Pretesting effects on incidental L2 vocabulary learning through reading

Eva Puimège[a]*, Eva Caltabellotta[a], and Elke Peters[a]

[a] *Research group Language, Education, and Society, KU Leuven, Belgium*

Eva Puimège

Junior Postdoctoral Fellow of the Research Foundation – Flanders (FWO)

KU Leuven

Faculty of Arts

Sint-Jacobsmarkt 49-51

2000 Antwerp

eva.puimege@kuleuven.be

Eva Caltabellotta

KU Leuven

Faculty of Arts

Sint-Jacobsmarkt 49-51

2000 Antwerp

eva.caltabellotta@kuleuven.be

Elke Peters

KU Leuven

Faculty of Arts

Sint-Jacobsmarkt 49-51

2000 Antwerp

elke.peters@kuleuven.be

ABSTRACT

An increasing number of studies have shown that pretesting L2 word knowledge before a study phase can enhance subsequent learning. However, little is known about pretesting effects in the context of incidental L2 vocabulary acquisition. This study explores the effects of pretesting on L2 vocabulary learning through reading, focusing on the moderating effect of the pretest format. One hundred and forty-three participants were randomly assigned to a non-pretested condition or three pretested conditions (meaning recall, meaning recognition, form recognition). In the pretested conditions, participants completed a vocabulary pretest, followed by a meaning-focused reading task and three vocabulary posttests. The findings show that the meaning recall and form recognition groups were impacted most by pretesting in terms of learning outcomes and perceptions of the learning intervention. However, the pretesting effect on posttest scores was small and statistically nonsignificant, suggesting a minimal impact of pretesting on incidental learning outcomes.

BACKGROUND

In recent years, vocabulary has become widely recognized as a central element of second language acquisition. This has led to the development of a large number of vocabulary tests aimed at measuring L2 learners' vocabulary knowledge (e.g., Schmitt et al., 2020). Keeping track of vocabulary knowledge through testing can provide teachers with a rich account of their students' L2 performance and growth. Recent evidence suggests that tests can also be employed to enhance learning, as evidenced by retrieval practice effects on L2 vocabulary retention (e.g., Barcroft, 2007; Nakata et al., 2021; van den Broek et al., 2018). These studies have found that taking practice tests can help consolidate and increase the accessibility of knowledge.

A less commonly studied effect of testing is its potential to enhance learning of new information, even before that information has been studied (Pan & Carpenter, 2023). Pretesting

effects, also labelled pre-question effects, occur when learners attempt to respond to test items before a study phase. Evidence that pretesting can enhance future learning has been observed for various study materials, including word lists (e.g., Yang et al., 2017), paired associates (e.g., Grimaldi & Karpicke, 2012), and trivia facts (e.g., Richland et al., 2009). Although pretesting is a less common practice in the L2 classroom, recent findings suggest that testing before studying can likewise enhance L2 vocabulary gains (e.g., Seabrooke et al., 2019).

Pretesting effects are not only pedagogically relevant, but may also hold methodological implications for studies adopting a pretest-posttest design to measure the effect of an L2 learning intervention. The idea that pretesting may inflate L2 vocabulary gains is recognized by many SLA researchers (e.g., Durrant et al., 2022), but very little research has been done to examine its impact on learning outcomes. Inadvertent pretesting effects may be particularly impactful in the context of incidental vocabulary learning, where reported learning effects from one-off interventions are typically small and highly variable (e.g., Webb, 2020). This means that any error introduced by pretesting effects could have a strong impact on research findings. The current study explores pretesting effects on incidental vocabulary learning outcomes from L2 reading. By comparing pretested and non-pretested learning conditions in an experimental setting, our study examines the interaction between testing and subsequent word learning during meaning-focused reading.

THE PRETEST-POSTTEST DESIGN IN INCIDENTAL L2 VOCABULARY LEARNING

L2 learning activities can be considered incidental when their main focus is on meaning comprehension or communication, rather than on linguistic form (e.g., Schmitt, 2010). Examples include reading books, watching television for leisure, or attending academic lectures. During these activities, the main focus is on input comprehension, but vocabulary knowledge may be picked up from context, either subconsciously, or by temporarily shifting focus, for example in order to infer the meaning of a novel word from context (Webb, 2020). There is

now a considerable amount of evidence that L2 learners can acquire new words from activities such as tv viewing (e.g., Peters & Webb, 2018) or listening to songs (e.g., Pavia et al., 2019). By far the most studied type of incidental learning activity has been L2 reading (see Webb et al., 2023).

The vast majority of studies investigating the effects of reading on L2 vocabulary knowledge have employed a one-off intervention exposing learners to a relatively short reading passage before testing their knowledge of L2 vocabulary from the input. An important methodological element of such intervention studies is the control of pre-existing vocabulary knowledge. Prior knowledge control ensures that observed learning gains result from learners' engagement with the reading materials presented during the experiment, not from knowing the L2 words prior to the study (Durrant et al., 2022; Nation & Webb, 2011; Schmitt, 2010).

The most frequent approach to prior knowledge control in incidental L2 vocabulary learning has been the pretest-posttest design (Uchihara et al., 2019). As opposed to the use of pilot tests or pseudowords, measuring gains from pre- to posttest can track learning for partially known lexical items. This is particularly useful when researchers are interested in measuring the development of various word knowledge aspects or the development from partial to full word knowledge, rather than treating vocabulary knowledge as an all-or-nothing phenomenon (e.g., Schmitt, 2019). Pretesting may provide a fine-grained and accurate measurement of gradual vocabulary development. A potential disadvantage of using pretests is the risk of inadvertent pretesting effects. Answering test items prior to studying can improve memory for those items in a posttest, compared to a condition in which the same materials are studied without a pretest (Pan & Carpenter, 2023). In cognitive psychology, this type of testing effect has been labelled the prequestion effect, or the pretesting effect (James & Storm, 2019).

PRETESTING EFFECTS ON L2 VOCABULARY LEARNING

Pretesting effects have been reported for a wide range of study materials, including written texts (e.g., James & Storm, 2019), academic lectures (e.g., St. Hilaire & Carpenter, 2020), paired associates (e.g., Hays et al., 2013), and trivia facts (e.g., Kornell, 2014). Testing before studying tends to be more, or at least equally effective as studying the same information multiple times (e.g., Hays et al., 2013) or for a prolonged amount of time (e.g., Kornell et al., 2009; Little & Bjork, 2016). For example, Kornell et al. (2009) found that participants who made incorrect guesses on invented trivia questions (e.g., "What peace treaty ended the Calumet War?") before studying the correct responses (e.g., "Harris treaty") performed similarly on posttests as participants who had twice as much study time. Findings such as these suggest that attempting a response before studying could be an efficient way to enhance learning.

Little is known about pretesting effects in relation to L2 vocabulary learning, and only a few studies have examined these effects directly using an experimental design. The current evidence for pretesting effects on L2 vocabulary acquisition is situated in memory research and focuses mainly on deliberate learning of translation pairs and paired associates, not on incidental L2 vocabulary learning (e.g., Potts & Shanks, 2014; Seabrooke et al., 2022).

In a study by Potts and Shanks (2014), participants learned unknown Euskara words and their English (L1) translations in three conditions: (a) by guessing the word's meaning followed by corrective feedback, (b) by reading each word and its corresponding meaning, or (c) by guessing the meaning from two options, followed by feedback. On a multiple-choice posttest, participants recognized a higher number of meanings for words in the guessing (i.e., pretesting) conditions, compared to words that had been pre-read along with their translation or definition. The effect was particularly evident for words pretested at the level of meaning recall, with items pretested at the recall level yielding higher scores than multiple-choice guessing ($d = 0.42$) and

pre-reading ($d = 0.47$). The authors proposed that the generation of errors benefited learning by affecting the degree of effort or attention to corrective feedback.

