with the support of vocabulary knowledge. Conversely, for learners with limited vocabulary knowledge, the difficulty could exceed the threshold of a "desirable" level of cognitive difficulty, and it was likely perceived as overwhelmingly demanding. In contrast, the LCD condition may have been relatively undemanding for most learners, since genre and topic were coherent across the three texts. This consistency likely allowed readers to reuse similar background knowledge and discourse schemas, reducing cognitive effort. In this sense, LCD may have been neither too difficult for lower-proficiency learners nor sufficiently challenging for higher-proficiency learners, which could explain the weaker moderating role of vocabulary knowledge in this condition.

Conclusion

This study examined the effects of CD on L2 incidental vocabulary learning, and the findings suggested that CD potentially affects L2 vocabulary learning from reading, with

https://doi.org/10.1017/S0272263126101818 Published online by Cambridge University Press

the effect varying depending on learners' prior vocabulary knowledge. Although HCD and LCD produced comparable gains overall, the benefits of HCD were limited to learners with larger vocabulary knowledge. In contrast, learners in the LCD condition showed robust learning gains regardless of vocabulary size, highlighting LCD input as a broadly facilitative and pedagogically valuable context for learners across proficiency levels.

These findings offer useful insights into how to optimize reading-based vocabulary instruction. First, regarding the influence of variability of reading materials, given that LCD led to learning gains comparable to HCD and that CD showed a significant interaction with uVLT scores, repeated exposure to thematically related content may be particularly beneficial for lower-proficiency learners. Reading texts with coherent genres and topics can activate consistent background knowledge, thereby supporting comprehension and retention (Chang, 2019). Once learners have built a sufficient lexical foundation and discourse familiarity through narrow reading, they may be better prepared to benefit from contextually diverse texts involving varied genres and topics. Second, another promising approach is to scaffold reading input by sequencing activities from narrow reading to more diverse reading contexts. Narrow reading, where learners repeatedly encounter shared vocabulary across texts, can reduce lexical load and facilitate word recognition (Schmitt & Carter, 2000). Therefore, for maximizing incidental vocabulary gains, particularly among lower-proficiency learners, it may be pedagogically beneficial to initiate reading programs with low CD materials first and gradually transition to high CD input as learners' vocabulary knowledge and comprehension skills develop. Teachers also need to take students' background knowledge and lexical proficiency into consideration for designing instruction.

Despite its methodological strengths, the present study has several limitations. First, the findings are based on a small number of target words (k = 10), which may limit the generalizability of the results and should be further tested in future research. This number was a result of our effort to ensure precise control over lexical coverage while maintaining enough repetitions of each target word (i.e., six times). These are the methodological issues identified in a review of previous studies (e.g., Frances et al., 2020) that we tried to address in the current study. To address this issue, future studies should adopt a longitudinal design using longer texts while maintaining sufficient repetition for each target item (e.g., Webb & Chang, 2015a, 2015b). In addition, future research needs to investigate the CD effects longitudinally to clarify how extended exposure to different text types influences vocabulary acquisition over time. Learners may become familiar with a particular discourse type through repeated exposure (Chang, 2019), which could affect the effectiveness of contextually diverse input. Indeed, Mak et al. (2021) found that LCD was more effective for initial word learning, whereas HCD yielded greater benefits once learners had first encountered new words in coherent and familiar contexts. This suggests that the effects of CD may vary depending on the timing and sequencing of exposure, highlighting the need for future research to examine such time-dependent influences of CD.

Supplementary material. The supplementary material for this article can be found at http://doi.org/10.1017/S0272263126101818.

Data availability statement. Data and R code are available in OSF: https://osf.io/5xzj6/overview

Acknowledgments. We would like to express our sincere gratitude to the audience at AAAL in Denver for their helpful feedback. We are also deeply grateful to Dr. Ryuya Komuro and Dr. Ryo Maie for their continued support with data analysis and encouraging feedback. We also thank Dr. Eva Puimège for her valuable comments on refining the research design and Dr. Aline Godfroid for her helpful feedback on interpreting the

https://doi.org/10.1017/S0272263126101818 Published online by Cambridge University Press

results. We are grateful to Victoria-Anne Flood and Jessica Laine Rose for their contributions to material development. Finally, we thank the students who participated in this study.

