Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

The Effects of Synonym-Embedment in Reading Passages on Vocabulary Acquisition and Retention among EFL Freshmen

Raju Dhuli

raju.dhuli@shardauniversity.uz

Sharda University Uzbekistan https://orcid.org/0000-0001-8909-5103

Research Article

Keywords: synonym-embedment, incidental-with-support learning, reading-based vocabulary learning, EFL freshmen, vocabulary retention

Posted Date: September 21st, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11082828/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: The authors declare no competing interests.

Abstract

Vocabulary knowledge underpins effective communication and academic language use, yet gains from meaning-focused reading are often limited in non-immersive EFL contexts. This study examined whether synonym-embedment (parenthetical synonym cueing within running text) improves EFL freshmen's vocabulary acquisition and two-week retention compared with reading the same passages without synonym support. Participants were 110 first-year college EFL students at a private university in Andijan, Uzbekistan, assigned to an experimental group (n = 55) and a control group (n = 55). Using a quasi-experimental pretest–posttest–delayed posttest design, target items were selected through a two-stage Vocabulary Knowledge Scale screening and assessed with a 0–100 multiple-choice meaning-recognition test covering 100 target synonyms administered at three time points. Data were analysed using a mixed-design ANOVA with Greenhouse–Geisser correction and follow-up Welch t tests under variance heterogeneity, supplemented by a retention index to quantify gain maintenance. Results showed that the experimental group outperformed the control group at both posttest and delayed posttest, retaining a substantial proportion of posttest gains after two weeks. The findings suggest that embedding unfamiliar target synonyms alongside familiar lexical anchors may facilitate vocabulary learning beyond exposure to the target synonyms alone. Implications are discussed for materials design and classroom implementation, and future research is recommended across genres, proficiency levels, and transfer outcomes beyond meaning recognition. Because semantic cues were embedded in the running text, the intervention is best interpreted as incidental-with-support rather than purely incidental learning.

1. Introduction

Vocabulary constitutes a foundational component of language competence, shaping learners' ability to communicate effectively across diverse social and academic contexts. For English as a Foreign Language (EFL) learners, especially freshmen beginning formal academic study, developing fluency depends heavily on acquiring and retaining vocabulary. Research in both L1 and L2 contexts identifies reading as an important pathway for lexical development, with extensive and repeated exposure to meaningful texts supporting incidental and context-based vocabulary growth (Chandy et al., 2024; Zhu et al., 2024). Because coherent spoken and written discourse requires flexible access to lexical knowledge, vocabulary limitations can constrain learners' ability to comprehend texts and produce meaning in academic and authentic settings (Zeng et al., 2025).

Vocabulary knowledge and reading ability are widely described as reciprocal and mutually reinforcing. Vocabulary growth supports reading comprehension, while reading activity creates opportunities for further vocabulary acquisition (Binhomran & Altalhab, 2023; Georgiou et al., 2023). Pellicer-Sánchez et al. (2024) characterize this relation as bidirectional, and empirical work consistently reports positive associations between vocabulary knowledge and reading comprehension (Lestiana et al., 2025). Importantly, this relationship is not reducible to isolated word knowledge; comprehension reflects multiple interacting pathways, including direct and reciprocal influences between lexical knowledge and reading processes (Wagner & Meros, 2010).

Despite its centrality, vocabulary learning remains challenging in many EFL contexts. Learners often struggle to expand vocabulary because of memory constraints, limited exposure, and difficulty recognizing forms and meanings in context (Alsharif, 2022). Nation (2001) argues that vocabulary learning strategies support learners at all levels, and that reliance on isolated definitional learning is typically insufficient for sustained growth. In line with this view, strategy-based training, repeated exposure, and contextualized learning are frequently associated with stronger outcomes than decontextualized memorization (Aljasir, 2025; Shahiwala et al., 2024). McKeown (2019) similarly emphasizes that effective vocabulary instruction should move beyond definitions to support flexible, context-sensitive understanding of words.

Building on evidence that contextualized exposure can enhance uptake and retention (Van den Broek et al., 2022; Eskandari et al., 2024), the present study examines a specific in-text semantic support technique: synonym-embedment within reading passages. Although previous research has established the value of lexical glossing and other forms of meaning support for vocabulary learning through reading, less is known about the effects of embedding familiar synonym cues directly within running text alongside unfamiliar target words. The present study therefore focuses specifically on parenthetical synonym cueing as an inline form of semantic support rather than treating it as an alternative to, or a direct comparison with, conventional glossing. The study examines whether learners who encounter unfamiliar target words accompanied by familiar synonyms demonstrate greater immediate vocabulary acquisition and delayed retention than learners who encounter the same target words without synonym support. In this way, the study extends the literature on meaning support in reading by examining a specific form of in-text semantic enrichment and its potential contribution to vocabulary learning and retention.

2. Literature Review

2.1 Reading-based vocabulary learning and in-text semantic support

Vocabulary learning through reading is shaped by both (a) the frequency and distribution of encounters with target items and (b) the quality of processing those encounters elicit. Reading provides sustained lexical input and supports vocabulary development through contextualized meaning control, exposure conditions, task demands, and outcome measures (Webb et al., 2023). These findings have a direct pedagogical implication: if classroom reading is a major input source in tertiary EFL contexts, then materials-design decisions that preserve a comprehension-first orientation while increasing the likelihood that word encounters become learning opportunities may meaningfully alter vocabulary outcomes. This motivates the present study's focus on synonym-embedment as an integrated, in-text semantic support intended to strengthen both immediate learning and retention over time.