In a later study by Potts, Davies and Shanks (2019), participants either attempted to retrieve the translations of L2 words immediately before or after receiving corrective feedback. The aim of this manipulation was to find out whether guessing would enhance learning either through semantic pre-activation (e.g., Carpenter, 2009), or by increasing learners' curiosity and learning efforts (e.g., Geller et al., 2017). The study replicated Potts and Shanks' (2014) finding that pre-guessing the meanings of words could enhance the effect of studying the form-meaning link with more than 4% compared to a study-only condition, whereas post-guessing (i.e., retrieval) did not (Experiment 1). Furthermore, curiosity ratings for learning the correct translation of a word were higher after incorrectly guessing the meaning, indicating that generating errors can arouse learners' curiosity to learn the correct meaning of the word.

In five experiments, Seabrooke et al. (2019) examined whether Potts and Shanks' (2014) findings would extend to other aspects of knowledge, including cued recall, two-alternative forced choice tests, and associative recognition. The study found that pretesting led to improved recognition of unfamiliar English and Euskara words, in line with Potts and Shanks' (2014) findings. The study also extended the pretesting effect to tests of multiple-choice meaning recognition, and a two-alternative forced-choice test. However, the effect did not extend to cued recall, associative recognition, or a multiple-choice test containing familiar distractors. These results suggest that pretesting effects on L2 vocabulary acquisition can depend on the word knowledge aspect that is tested.

Seabrooke et al. (2022) likewise found pretesting effects on translation recognition. Participants (English native speakers) guessed both the semantic categories and specific meanings of unknown Finnish words before receiving corrective feedback. Both correct and incorrect guesses led to higher recognition scores than pre-exposure to the correct form-

meaning pairs ($d = 0.67$). The authors proposed that making small errors might trigger learners' curiosity and motivation, leading to enhanced processing and learning of the correct word meanings.

Similar effects of guessing have been reported for learning L2 multiword expressions (e.g., Strong & Leeming, 2024). However, recent evidence suggests that pretesting can also be detrimental to learners' memory of multiword combinations, which tend to be easily confused with distractors or erroneous guesses (e.g., Li & Boers, 2024). It seems, therefore, that no single mechanism can explain all pretesting effects in L2 vocabulary learning, and that item and task variables may underlie differences in vocabulary processing and learning outcomes.

PRETEST FORMAT EFFECTS

Pretesting can lead to different learning outcomes depending on the pretested word knowledge component or pretest format. In addition to raising overall curiosity, the expectations raised by a particular test format could induce test-relevant processing of the study materials (Chan et al., 2018). For example, when learners expect a free (unprompted) recall test after studying a list of target words, they tend to score higher on both free recall of word forms and multiple-choice posttests than when they expect a multiple-choice test (e.g., Balota & Neely, 1980; Schmidt, 1988). One explanation for this is that the higher perceived difficulty of a free recall test induces stronger engagement with the tested words or their semantic associations (e.g., Leonard & Whitten, 1983; Schmidt, 1988).

Test format differences have also been found in the opposite direction. Compared to cued recall, multiple-choice tests may induce more efficient learning of competitive related information (e.g., Little & Bjork, 2016). For example, in Little and Bjork's (2016) study, participants completed a pretest on trivia facts before reading a text passage about Yellowstone National Park and the planet Saturn. Half of the participants completed multiple-choice

questions in the pretest (e.g., What geyser is thought to be the oldest in the world? a. Old Faithful, b. Steamboat Geyser, c. Castle Geyser, d. Daisy Geyser), whereas the other half received the same questions without multiple-choice options. The results showed that the multiple-choice group outperformed the cued recall group on posttest questions related to the distractors in the multiple-choice test, suggesting that processing of the multiple-choice options influenced the learners' engagement with relevant information in the reading passage.

Different pretest formats may also induce qualitatively different encoding strategies (i.e., transfer-appropriate processing), whereby the expectation of one format leads to increased performance in that format but decreased performance in another format (e.g., Blaxton, 1989; Finley & Benjamin, 2012). For example, in a study focusing on word pair associates, Finley and Benjamin (2012) manipulated learners' posttest expectations when studying a list of associated word pairs. Participants who were led to expect a cued recall test (i.e., retrieving one word based on its paired associate) spent more time and effort focusing on cue-target associations, whereas learners expecting a free recall test (i.e., listing as many of the target words as possible) focused more strongly on individual target words as well as semantic relations across different word pairs. In both groups, learning gains tended to be higher on the expected test format. Although Finley and Benjamin's study did not focus on the effects of pretesting, transfer-appropriate processing might also play a role when different pretest formats are compared.

Taken together, the findings presented above show that not knowing the answer to a question in a pretest can lead to both quantitative and qualitative changes in processing during a subsequent study phase. Being pre-exposed to target words in a pretest may direct learners' attention to those words when they are later encountered in the study phase. In addition, the process of making errors can affect learning efforts and strategies, and the nature of the test format can alter learners' engagement with unknown words.

PRETESTING EFFECTS ON INCIDENTAL L2 VOCABULARY LEARNING

Very little is known about the scale and nature of pretesting effects in the context of incidental L2 vocabulary acquisition. In a recent meta-analysis, Webb et al. (2023) examined the moderating effect of prior knowledge control on incidental L2 vocabulary acquisition. Their results showed that the pretest-posttest design was associated with larger effects compared to other methods of prior knowledge control such as pilot testing. This suggests that pretesting effects might impact the outcomes of studies examining incidental L2 vocabulary acquisition. However, few (if any) studies on incidental L2 vocabulary learning have directly controlled pretesting effects through the inclusion of a non-pretested comparison group. As a result, current estimates of pretesting effects might be confounded by variables such as item difficulty or partial item knowledge, which are likely to covary with the method of prior knowledge control.

One way of countering pretesting effects on incidental L2 vocabulary learning is by inserting a time gap of minimally one week between the pretest and the learning intervention, in order to reduce learners' memory of the pretest items before the start of the treatment (e.g., Durrant et al., 2022; Hulstijn, 2003; Swanborn & De Glopper, 1999). Another widely used strategy is the inclusion of distractor items in the pretest to reduce the salience of pretested words (e.g., Nation & Webb, 2011; Schmitt, 2010). Although many intervention studies on incidental L2 vocabulary acquisition have employed such strategies, there is very little empirical evidence for their effectiveness.

THE CURRENT STUDY

Pretesting effects have not yet been empirically examined in the context of incidental L2 vocabulary learning, and very little is still known about how pretesting might affect the learning outcomes of meaning-focused L2 activities such as reading. There is good reason to assume

that findings for L2 word - translation pairs may not extend to the context of incidental word learning through reading. Incidental learning conditions, which may involve contextual inferencing and an overall focus on meaning comprehension, are very different from studying words and their translations in isolation. Further, because a one-week delay is often inserted between the pretest and the treatment in incidental vocabulary learning experiments, the impact of pretesting effects on learning outcomes in these studies might be less outspoken than those reported in memory research (e.g., Potts & Shanks, 2014).

The present study provides an initial exploration of pretesting effects on incidental L2 vocabulary acquisition from reading. The study aims to examine the effects of pretesting on the outcomes of incidental vocabulary learning through reading. More concretely, the present study aims to estimate how strongly vocabulary posttest scores are influenced by combining a pretest and an incidental learning treatment, in comparison with a non-pretested condition, while considering the potential moderating effect of the pretest format. Further, our study aims to gain initial insight into pretesting effects on learners' perceptions of the treatment and target items. To match realistic experimental conditions, the materials, instruments, and experimental procedures were designed to reflect common practices in studies examining incidental L2 vocabulary learning. In so doing, the study aims to provide greater insight into the methodological uses of pretests in L2 vocabulary research.

RESEARCH QUESTIONS

The study addresses the following research questions:

1. To what extent does pretesting L2 vocabulary knowledge affect incidental vocabulary learning from L2 reading?
2. To what extent is the pretesting effect on incidental L2 vocabulary gains moderated by the pretest format?

3. How does pretesting in different test formats affect learners' perceptions of the treatment and expectations of a posttest?