Funding statement. This project was supported by the JSPS KAKENHI Grant Number 24K00080.

Competing interests. The author declares none.

References

Adelman, J. S., Brown, G. D. A., & Quesada, J. F. (2006). Contextual diversity, not word frequency, determines word-naming and lexical decision times. *Psychological Science*, *17*(9), 814–823. https://doi.org/10.1111/j.1467-9280.2006.01787.x

Arai, Y., & Takizawa, K. (2024). Text and reading task variables in incidental L2 vocabulary learning from reading: A methodological synthesis. *Research Methods in Applied Linguistics*, *3*(2), 100110. https://doi.org/10.1016/j.rmal.2024.100110

Biber, D. (1992). The multi-dimensional approach to linguistic analyses of genre variation: An overview of methodology and findings. *Computers and the Humanities*, *26*(5/6), 331–345.

Biber, D., & Conrad, S. (2009). *Register, genre, and style*. Cambridge University Press. https://doi.org/10.1017/CBO9780511814358

Bjork, R. A., & Kroll, J. F. (2015). Desirable difficulties in vocabulary learning. *The American Journal of Psychology*, *128*(2), 241–252. https://doi.org/10.5406/amerjpsyc.128.2.0241

Bolger, D. J., Balass, M., Landen, E., & Perfetti, C. A. (2008). Context variation and definitions in learning the meanings of words: An instance-based learning approach. *Discourse Processes*, *45*(2), 122–159. https://doi.org/10.1080/01638530701792826

Bolker, B. M., Brooks, M. E., Clark, C. J., Geange, S. W., Poulsen, J. R., Stevens, M. H. H., & White, J.-S. S. (2009). Generalized linear mixed models: A practical guide for ecology and evolution. *Trends in Ecology & Evolution*, *24*(3), 127–135. https://doi.org/10.1016/j.tree.2008.10.008

Cancino, M. (2023). Topic interest increases L2 incidental vocabulary learning and effective dictionary look-up behaviour. *Learning and Motivation*, *83*, 101920. https://doi.org/10.1016/j.lmot.2023.101920

Chandy, R., Serrano, R., & Pellicer-Sánchez, A. (2025). The effect of context on the processing and learning of novel L2 vocabulary while reading. *Applied Psycholinguistics*, *45*(6), 1086–1113. https://doi.org/10.1017/S0142716424000407

Chang, A. C.-S. (2019). Effects of narrow reading and listening on L2 vocabulary learning: Multiple dimensions. *Studies in Second Language Acquisition*, *41*(4), 769–794. https://doi.org/10.1017/S0272263119000032

Chinn, S. (2000). A simple method for converting an odds ratio to effect size for use in meta-analysis. *Statistics in Medicine*, *19*(22), 3127–3131. https://doi.org/10.1002/1097-0258(20001130)19:22<3127::AID-SIM784>3.0.CO;2-M

Cobb, T. (n.d.). *Compleat Lexical Tutor*. Retrieved https://www.lextutor.ca/

Craik, F. I. M., & Lockhart, R. S. (1972). Levels of processing: A framework for memory research. *Journal of Verbal Learning and Verbal Behavior*, *11*(6), 671–684. https://doi.org/10.1016/S0022-5371(72)80001-X

Craik, F. I. M., & Tulving, E. (1975). Depth of processing and the retention of words in episodic memory. *Journal of Experimental Psychology: General*, *104*(3), 268–294. https://doi.org/10.1037/0096-3445.104.3.268

de Bot, K., Paribakht, T. S., & Wesche, M. B. (1997). Toward a lexical processing model for the study of second language vocabulary acquisition. *Studies in Second Language Acquisition*, *19*(3), 309–329.