PARTICIPANTS

One hundred and forty-three students took part in the experiment. The participants were quasi-randomly assigned to four experimental groups: a form recognition pretest group (n = 36), a meaning recall pretest group (n = 35), a meaning recognition pretest group (n = 35), and a no-pretest group (n = 37). According to the background questionnaire data (see Instruments), 120 participants identified as female, 19 as male, and 2 participants as non-binary (two non-responses). Participants were between 17-59 years old (M = 21.05 years old, SD = 4.92). The students started formal English instruction at the average age of 12 (SD = 1.83). All participants were Dutch-speaking university students (L1 = Dutch) from various undergraduate programs in Flanders, including 64 English majors. Participants' total score on the Vocabulary Levels Test (VLT, Schmitt et al., 2001) served as a proxy of their English lexical proficiency. The VLT was also used to estimate participants' lexical coverage of the reading materials (see Materials). This test was deemed appropriate for the current study because it covers a wide range of frequency levels and can therefore better differentiate between learners from higher proficiency levels than other vocabulary levels tests. Participants' scores on the VLT are summarized in Table 1. The results of an ANOVA indicated that the four groups were matched in terms of total VLT score (F[3,139] = 0.57, p = .638). Some participants completed the experiment for partial course credit (n = 45), whereas others received a monetary reward.

Table 1. Mean scores and standard deviations (in parentheses) for the Vocabulary Levels Test per experimental group.

	Form recognition (n = 36)	Meaning recall (n = 35)	Meaning recognition (n = 35)	No pretest (n = 37)
2K	28.33 (3.31)	29.37 (1.11)	28.60 (1.58)	28.46 (3.01)
3K	27.31 (4.75)	28.43 (2.72)	28.23 (2.49)	27.24 (4.43)
5K	24.36 (5.30)	25.03 (3.71)	24.51 (3.20)	25.05 (4.91)
10K	16.83 (7.89)	16.54 (6.29)	14.31 (6.15)	14.92 (6.76)
Academic	26.75 (4.84)	27.31 (3.29)	27.57 (2.55)	26.22 (4.04)
Total	123.58 (23.89)	126.87 (15.40)	123.23 (13.46)	120.89 (20.84)

MATERIALS

Twenty-four target words were used in the study. To minimize prior knowledge of the target items ahead of the learning phase, we selected low-frequency words that were unlikely to be familiar to the participants (e.g., *nurdle, aphid, osier*). The inclusion of rare words ensured minimal pre-existing item knowledge, while maintaining comparability with existing pretest-posttest studies including real words as opposed to pseudowords. The target words had an average lemma frequency of 0.42 per million occurrences in the Corpus of Contemporary American English (COCA, Davies, 2008) with a maximum frequency of 1.1 per million. All target words were concrete nouns between five and seven letters long. Each of the target words appeared three times in a reading passage of 4,524 tokens. Two versions of the passage were created, each version containing a different set of 12 target words. The remaining 12 words were only included in the pre- and posttests as control items, in order to control for practice effects and learning outside of the experiment. To control for item effects, the item sets were counterbalanced so that all words served in the experimental and control conditions.

Participants were randomly assigned to one of the two item sets, so that they were pre- and post-tested on 24 target items, but encountered only 12 of those items in the reading passage.

The reading passage was a fictional story about space travel, inspired by Douglas Adams' (1979) The Hitchhiker's Guide to the Galaxy. The first version of the passage was generated by Chat GPT (OpenAI, n.d.), based on the following prompt: "Could you please write a short story about space travelers visiting a strange planet in the style of The Hitchhiker's Guide to the Galaxy using the following words? [three target words]". The prompt was repeated until all target words had been incorporated into different story parts, each relating the characters' visit to a different planet. The resulting text passages were then substantially edited to control occurrence frequency of the target items and to increase the coherence and narrative flow of passages.

Each text passage was edited to ensure that 95% of running words belonged to the 3,000 most frequent English word families (Cobb, n.d.). However, some low-frequency words (less than 1% of tokens in the text, e.g., *voodoo, malfunction, chirp*) were retained to decrease the target words' perceptual salience, as well as to make the reading materials seem more authentic and natural. Because Flemish learners are regularly exposed to various sources of authentic English input (e.g., Peters et al., 2019), texts that are too heavily manipulated are likely to come across as unnatural, which may negatively impact learners' motivation during reading (Peters et al., 2023). The contexts of the target words were edited so that their meanings could easily be inferred. The final story, which consisted of seven parts, was piloted with 9 native speakers of English[1] and 4 Dutch-English bilinguals. The text was piloted for naturalness, level of difficulty, level of enjoyment, and contextual information. The full list of target items, both versions of the final reading passage, and a description of the piloting procedure and can be found in the OSF: https://osf.io/t47gq/.

Each section of the reading passage was followed by two true-or-false comprehension questions about the content of the story (14 questions in total). The questions served as an attention check throughout the treatment. They were piloted along with the reading passage to ensure that each question was unambiguous and could be answered without knowledge of the target words.

INSTRUMENTS

Background questionnaire

Participants completed a background questionnaire about demographic information (e.g., gender, age, education level, first language) and contact with English at university (i.e., number of hours of English class per week) and at home (e.g., TV viewing with or without subtitles, gaming, etc.). Data pertaining to the participants' amount of contact with English inside and outside the classroom were not analyzed as part of the present study.

Pretests

Three test formats commonly used in intervention studies on incidental L2 vocabulary acquisition were included as pretests in the present study: a meaning recall test, a meaning recognition test, and a form recognition test (see e.g., Webb et al., 2023). All pretests comprised 40 items in total. Along with the 24 target and control words, 16 high-frequency fillers were included to raise the number of distractors to the 1:3 ratio recommended by Schmitt (2010). Each participant was assigned to one of the pretest formats, except for the no pretest comparison group. All pretests can be accessed in the OSF: https://osf.io/t47gq/.

The meaning recall test measured the participants' ability to retrieve the correct meaning of the target words from memory, based on the isolated word form and without any additional cues. The test used a translation format in which participants were presented with the target word (e.g., *nurdle*) and were instructed to provide a Dutch translation, a definition or a

description of the word's meaning. Participants were encouraged to write down any meaning that came to mind, even if they did not know the precise meaning of the word, and to type "I don't know", "/", or "x" if they had no clue about the word's meaning.

The meaning recognition test measured the participants' ability to select the correct meaning of each target word from three options. Participants were instructed to select the meaning of the target word out of three options (the key and two distractors) and an "I don't know" option. Participants were encouraged to select a response even if they were not entirely certain of the correct meaning, and to select the "I don't know" option if they had no clue which was the correct answer. The key of each item was a Dutch translation of a high-frequency synonym of the target word, because close translations of the target words were of low frequency and therefore possibly unknown to the participants (e.g., "landengte" for *isthmus*). Distractors were Dutch words of similar frequency as the key (within 300 occurrences in the SUBTLEX NL word list, Keuleers et al., 2010). Care was taken to select distractors that did not occur in the story but were semantically plausible within the context of the story (e.g., words referring to animals or plants). None of the distractors were used more than once or appeared in the reading passage.

The form recognition test measured participants' recognition of the target word form. In this test, participants were instructed to select all and only words they had seen or heard before. The pretest did not contain any pseudoword foils, to match the length of the other test formats. Because prior knowledge control was not the primary aim of the pretests in the current study, the inclusion of pseudoword distractors was not deemed crucial to answer the research questions. Consequently, the pre- and posttest scores for this format might somewhat overestimate participants' recognition of the target words.

Posttests

All participants completed three posttests, corresponding to the three formats used in the pretest. The form recognition posttest was administered first, followed by the meaning recall test and the meaning recognition test. The same 40 target words occurred in all three tests. The instructions of the form recognition test were edited slightly in the posttest: participants were instructed to indicate which words they remembered encountering in the reading passage.

Post-reading questionnaire

Participants completed a short questionnaire enquiring about their perceptions of the reading task and their noticing of target words during reading. The questionnaire contained six statements with a 5-point Likert-scale ranging from "strongly disagree" to "strongly agree". The statements were as follows: "I enjoyed reading the story", "I found the content of the story easy to follow", "I found the language of the story easy to understand", "While reading, I mainly paid attention to the content of the story", "While reading, I paid attention to difficult words in the text", and "I expected to receive a vocabulary test after reading the text". Participants in the pretested groups received one additional statement, "I noticed that words from last week's test appeared in the story". These participants were also asked if they had looked up any of the pretested items at home, and if so, which ones.

PROCEDURE

The experiment was programmed in the Gorilla Experiment Builder (www.gorilla.sc, Anwyl-Irvine et al., 2020). Participants completed the experiment individually or in small groups on computers under supervision of the first or second author. A one-week delay was inserted between the pretest and treatment in order to reflect a common time course of intervention studies in incidental L2 vocabulary research (e.g., Horst et al., 1998; Feng & Webb, 2020; Peters, 2019). This time delay was chosen to avoid overestimating pretesting effects likely to affect learning outcomes in L2 vocabulary studies.

In the first session, participants first completed the background questionnaire, followed by the pretest corresponding to the assigned experimental condition. The pretest was announced as a measure of general English lexical proficiency. There were no time constraints on test completion. The pretest was followed immediately by the VLT, which was also completed without time constraints. After a one-week delay, participants returned to the lab to read the experimental text. They were instructed to carefully read the fictional story at their own pace. Participants were not forewarned of a vocabulary posttest, but were told that they would receive a comprehension test about the content of the reading passage. The reading passage was divided into 7 parts, each followed by two true-or-false comprehension questions. After reading, participants completed the post-reading questionnaire and the three unannounced posttests.

The order of the two sessions was reversed for participants in the no-pretest comparison group, in order to avoid any awareness-raising effects caused by taking the Vocabulary Levels Test. Participants in the no-pretest group thus completed the reading task and posttests in the first session, and the background questionnaire and VLT in the second session. All participants were debriefed both in spoken and written form at the end of the experiment. The full procedure is shown in Table 2.

Table 2. Procedure.

	Pretest (*n* = 93)	**No pretest (*n* = 33)**
Session 1	1. Background questionnaire 2. Pretest 3. VLT	1. Reading task 2. Post-reading questionnaire 3. Posttests
Session 2	4. Reading task 5. Post-reading questionnaire 6. Posttests	4. Background questionnaire 5. VLT

SCORING AND ANALYSIS

Items of the pre- and posttests were scored dichotomously (1 point per correct response). In the meaning recall test, responses were coded as 1 (“correct”) when participants could provide a description, translation, or synonym in line with the contextual information provided in the reading passage. Responses referring to a vague semantic category (e.g., “an animal” or “a plant”) were coded as 0 (“incorrect”), because some target words belonged to the same category. Pre- and posttest scores of 35 participants (24%) were scored by two independent raters, with interrater agreement of 99.2%. In the meaning recognition test, correct responses were coded as 1 (“correct”) and selecting a distractor or the “I don’t know” option was coded as 0 (“incorrect”). In the form recognition test, items indicated as known or familiar received a score of 1 and new items received a score of 0.

Comparing posttest scores of target words and test-only control words enabled us to account for guessing or learning outside the experiment. Because our main independent variable was the presence or absence of a pretest (research question 1), we could not calculate gain scores or make a direct comparison between pre- and posttest scores for each of the conditions. Instead, differences in posttest scores between target and control items served as the primary source of evidence for learning from the experimental treatment. To reduce extraneous effects of look-up behavior on the learning outcomes, 39 data points (0.003%) corresponding to items that participants had looked up at home (according to the post-reading questionnaire) were excluded from the analysis.

To examine the effect of pretesting (RQ 1) and the pretest format (RQ 2) on the learning outcomes of the experimental treatment, we conducted generalized linear mixed models with binary item-level posttest scores as the dependent variable. The analyses were conducted using the *glmer* function in the *lme4* package (Bates et al., 2015) in R (version 4.2.1, R Core Team, 2022). Three separate models were built, one for each posttest format. All models included

fixed effects for Item type (target word or control word), Pretest condition (form recognition, meaning recognition, meaning recall, no pretest), and their interaction, as well as random intercepts for items and participants, and a by-participant random slope for Item Type. To interpret the interaction between Item type and Pretest condition, we used Bonferroni-adjusted pairwise comparisons with the *emmeans* function in the *emmeans* package (version 1.4, Lenth 2018). Effect sizes were calculated following the procedure outlined by Chinn (2000). Model assumptions were checked using the *DHARMa* package (Hartig, 2022), and did not reveal any issues with collinearity or normality of the scaled residuals.

To examine the effects of pretesting on learners' perceptions of the experimental treatment (RQ 3), we analyzed the results of the post-reading questionnaire. The proportion of responses for different points on the Likert scale were counted for each statement and compared across pretesting conditions. We only report descriptive results for the questionnaire data, using dot plots generated in JASP (JASP Team, 2024).

RESULTS

PRE- AND POSTTESTS

The descriptive results of the pre- and posttests are presented in Table 3. In the meaning recall test, the average difference between pre- and posttest scores was 2.11 (*SD* = 2.14, range = 0-9). In the meaning recognition format, average gains were slightly higher: 2.97 (*SD* = 2.60, range = 0-9) new meanings were recognized in the multiple-choice test. In the form recognition test, participants could on average identify 6.08 out of 12 new target words (*SD* = 2.74, range = 1-11).

Table 3. Average percentage of correct responses on pre- and posttests and standard deviations (in parentheses) by Pretest format and Item type.

			Posttest		
Pretest format	Item type	Pretest	Form recognition	Meaning recognition	Meaning recall
Form recognition	Target	**21.53 (17.18)**	**72.22 (19.31)**	46.76 (22.91)	21.76 (18.40)
	Control	**18.75 (14.00)**	**13.19 (14.96)**	21.30 (14.70)	4.40 (5.07)
Meaning recognition	Target	**15.48 (13.43)**	64.29 (20.77)	**40.24 (24.13)**	20.24 (20.24)
	Control	**15.48 (14.02)**	11.91 (14.19)	**25.00 (22.60)**	8.57 (13.93)
Meaning recall	Target	**4.05 (5.49)**	66.19 (20.21)	47.86 (24.28)	**21.67 (20.03)**
	Control	**2.62 (5.63)**	10.00 (13.53)	25.00 (20.71)	**4.76 (9.33)**
No pretest	Target	NA	61.04 (22.05)	41.67 (20.03)	16.22 (15.71)
	Control	NA	11.26 (12.91)	26.58 (16.53)	5.41 (7.66)

Note. Pretest and posttest scores that are directly comparable due to their matching format appear in boldface.

Results of the mixed models indicated a statistically significant main effect of Item type on all outcome measures. In the form recognition posttest, the odds of obtaining a correct response were 22.45 times ($d = 1.72$, $p < .001$) higher for target words seen in the experimental texts than for test-only control words. Likewise, odds of a correct response in the meaning recognition test ($OR = 2.51$, $d = 0.51$, $p < .001$) and in the meaning recall test ($OR = 6.48$, $d = 1.03$, $p < .001$) were significantly higher for target items than for control items.

In terms of a pretesting effect, there was no significant main effect of Pretest condition in any of the models, and no significant interactions between Item type and Pretest condition in the form recognition posttest and in the meaning recall posttest. In the meaning recognition posttest, the interaction between Item type and Pretest condition reached significance ($OR = 1.94$, $d = 0.37$, $p = .026$) for the comparison between the no pretest group and the form

recognition group, and approached significance for the comparison between the no pretest group and the meaning recall group (*OR* = 1.75, *d* = 0.31, *p* = 0.064). However, pairwise comparisons with Tukey adjustment did not reveal any significant differences between pretesting conditions for target or control items. The results are summarized in Figures 1-3, and the full model output can be found in the OSF: https://osf.io/t47gq/.