Durrant, P., & Schmitt, N. (2010). Adult learners' retention of collocations from exposure. *Second Language Research*, *26*(2), 163–188. https://doi.org/10.1177/0267658309349431

Durrant, P., Siyanova-Chanturia, A., Kremmel, B., & Sonbol, S. (2022). *Research methods in vocabulary studies* (1st ed.). John Benjamins Publishing Company.

Engelbart, S. M., & Theuerkauf, B. (1999). Defining context within vocabulary acquisition. *Language Teaching Research*, *3*(1), 57–69. https://doi.org/10.1177/136216889900300104

Frances, C., Martin, C. D., & Duñabeitia, J. A. (2020). The effects of contextual diversity on incidental vocabulary learning in the native and a foreign language. *Scientific Reports*, *10*(1), Article 1. https://doi.org/10.1038/s41598-020-70922-1

https://doi.org/10.1017/S0272263126101818 Published online by Cambridge University Press

Graesser, A. C., McNamara, D. S., & Kulikowich, J. M. (2011). Coh-Metrix: Providing multilevel analyses of text characteristics. *Educational Researcher*, *40*(5), 223–234. https://doi.org/10.3102/0013189X11413260

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

Hair, J. F., Babin, B. J., Anderson, R. E., & Black, W. C. (2019). *Multivariate data analysis* (8th ed.). Pearson.

Horst, M., Cobb, T., & Meara, P. (1998). Beyond a clockwork orange: Acquiring second language vocabulary through reading. *Reading in a Foreign Language*, *11*, 207–223.

Hulme, R. C., Begum, A., Nation, K., & Rodd, J. M. (2023). Diversity of narrative context disrupts the early stage of learning the meanings of novel words. *Psychonomic Bulletin and Review*, *30*, 2338–2350. https://doi.org/10.3758/s13423-023-02316-z

Hulstijn, J. H. (2003). Incidental and intentional learning. In Catherine. J. Doughty & M. H. Long (Eds.), *The handbook of second language acquisition* (pp. 349–381). John Wiley & Sons, Ltd. https://doi.org/10.1002/9780470756492.ch12

Hyland, K. (2002). Genre: Language, context, and literacy. *Annual Review of Applied Linguistics*, *22*, 113–135. https://doi.org/10.1017/S0267190502000065

Hyland, K. (2021). *Teaching and researching writing* (4th ed.). Routledge. https://doi.org/10.4324/9781003198451

Jenkins, J. R., Stein, M. L., & Wysocki, K. (1984). Learning vocabulary through reading. *American Educational Research Journal*, *21*(4), 767–787. https://doi.org/10.3102/00028312021004767

Joe, A. (1998). What effects do text-based tasks promoting generation have on incidental vocabulary acquisition? *Applied Linguistics*, *19*(3), 357–377. https://doi.org/10.1093/applin/19.3.357

Johns, B. T., Dye, M., & Jones, M. N. (2016). The influence of contextual diversity on word learning. *Psychonomic Bulletin & Review*, *23*(4), 1214–1220. https://doi.org/10.3758/s13423-015-0980-7

Joseph, H., & Nation, K. (2018). Examining incidental word learning during reading in children: The role of context. *Journal of Experimental Child Psychology*, *166*, 190–211. https://doi.org/10.1016/j.jecp.2017.08.010

Kintsch, W., & Welsch, D. M. (1991). The construction-integration model: A framework for studying memory for text. In W. E. Hockley & S. Lewandowsky (Eds.), *Relating theory and data: Essays on human memory in honor of Bennet B. Murdock* (pp. 367–385). Lawrence Erlbaum Associates, Inc.

Koda, K. (2018). Integrated communication skills approach: Reading to learn as a basis for language and content integration. In K. Koda & J. Yamashita (Eds.), *Reading to Learn in a Foreign Language*. Routledge. https://doi.org/10.4324/9781315183077

Kremmel, B., Indrarathne, B., Kormos, J., & Suzuki, S. (2023). Unknown vocabulary density and reading comprehension: Replicating Hu and Nation (2000). *Language Learning*, *73*(4), 1127–1163. https://doi.org/10.1111/lang.12622

Kremmel, B., & Schmitt, N. (2016). Interpreting vocabulary test scores: What do various item formats tell us about learners' ability to employ words? *Language Assessment Quarterly*, *13*(4), 377–392. https://doi.org/10.1080/15434303.2016.1237516

Landis, J. R., & Koch, G. G. (1977). The measurement of observer agreement for categorical data. *Biometrics*, *33*(1), 159–174. https://doi.org/10.2307/2529310

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

Laufer, B. (2020). Lexical coverages, inferencing unknown words and reading comprehension: How are they related? *TESOL Quarterly*, *54*(4), 1076–1085. https://doi.org/10.1002/tesq.3004

Lee, D. (2001). Genres, registers, text types, domain, and styles: Clarifying the concepts and navigating a path through the BNC jungle. *Language Learning & Technology*, *5*(3), 37–72. https://doi.org/10.64152/10125/44565

Li, Z., Wu, S., & Liu, D. (2023). The effects of context, story-type, and language proficiency on EFL word learning and retention from reading. *International Review of Applied Linguistics in Language Teaching*, *63*(2), 1507–1531. https://doi.org/10.1515/iral-2023-0155

Liu, Y.-T., & Todd, A. G. (2016). Implementation of assisted repeated reading techniques for the incidental acquisition of novel foreign vocabulary. *Language Teaching Research*, *20*(1), 53–74. https://doi.org/10.1177/1362168814559802

Mackey, A., & Gass, S. M. (2022). *Second language research: Methodology and design* (3rd ed.). Routledge. https://doi.org/10.4324/9781003188414

https://doi.org/10.1017/S0272263126101818 Published online by Cambridge University Press

Maie, R., Eguchi, M., & Uchihara, T. (2024). Arbitrary choices, arbitrary results: Three cases of multiverse analysis in L2 research. *Research Methods in Applied Linguistics*, *3*(2), 100124. https://doi.org/10.1016/j.rmal.2024.100124

Mak, M. H. C., Hsiao, Y., & Nation, K. (2021). Anchoring and contextual variation in the early stages of incidental word learning during reading. *Journal of Memory and Language*, *118*, 104203. https://doi.org/10.1016/j.jml.2020.104203

Mar, R. A., Li, J., Nguyen, A. T. P., & Ta, C. P. (2021). Memory and comprehension of narrative versus expository texts: A meta-analysis. *Psychonomic Bulletin & Review*, *28*(3), 732–749. https://doi.org/10.3758/s13423-020-01853-1

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

Nation, I. S. P. (2012). *The BNC/COCA word family lists*. https://www.wgtn.ac.nz/lals/resources/paul-nations-resources/vocabulary-lists

Nation, I. S. P. (2016). *Making and using word lists for language learning and testing*. John Benjamins Publishing Company.

Nation, I. S. P., & Webb, S. (2011). *Researching and analyzing vocabulary*. Cengage Learning.

Nguyen, T. Q. (2023). Impact of processing various textual genres on the incidental acquisition and retention of L2 vocabulary. *English Teaching*, *78*(2), 139–164. https://doi.org/10.15858/engtea.78.2.202306.139

Norman, R., Hulme, R. C., Sarantopoulos, C., Chandran, V., Shen, H., Rodd, J. M., Joseph, H., & Taylor, J. S. H. (2023). Contextual diversity during word learning through reading benefits generalisation of learned meanings to new contexts. *Quarterly Journal of Experimental Psychology*, *76*(7), 1658–1671. https://doi.org/10.1177/17470218221126976

Pagán, A., & Nation, K. (2019). Learning words via reading: Contextual diversity, spacing, and retrieval effects in adults. *Cognitive Science*, *43*(1), e12705. https://doi.org/10.1111/cogs.12705

Pellicer-Sánchez, A., Webb, S., & Wang, A. (2024). How does lexical coverage affect the processing of L2 texts? *Applied Linguistics*, *45*(6), 953–972. https://doi.org/10.1093/applin/amae062