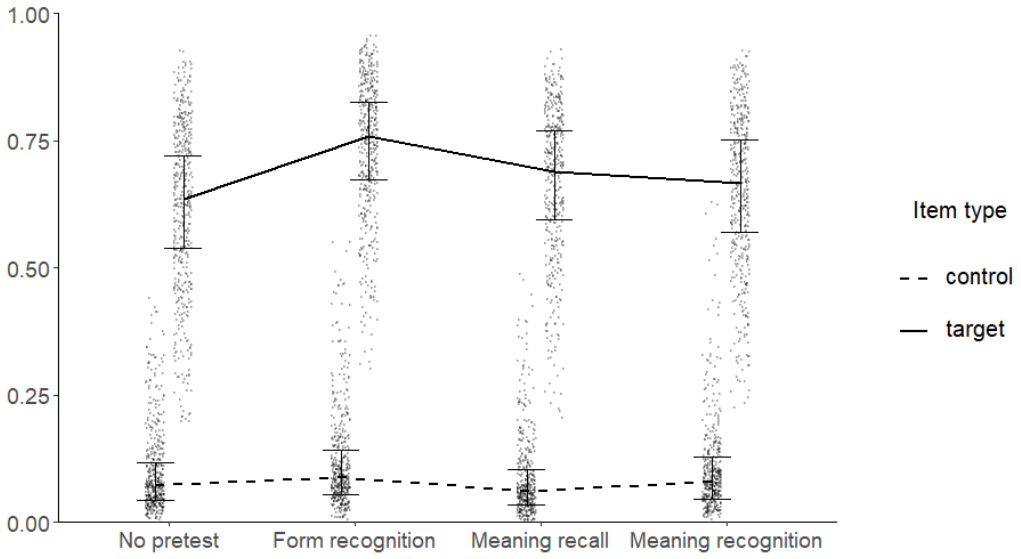


Figure 1. Predicted probabilities of a correct response in the form recognition posttest by group and item type. Error bars show 95% confidence intervals. Dots represent item variability jittered within groups.

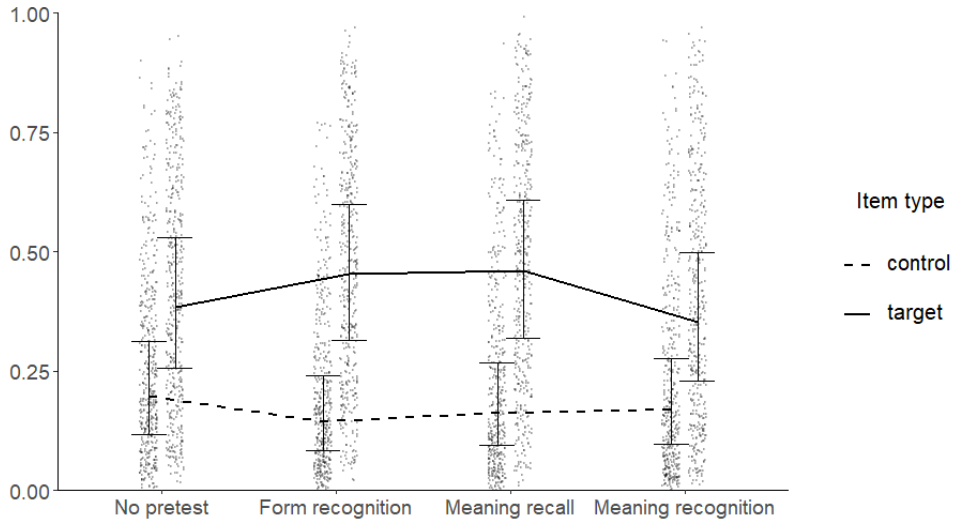


Figure 2. Predicted probabilities of a correct response in the meaning recognition posttest by group and item type. Error bars show 95% confidence intervals. Dots represent item variability jittered within groups.

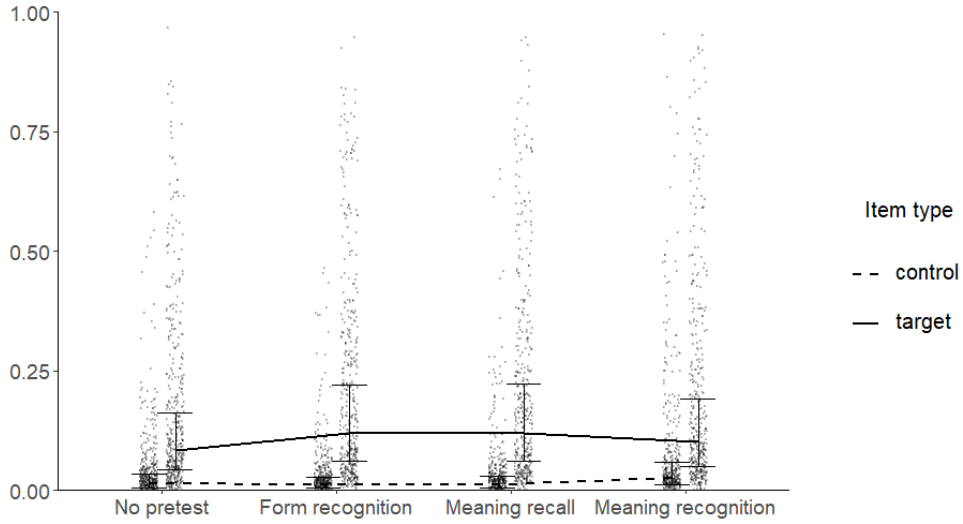


Figure 3. Predicted probabilities of a correct response in the meaning recall posttest by group and item type. Error bars show 95% confidence intervals. Dots represent item variability jittered within groups.

POST-READING QUESTIONNAIRE

The results of the post-reading questionnaire indicate that there was only minor variation between groups in terms of participants' level of enjoyment of the reading passage, and in the perceived level of difficulty of the text in terms of content and language. All participants reported a high degree of attention to the text's content as well as the difficult words in the text. Participants in all pretested groups looked up target words at home, but there were some differences between groups in terms of self-reported look-up behavior. In the form recognition and meaning recognition groups, only 11.77% and 10.71% of participants reported looking up target words at home, compared to 46.43% in the meaning recall group. Furthermore, responses to the five-point Likert items indicated that 45.71% of participants in the meaning recall group agreed (4) or strongly agreed (5) that they had recognized target words from the pretest while reading the experimental texts, compared to 30.55% in the form recognition group, and 28.56% in the meaning recognition group. Finally, 40% of participants in the meaning recall group agreed (4) or strongly agreed (5) that they had expected a vocabulary posttest, compared to 33.32% in the form recognition group, 26.70% in the meaning recognition group, and 35.12% in the no pretest group. Detailed results of the post-reading questionnaire can be found in the OSF: https://osf.io/t47gq/.

DISCUSSION

The present study set out to explore the effects of pretesting on the outcomes of incidental L2 vocabulary learning through reading. Firstly, the results show that posttest scores were significantly higher for target words encountered in the experimental reading passage than for test-only control items. This finding adds to existing evidence that significant vocabulary gains can arise from meaning-focused reading (e.g., Pellicer-Sánchez & Schmitt, 2010). The descriptive results further show that gains from pre- to posttest were generally small and highly variable across individuals, as is typical of incidental vocabulary learning (Webb, 2020). In line

with previous studies (e.g., Pellicer-Sánchez & Schmitt, 2010), gains were lower for meaning recall than for meaning and form recognition.

Regarding pretesting effects, the no pretest and meaning recognition groups had the smallest differences in posttest scores between target and control items, and the lowest posttest scores overall. However, looking at the target items, there were only minor differences in posttest scores across test formats. The descriptive results suggest that the inclusion of a form recognition or meaning recall pretest amounted to a difference of only one additional correct item out of twelve, two at the most. Further, no clear differences in learning outcomes were found between the different pretest conditions in the generalized linear mixed models. Although a significant interaction between pretesting condition and item type emerged in the meaning recognition posttest, pairwise comparisons did not reveal any significant differences between pretesting conditions. It is also noteworthy that the difference in posttest scores between target and control words was much more evident than the difference between pretest conditions. This shows that a treatment effect can clearly be discerned irrespective of the pretesting condition. The strong test-potentiated learning effects reported by studies focusing on word lists and translation pairs (e.g., Seabrooke et al., 2019), were thus not found in the current study.

Several components of our experimental design could help explain why the results do not match those of previous studies. Firstly, the aim of the present study was to closely reflect the conditions of real-life experiments on incidental L2 vocabulary learning (e.g., Feng & Webb, 2020; Peters, 2019). Unlike most studies examining pretesting effects on L2 vocabulary learning (e.g., Potts & Shanks, 2014), forced guesses were therefore not applied in any of the pretests. Because theoretical accounts of pretesting effects have emphasized the role of effortful guessing, or attempted retrieval (e.g., Kornell et al., 2009), the absence of forced guesses in the present study design might in part explain the nonsignificant pretesting effect, as well as the large amount of individual variability in both the posttest scores and the questionnaire responses.

It seems plausible that some participants took greater effort to attempt responding to unknown target words in the pretest than others, potentially leading to attempted retrieval benefits in some, but not all, learners.

Furthermore, our study procedure included a one-week interval between the pretest and the treatment. Some studies on pretest-potentiated effects have found that pretesting effects all but disappear when a delay of more than 24 hours is inserted between the pretest and study phase, depending on the nature of the studied materials (e.g., Kornell, 2014). It seems plausible, therefore, that pretesting effects would be more pronounced if the pretest were directly followed by the reading treatment. Another explanation for the small between-group differences could be related to the low frequency of the target items used in the current study. It has been argued that pretesting effects are mediated by the semantic quality or meaningfulness of the information in the pretest (Kornell, 2014). It is possible that pretesting partially known vocabulary would trigger greater semantic pre-activation or retrieval efforts than pretesting entirely new lexical items. Furthermore, rare words tend to attract attention during meaning-focused reading (e.g., Godfroid et al., 2013) and the presence of low-frequency items might cause learners to expect a vocabulary test regardless of prior tasks or instructions. The difficulty of the target items, in combination with a lab-based research setting, might also explain the highly variable responses in the post-reading questionnaire: even in the group that did not receive a pretest, 36% of participants reported that they had expected to receive a vocabulary posttest. Finally, a potential source of noise in our data set is associated with the inclusion of real word target items instead of pseudowords. Because of the one-week gap between the pretest and treatment, differences in learning gains may partly stem from participants looking up target words outside the experiment. Although we excluded items that participants remembered looking up from the regression analysis, undetected look-up behaviors might account for some

degree of variance in the data, and could thus help to explain why no clear pretesting effects were found in the present study.

Considering the small and non-significant differences between different pretest formats, no clear conclusions can be drawn regarding differences between test formats, or the cognitive explanations underlying pretest format effects. However, results of the post-reading questionnaire suggest that pretesting may have altered learners' perceptions of the target items and the learning intervention. For example, participants in the meaning recall and form recognition groups more often recognized pretested items in the experimental text than the other pretesting conditions. This may have to do with the way in which items were presented in these pretests: each test item contained only the target words and fillers, unlike the multiple-choice tests, which contained a higher amount of distracting information that might have confused learners' memory of the test items (see also Li & Boers, 2024). Previous studies have found that when distractors are interchangeable or easily confused, pretesting effects tend to be less robust (e.g., Seabrooke et al., 2019; 2022). Learners who completed the meaning recall pretest were also more prone to looking up target words at home. It is possible that learners' inability to retrieve the correct meaning of target words from memory aroused their curiosity about the words' meanings, especially for words that were already partly known or familiar (see also Seabrooke et al., 2022).

Interestingly, between-group differences in perceptions of the experimental treatment did not seem to translate into different learning outcomes. For example, even though posttest anticipation was considerably higher in the meaning recall condition (40%) than in the other conditions, this group did not obtain significantly higher posttest scores. This apparent discrepancy merits further investigation, given that vocabulary learning outcomes in incidental conditions have been widely assumed to depend on affective and cognitive variables such as motivation and attention (e.g., Laufer & Hulstijn, 2001; Yanagisawa & Webb, 2021). The

results seem to align with the findings of recent eye-tracking studies that brief pre-exposure to L2 words can impact learners' attention to those words during reading, but does not necessarily affect incidental learning outcomes (Elgort & Beyersmann, 2024; Elgort et al., 2023). Future studies could more closely examine how pretesting differentially affects learning processes and outcomes, and what role individual variation plays in explaining pretesting effects. For example, it is plausible that learners' personal goals and motivations could impact how they react to the presence of pretested vocabulary in a meaning-focused task.

Finally, results of the post-reading questionnaire revealed that learners in all groups, not just pretested groups, paid attention to L2 vocabulary while reading, and that many learners in the non-pretested group anticipated a vocabulary posttest. This finding calls into question the incidental nature of the present learning intervention, and of incidental learning experiments more generally. Researchers have argued that pretesting can alter the nature of a learning intervention by raising learners' awareness of the study aims or by enhancing their attention to L2 vocabulary (e.g., Swanborn & De Glopper, 1999). However, the current findings suggest that non-pretested conditions may likewise trigger focus on L2 word forms, perhaps in particular when target words are of low frequency and encountered in an experimental setting. In line with recent calls for a more cautious approach to defining incidental vocabulary learning (e.g., Webb, 2020; Webb et al., 2023), our findings highlight that a complete lack of intention cannot be assumed when measuring incidental vocabulary learning. More research is needed to obtain a better understanding of the cognitive processes that underlie incidental vocabulary learning, and how pretests and other methodological variables might affect those processes.

METHODOLOGICAL IMPLICATIONS

The present study was intended as a first empirical exploration of pretesting effects in incidental L2 vocabulary acquisition through reading. The small and nonsignificant differences between pretest conditions suggest that pretesting might not substantially impact incidental vocabulary

learning outcomes, in particular when target items are low-frequency words and when there is a one-week gap between the pretest and treatment. Our results also suggest that even in the non-pretested condition, participants were likely to expect a vocabulary posttest. It thus seems that administering a vocabulary pretest one week in advance may not fundamentally alter learners' expectations or perceptions of the learning intervention.

On the one hand, these findings suggest that the pretest-posttest design does not significantly affect the outcomes of incidental L2 vocabulary learning through reading, although more evidence is needed to support this conclusion. Furthermore, a relatively short interval between pretest and treatment might suffice to minimize pretesting effects on incidental L2 vocabulary learning. In other words, a gap of four weeks, which has been recommended by some researchers (e.g., Durrant et al., 2022), might not be necessary. As larger intervals provide for more opportunity to learn the target vocabulary elsewhere, limiting the size of the interval might in certain contexts be more interesting, as well as more practically feasible. However, as the current study did not manipulate the time interval between the pretest and treatment, the effect of timing remains an empirical question for future investigation.

On the other hand, our results suggest that interventions produce similar learning outcomes regardless of whether prior knowledge is controlled through pretesting. Therefore, not using a pretest seems a legitimate approach when target words are carefully selected to be unfamiliar to participants. However, this does not mean that low-frequency target items can or should serve as a substitute for pretests in all research contexts. Considering the incremental nature of L2 vocabulary acquisition (e.g., Schmitt, 2019), incidental vocabulary learning may continue long after the very first exposure to a new word, for example, when forming nuanced semantic and collocational networks based on varied contextual exposures (e.g., Webb, 2020). Pretesting may therefore still be beneficial when measuring gains for already partially known vocabulary. In addition, the questionnaire results suggest that the high salience of novel word

forms might in itself trigger expectations of a vocabulary posttest. It seems, therefore, that incorporating low-frequency words is not necessarily a superior approach in terms of lowering participants' awareness of the study aims.