Pulido, D. (2003). Modeling the role of second language proficiency and topic familiarity in second language incidental vocabulary acquisition through reading. *Language Learning*, *53*(2), 233–284. https://doi.org/10.1111/1467-9922.00217

Pulido, D. (2007). The relationship between text comprehension and second language incidental vocabulary acquisition: A matter of topic familiarity? *Language Learning*, *57*(s1), 155–199. https://doi.org/10.1111/j.1467-9922.2007.00415.x

Reynolds, B. L. (2015). The effects of word form variation and frequency on second language incidental vocabulary acquisition through reading. *Applied Linguistics Review*, *6*(4), 467–497. https://doi.org/10.1515/applirev-2015-0021

Robinson, P. (2003). Attention and memory during SLA. In C. J. Doughty & M. H. Long (Eds.), *The handbook of second language acquisition* (pp. 631–678). John Wiley & Sons, Ltd. https://doi.org/10.1002/9780470756492.ch19

Rosa, E., Salom, R., & Perea, M. (2022). Contextual diversity favors the learning of new words in children regardless of their comprehension skills. *Journal of Experimental Child Psychology*, *214*, 105312. https://doi.org/10.1016/j.jecp.2021.105312

Rosa, E., Tapia, J. L., & Perea, M. (2017). Contextual diversity facilitates learning new words in the classroom. *PLoS ONE*, *12*(6), e0179004. https://doi.org/10.1371/journal.pone.0179004

Rott, S. (1999). The effect of exposure frequency in 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

Schmitt, N., & Carter, R. (2000). The lexical advantages of narrow reading for second language learners. *TESOL Journal*, *9*(1), 4–9. https://doi.org/10.1002/j.1949-3533.2000.tb00220.x

Schmitt, N., Jiang, X., & Grabe, W. (2011). The percentage of words known in a text and reading comprehension. *The Modern Language Journal*, *95*(1), 26–43. https://doi.org/10.1111/j.1540-4781.2011.01146.x

Schmitt, N., Nation, P., & Kremmel, B. (2020). Moving the field of vocabulary assessment forward: The need for more rigorous test development and validation. *Language Teaching*, *53*(1), 109–120. https://doi.org/10.1017/S0261444819000326

https://doi.org/10.1017/S0272263126101818 Published online by Cambridge University Press

Smith, R., Snow, P., Serry, T., & Hammond, L. (2021). The role of background knowledge in reading comprehension: A critical review. *Reading Psychology*, *42*(3), 214–240. https://doi.org/10.1080/02702711.2021.1888348

Suchomel, V. (2021). Genre annotation of web corpora: Scheme and issues. Proceedings of the Future Technologies Conference (FTC) 2020, 1, 738–754. https://doi.org/10.1007/978-3-030-63128-4_55

Suzuki, Y., Nakata, T., & Dekeyser, R. (2019). The desirable difficulty framework as a theoretical foundation for optimizing and researching second language practice. *The Modern Language Journal*, *103*(3), 713–720. https://doi.org/10.1111/modl.12585

Sweller, J. (1988). Cognitive load during problem solving: Effects on learning. *Cognitive Science*, *12*(2), 257–285. https://doi.org/10.1207/s15516709cog1202_4

Sweller, J., van Merriënboer, J. J. G., & Paas, F. (2019). Cognitive architecture and instructional design: 20 years later. *Educational Psychology Review*, *31*(2), 261–292. https://doi.org/10.1007/s10648-019-09465-5

Teng, M. F. (2024a). Incidental vocabulary learning from listening, reading, and viewing captioned videos: Frequency and prior vocabulary knowledge. *Applied Linguistics Review*, *16*(1), 477–507. https://doi.org/10.1515/applirev-2023-0106

Teng, M. F. (2024b). Working memory and prior vocabulary knowledge in incidental vocabulary learning from listening, reading, reading-while-listening, and viewing captioned videos. *System*, *124*, 103381. https://doi.org/10.1016/j.system.2024.103381