LIMITATIONS AND FUTURE DIRECTIONS

The present study explored pretesting effects in a highly specific research context. Consequently, methodological and theoretical implications are limited to the setting and design examined in this experiment. There is still much to be learned about the magnitude and nature of these effects under different circumstances. Firstly, more evidence is needed to compare different time intervals between pretest and treatment. Pretesting effects are likely to be more pronounced when the delay is shorter than one week, or when a distractor activity is inserted between the pretest and the treatment. Further, the present study focused on low-frequency words, which served as an additional form of prior knowledge control to ensure a high degree of comparability in learning outcomes between the pretested and non-pretested groups. Conversely, in studies employing a pretest-treatment-posttest design, target words can belong to a variety of frequency levels, enabling researchers to track incremental gains of partially known words. Examining pretesting effects for partially known words could be an interesting avenue for future research.

An important limitation of the present study is that no a priori power analysis was conducted to estimate the required sample size. The relatively small sample size combined with the between-participants comparison implies that the null effect found for pretesting conditions needs to be interpreted with caution. Future studies could adopt a full within-participants design, which tends to have higher statistical power, in order to verify the validity of the current findings. However, such a study would have to be carefully designed to account for global attention-raising effects (e.g., raising learners' overall curiosity for L2 vocabulary), which may be more difficult to detect in a within-participants design.

Further, our study focused primarily on learning outcomes. Although the results of the post-reading questionnaire provide some initial insight into potential causes of pretesting effects, the cognitive mechanisms that drive these effects remain poorly understood. To obtain a clear understanding of what cognitive factors might explain pretesting effects, other experimental designs and methods are needed. For instance, think-aloud protocols, stimulated recall, and eye-tracking technology could be useful in determining learners' attentional processes and levels of engagement both during pretesting and reading. Finally, given the highly variable designs adopted to examine incidental L2 vocabulary acquisition, replication could be a fruitful approach to provide greater insight into pretesting effects, while keeping the experimental conditions and procedures of original studies largely intact. Future studies could replicate and extend the design of existing pretest-posttest studies through the inclusion of a no-pretest control group. In the long run, this could allow for a more large-scale systematic analysis of pretesting effects on L2 vocabulary learning.

CONCLUSION

The present study is the first to empirically examine the effects of pretesting on incidental vocabulary learning outcomes from L2 reading. The findings suggest that giving learners a vocabulary pretest one week before they encounter the pretested words in a meaning-focused reading task does not significantly affect incidental vocabulary gains, although pretesting may enhance learners' curiosity for learning target words, as evidenced by learners' responses in a post-reading questionnaire. Overall, the presence or absence of a pretest did not interact strongly with the treatment's outcomes. It thus seems that, when enough time is left between the pretest and the treatment, pretesting may present only a minor source of error in incidental L2 vocabulary acquisition through reading. Nonetheless, the widespread use of pretesting in the field underscores the need for further research into its effects on incidental L2 vocabulary acquisition.

NOTES

1. The rating task was sent out to 10 native speakers, but one rater did not complete the task and was therefore excluded.

REFERENCES

Adams, D. (1979). *The Hitchhiker's Guide to the Galaxy*. Pan Books.

Anwyl-Irvine, A. L., Massonié, J., Flitton, A., Kirkham, N. Z., & Evershed, J. K. (2020). Gorilla in our midst: an online behavioural experiment builder. *Behavior Research Methods*, *52*, 388–407.

Balota, D. A., & Neely, J. H. (1980). Test-Expectancy and Word-Frequency Effects in Recall and Recognition. In *Journal of Experimental Psychology: Human Learning and Memory* (Vol. 6, Issue 5).

Barcroft, J. (2007). Effects of opportunities for word retrieval during second language vocabulary learning. *Language Learning*, *57*(1), 35–56. https://doi.org/10.1111/j.1467-9922.2007.00398.x

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software, 67,* 1–48. https://doi.org/10.18637/jss.v067.i01

Blanca, M. J., Arnau, J., García-Castro, F. J., Alarcón, R., & Bono, R. (2023). Non-normal Data in Repeated Measures ANOVA: Impact on Type I Error and Power. *Psicothema*, *35*(1), 21–29. https://doi.org/10.7334/psicothema2022.292

Blaxton, T. A. (1989). Investigating Dissociations Among Memory Measures: Support for a Transfer-Appropriate Processing Framework. In *Journal of Experimental Psychology: Learning. Memory, and Cognition* (Vol. 15, Issue 4).

Carpenter, S. K. (2009). Cue Strength as a Moderator of the Testing Effect: The Benefits of Elaborative Retrieval. *Journal of Experimental Psychology: Learning Memory and Cognition*, *35*(6), 1563–1569. https://doi.org/10.1037/a0017021

Chinn, S. (2000). A simple method for converting an odds ratio to effect size for use in meta-analysis. *Statistics in Medicine, 19,* 3127–3131. https://doi.org/10.1002/1097-0258(20001130)19:22<3127::aid-sim784>3.0.co;2-m

Cobb, T. (n.d.). *Range for texts v.3 [computer program]. Accessed 23 April 2023 at https://www.lextutor.ca/cgi-bin/range/texts/index.pl* .

Davies, M. (2008). *The Corpus of Contemporary American English (COCA). Available online at https://www.english-corpora.org/coca/.*

Day, R., Omura, C., & Hiramatsu, M. (1991). Incidental EFL Vocabulary Learning and Reading. *Reading in a Foreign Language*, *7*(2), 541–551.

Durrant, P., Siyanova-Chanturia, A., Kremmel, B., & Sonbul, S. (2022). *Research Methods in Vocabulary Studies*. John Benjamins. https://doi.org/10.1075/rmal.2

Elgort, I., & Beyersmann, E. (2024). Do reading times predict word learning? An eye–tracking study with novel words. *Studies in Second Language Acquisition, 46*(4), 1282–1297. https://doi.org/10.1017/S0272263124000585

Elgort, I., Brysbaert, M., Stevens, M., & Van Assche, E. (2018). Contextual word learning during reading in a second language: An eye-movement study. *Studies in Second Language Acquisition*, *40*(2), 341–366. https://doi.org/10.1017/S0272263117000109

Elgort, I., Wetering, R.V.D., Arrow, T. and Beyersmann, E. (2024), Previewing novel words before reading affects their processing during reading: An eye-movement study with first

and second language readers. *Language Learning, 74*(1), 78-110. https://doi-org.helicon.vuw.ac.nz/10.1111/lang.12579

Feng, Y., & Webb, S. (2020). Learning vocabulary through reading, listening, and viewing. *Studies in Second Language Acquisition*, *42*(3), 499–523. https://doi.org/10.1017/S0272263119000494

Finley, J. R., & Benjamin, A. S. (2012). Adaptive and qualitative changes in encoding strategy with experience: Evidence from the test-expectancy paradigm. *Journal of Experimental Psychology: Learning Memory and Cognition*, *38*(3), 632–652. https://doi.org/10.1037/a0026215

Grimaldi, P. J., & Karpicke, J. D. (2012). When and why do retrieval attempts enhance subsequent encoding? *Memory and Cognition*, *40*(4), 505–513. https://doi.org/10.3758/s13421-011-0174-0

Hartig, F. (2022). DHARMa: Residual diagnostics for hierarchical (multi-level/ mixed) regression models (version 0.4.5). Retrieved from https://CRAN.R-project.org/package=DHARMa

Hays, M. J., Kornell, N., & Bjork, R. A. (2013). When and why a failed test potentiates the effectiveness of subsequent study. *Journal of Experimental Psychology: Learning Memory and Cognition*, *39*(1), 290–296. https://doi.org/10.1037/a0028468

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.

Hulstijn, J. H. (2003). Incidental and intentional learning. In C. Doughty & M. H. Long (Eds.), *The handbook of second language acquisition* (pp. 349–381). Blackwell.