Uchihara, T., Webb, S., Saito, K., & Trofimovich, P. (2022). The effects of talker variability and frequency of exposure on the acquisition of spoken word knowledge. *Studies in Second Language Acquisition*, *44*(2), 357–380. https://doi.org/10.1017/S0272263121000218

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

Vilkaitė, L. (2017). Incidental acquisition of collocations in L2: Effects of adjacency and prior vocabulary knowledge. *International Journal of Applied Linguistics*, *168*(2), 248–277. https://doi.org/10.1075/itl.17005.vil

Vu, D. V., & Peters, E. (2022). Incidental learning of collocations from meaningful input: A longitudinal study into three reading modes and factors that affect learning. *Studies in Second Language Acquisition*, *44*(3), 685–707. https://doi.org/10.1017/S0272263121000462

Vu, D. V., & Peters, E. (2023). A longitudinal study on the effect of mode of reading on incidental collocation learning and predictors of learning gains. *TESOL Quarterly*, *57*(1), 5–32. https://doi.org/10.1002/tesq.3111

Waring, R., & Takaki, M. (2001). At what rate do learners learn and retain new vocabulary from reading a graded reader? *Reading in a Foreign Language*, *15*(2), 130–163.

Webb, S. (2007). The effects of repetition on vocabulary knowledge. *Applied Linguistics*, *28*(1), 46–65. https://doi.org/10.1093/applin/aml048

Webb, S. (2021). Research investigating lexical coverage and lexical profiling: What we know, what we don't know, and what needs to be examined. *Reading in a Foreign Language*, *33*(2), 278–293.

Webb, S. (2024). Methodological features of studies of incidental vocabulary learning. In M. F. Teng & B. L. Reynolds (Eds.), *Researching incidental vocabulary learning in a second language* (pp. 53–68). Routledge.

Webb, S., & Chang, A. C.-S. (2015a). How does prior word knowledge affect vocabulary learning progress in an extensive reading program? *Studies in Second Language Acquisition*, *37*(4), 651–675. https://doi.org/10.1017/S0272263114000606

Webb, S., & Chang, A. C.-S. (2015b). Second language vocabulary learning through extensive reading with audio support: How do frequency and distribution of occurrence affect learning? *Language Teaching Research*, *19*(6), 667–686. https://doi.org/10.1177/1362168814559800

Webb, S., & Nation, I. S. P. (2017). *How vocabulary is learned* (1st ed.). Oxford University Press.

Webb, S., Pellicer-Sánchez, A., & Wang, A. (2025). Lexical coverage and reading comprehension revisited. *Reading in a Foreign Language*, *37*(1), 1–21. https://doi.org/10.64152/10125/67517

Webb, S., Sasao, Y., & Ballance, O. (2017). The updated Vocabulary Levels Test: Developing and validating two new forms of the VLT. *ITL—International Journal of Applied Linguistics*, *168*(1), 33–69. https://doi.org/10.1075/itl.168.1.02web

Webb, S., Uchihara, T., & Yanagisawa, A. (2023). How effective is second language incidental vocabulary learning? A meta-analysis. *Language Teaching*, *56*(2), 161–180. https://doi.org/10.1017/S0261444822000507

https://doi.org/10.1017/S0272263126101818 Published online by Cambridge University Press

Yamashita, J. (2022). Theory and research: Chapter 2. L2 reading comprehension. In E. H. Jeon & Y. In'nami (Eds.), *Understanding L2 proficiency: Theoretical and meta-analytic investigations* (pp. 5–28). John Benjamins Publishing Company. https://doi.org/10.1075/bpa.13.02yam

Zhang, S., & Zhang, X. (2022). The relationship between vocabulary knowledge and L2 reading/listening comprehension: A meta-analysis. *Language Teaching Research*, *26*(4), 696–725. https://doi.org/10.1177/1362168820913998

Cite this article: Oikawa, G., & Uchihara, T. (2026). The effects of contextual diversity on incidental vocabulary learning. *Studies in Second Language Acquisition,* 48: 757–783. https://doi.org/10.1017/S0272263126101818

https://doi.org/10.1017/S0272263126101818 Published online by Cambridge University Press