James, K. K., & Storm, B. C. (2019). Beyond the pretesting effect: What happens to the information that is not pretested? *Journal of Experimental Psychology: Applied*, *25*(4), 576–587. https://doi.org/10.1037/xap0000231

JASP Team. (2024). *JASP (Version 0.18.3) [Computer software].*

Keuleers, E., Brysbaert, M., & New, B. (2010). SUBTLEX-NL: A new measure for Dutch word frequency based on film subtitles. *Behavior Research Methods*, *42*(3), 643–650. https://doi.org/10.3758/BRM.42.3.643

Kornell, N. (2014). Attempting to answer a meaningful question enhances subsequent learning even when feedback is delayed. *Journal of Experimental Psychology: Learning Memory and Cognition*, *40*(1), 106–114. https://doi.org/10.1037/a0033699

Kornell, N., Hays, M. J., & Bjork, R. A. (2009). Unsuccessful Retrieval Attempts Enhance Subsequent Learning. *Journal of Experimental Psychology: Learning Memory and Cognition*, *35*(4), 989–998. https://doi.org/10.1037/a0015729

Laufer, B., & Hulstijn, J. H. (2001). Incidental vocabulary acquisition in a second language: The construct of task-induced involvement. *Applied Linguistics, 22,* 1-26. https://doi.org/10.1093/applin/22.1.1

Lenth, R. (2018). *emmeans: estimated marginal means, aka least-squares means.* R package, available at http://CRAN.R-project.org/ package=emmeans

Leonard, J. M., & Whitten, W. B. (1983). Information Stored When Expecting Recall or Recognition. In *Journal of Experimental Psychology: Learning, Memory, and Cognition* (Vol. 9, Issue 3).

Li, A. M., & Boers, F. (2024). A multiple-choice exercise on collocations: What do learners actually remember? *International Journal of Applied Linguistics.* https://doi.org/10.1111/ijal.12666

Little, J. L., & Bjork, E. L. (2016). Multiple-choice pretesting potentiates learning of related information. *Memory and Cognition*, *44*(7), 1085–1101. https://doi.org/10.3758/s13421-016-0621-z

Nakata, T., Tada, S., Mclean, S., & Kim, Y. A. (2021). Effects of Distributed Retrieval Practice Over a Semester: Cumulative Tests as a Way to Facilitate Second Language Vocabulary Learning. *TESOL Quarterly*, *55*(1), 248–270. https://doi.org/10.1002/tesq.596

Nation, I. S. P., & Webb, S. (2011). *Research and Analyzing Vocabulary*. Heinle, Cengage Learning.

OpenAI. (n.d.). *Chat GPT. [Large language model]. Accessed on April 10 at https://chat.openai.com/chat.*

Pan, S. C., & Carpenter, S. K. (2023). Prequestioning and Pretesting Effects: a Review of Empirical Research, Theoretical Perspectives, and Implications for Educational Practice. In *Educational Psychology Review* (Vol. 35, Issue 4). Springer. https://doi.org/10.1007/s10648-023-09814-5

Pavia, N., Webb, S., & Faez, F. (2019). Incidental vocabulary learning through listening to songs. *Studies in Second Language Acquisition*, *41*(4), 745–768. https://doi.org/10.1017/S0272263119000020

R Core Team (2022). *R: A language and environment for statistical computing.* R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.

Pellicer-Sánchez, A., & Schmitt, N. (2010). *Incidental vocabulary acquisition from an authentic novel: Do Things Fall Apart? 22*(1). http://nflrc.hawaii.edu/rfl

Peters, E. (2019). The Effect of Imagery and On-Screen Text on Foreign Language Vocabulary Learning From Audiovisual Input. *TESOL Quarterly*, *53*(4), 1008–1032. https://doi.org/10.1002/tesq.531

Peters, E., & Webb, S. (2018). Incidental vocabulary acquisition through viewing L2 television and factors that affect learning. *Studies in Second Language Acquisition*, *40*(3), 551–577. https://doi.org/10.1017/S0272263117000407

Peters, E., Noreillie, A., Heylen, K., Bulté, B., & Desmet, P. (2019). The impact of instruction and out-of-school exposure to foreign language input on learners' vocabulary knowledge in two languages. *Language Learning, 69*(3), 747-782. https://doi.org/10.1111/lang.12351

Peters, E., Puimège, E., & Szudarski, P. (2023). Repetition and incidental learning of multiword units: A conceptual multisite replication study of Webb, Newton, and Chang (2013). *Language Learning, 73*(4), 1211-1251. https://doi.org/10.1111/lang.12621

Potts, R., Davies, G., & Shanks, D. R. (2019). The benefit of generating errors during learning: What is the locus of the effect? *Journal of Experimental Psychology: Learning Memory and Cognition*, *45*(6), 1023–1041. https://doi.org/10.1037/xlm0000637

Potts, R., & Shanks, D. R. (2014). The benefit of generating errors during learning. *Journal of Experimental Psychology: General*, *143*(2), 644–667. https://doi.org/10.1037/a0033194

Richland, L. E., Kornell, N., & Kao, L. S. (2009). The Pretesting Effect: Do Unsuccessful Retrieval Attempts Enhance Learning? *Journal of Experimental Psychology: Applied*, *15*(3), 243–257. https://doi.org/10.1037/a0016496

Schmidt, S. R. (1988). Test Expectancy and Individual-Item versus Relational Processing. In *Source: The American Journal of Psychology* (Vol. 101, Issue 1).

Schmitt, N. (2010). *Researching Vocabulary*. Palgrave Macmillan UK. https://doi.org/10.1057/9780230293977

Schmitt, N. (2019). Understanding vocabulary acquisition, instruction, and assessment: A research agenda. *Language Teaching*, *52*(2), 261–274. https://doi.org/10.1017/S0261444819000053

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

Schmitt, N., Schmitt, D., & Clapham, C. (2001). *Developing and exploring the behaviour of two new versions of the Vocabulary Levels Test*.

Seabrooke, T., Hollins, T. J., Kent, C., Wills, A. J., & Mitchell, C. J. (2019). Learning from failure: Errorful generation improves memory for items, not associations. *Journal of Memory and Language*, *104*, 70–82. https://doi.org/10.1016/j.jml.2018.10.001

Seabrooke, T., Mitchell, C. J., Wills, A. J., Inkster, A. B., & Hollins, T. J. (2022). The benefits of impossible tests: Assessing the role of error-correction in the pretesting effect. *Memory & Cognition*, *50*, 296–311. https://doi.org/10.3758/s13421-021-01218-6/Published

St. Hilaire, K. J., & Carpenter, S. K. (2020). Prequestions enhance learning, but only when they are remembered. *Journal of Experimental Psychology: Applied*, *26*(4), 705–716. https://doi.org/10.1037/xap0000296

Strong, B., & Leeming, P. (2024). Evaluating the application of a gap-fill exercise on the learning of phrasal verbs: Do errors help or hinder learning? *TESOL Quarterly, 58*(2), 726-750. https://doi.org/10.1002/tesq.3248

Swanborn, M. S. L., & De Glopper, K. (1999). *Incidental Word Learning While Reading: A Meta-Analysis* (Vol. 69, Issue 3).

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

van den Broek, G. S. E., Takashima, A., Segers, E., & Verhoeven, L. (2018). Contextual Richness and Word Learning: Context Enhances Comprehension but Retrieval Enhances Retention. *Language Learning*, *68*(2), 546–585. https://doi.org/10.1111/lang.12285

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

Webb, S. (2020). Incidental Vocabulary Learning. In S. Webb (Ed.), *The Routledge Handbook of Vocabulary Studies* (pp. 225–239). Routledge. https://doi.org/10.4324/9780429291586

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

Yanagisawa, A., & Webb, S. (2021). To what extent does the involvement load hypothesis predict incidental L2 vocabulary learning? A meta-analysis. *Language Learning, 71*(2), 487-536. https://doi.org/10.1111/lang.12444

Yang, C., Potts, R., & Shanks, D. R. (2017). Metacognitive unawareness of the errorful generation benefit and its effects on self-regulated learning. *Journal of Experimental Psychology: Learning Memory and Cognition*, *43*(7), 1073–1092. https://doi.org/10.1037/xlm0000363