2.2 Understanding synonym-embedment and nearby semantic supports

In the present study, synonym-embedment is implemented as parenthetical synonym cueing in the running text (e.g., *crowded (packed)*). Conceptually, synonym-embedment involves inserting a

semantically related synonym or near-synonym adjacent to a familiar lexical anchor, providing meaning support within the sentence stream rather than through external resources. It can therefore be treated as a type of in-text meaning support akin to minimal in-text glossing, with a distinctive design intention: delivering semantic access at the point of encounter while minimizing disruption to sentence-level processing and discourse integration. This approach differs from widely studied supports in several ways. First, many glossing and annotation approaches provide meaning outside the running text (margins, footnotes, glossaries, pop-ups), which can prompt attentional switching and foster a "lookup routine" that may fragment discourse-level integration (Boers, 2022; Teng, 2020; Yanagisawa et al., 2020). Second, textual enhancement manipulates perceptual salience (e.g., bolding, highlighting) to increase noticing, but it does not itself provide semantic content unless paired with meaning cues and may raise cognitive load when used intensively (Godfroid, 2019). Third, synonym lists and decontextualized synonym instruction may support breadth but risk shallow mapping and do not necessarily strengthen contextual integration during reading, which is central to durable learning (Webb & Nation, 2017). The present study therefore isolates a specific design space: integrated semantic support embedded directly in discourse, delivered in a minimal form (parenthetical cueing) that aims to preserve comprehension-oriented reading while increasing the probability of successful form–meaning mapping.

2.3 Theoretical mechanisms of synonym-embedment in vocabulary learning

The rationale for synonym-embedment can be grounded in complementary accounts of vocabulary learning that emphasize attention, semantic processing, and the quality of lexical representations. From a theoretical perspective, the potential benefits of parenthetical synonym cueing may be understood in terms of how additional semantic information can influence learners' processing of unfamiliar lexical items during reading. Providing a familiar synonym alongside an unfamiliar target word may make relevant semantic information more readily available and may therefore support the establishment of form–meaning connections. Such an account is broadly consistent with the Noticing Hypothesis (Schmidt, 1990), levels-of-processing theory (Craik & Lockhart, 1972), and perspectives emphasizing semantic elaboration and lexical representation in vocabulary learning (Perfetti, 2007). Importantly, these perspectives are used here as explanatory frameworks rather than as directly tested mechanisms. Because the present study measured vocabulary-test performance rather than learners' noticing, attention, inferencing, depth of processing, or lexical integration, claims about these processes should be interpreted as theoretically plausible explanations for the observed learning outcomes rather than as empirically demonstrated mechanisms.

The following subsections therefore consider three theoretically plausible pathways through which synonym cueing may support vocabulary learning: attention to the target form, access to meaning and semantic elaboration, and connections between new lexical forms and existing lexical knowledge.

2.3.1 Noticing and attention allocation

Noticing-based accounts propose that attention to linguistic form during meaning-focused activity increases the likelihood of acquisition (Schmidt, 1990). During reading, learners often prioritize global comprehension, and unfamiliar items may be skipped, misinferred, or processed too shallowly to support stable learning. A consistent cueing format may make a novel lexical item more salient while simultaneously providing information relevant to its interpretation. In this respect, parenthetical synonym cueing differs from salience-only forms of input enhancement because it combines the target form with an immediate semantic cue. The present study does not directly measure learners' noticing or attentional allocation; rather, these concepts provide a theoretically plausible explanation for why the additional semantic information accompanying target words might contribute to vocabulary learning.

2.3.2 Meaning access, inferencing support, and semantic elaboration

Vocabulary learning from reading may be strengthened when learners can successfully connect an unfamiliar lexical form with its meaning during an encounter and integrate that information into the developing discourse representation (Van den Broek et al., 2022). From a levels-of-processing perspective, meaning-based processing may produce more durable memory traces than superficial exposure (Craik & Lockhart, 1972). Parenthetical synonym cueing may reduce uncertainty surrounding the meaning of an unfamiliar item by providing a familiar semantic equivalent within the immediate reading context. The availability of this additional semantic information could make form–meaning mapping more accessible and may provide an opportunity for greater semantic elaboration. This interpretation is consistent with work emphasizing that processing quality, rather than exposure alone, may contribute to durable vocabulary learning (Webb & Nation, 2017). However, because the present study did not directly assess inferencing, depth of processing, or semantic elaboration, these processes should be regarded as possible explanatory mechanisms rather than outcomes demonstrated by the study.

2.3.3 Lexical quality hypothesis and semantic network integration

The Lexical Quality Hypothesis holds that skilled comprehension depends on well-specified lexical representations characterized by precise connections between lexical form and meaning (Perfetti, 2007). When an unfamiliar lexical form is encountered alongside a familiar semantic equivalent, the learner may have an opportunity to connect the new form with an existing lexical representation rather than processing it as an isolated item. This network-based perspective is compatible with accounts of the mental lexicon that emphasize associative organization and structured semantic relations (Agustín-Llach & Rubio, 2024). From this perspective, parenthetical synonym cueing may provide an additional semantic connection that could facilitate subsequent access to the target word. Such an account offers a possible explanation for differences in delayed vocabulary performance, although the present study did not directly measure lexical quality, semantic-network integration, or retrieval processes (Read, 2004; Chaokongjakra, 2023).

Taken together, these theoretical accounts suggest that synonym cueing could influence vocabulary learning through multiple, potentially related pathways: increased attention to the target form, more

accessible meaning information during reading, and stronger connections between unfamiliar forms and existing lexical knowledge. The present study tests the resulting vocabulary-learning outcomes rather than these underlying processes directly.

2.3.4 Positioning synonym-embedment among alternative supports

Synonym-embedment shares with lexical glossing the general function of providing learners with immediate semantic support during reading. However, the present study does not directly compare synonym-embedment with conventional glossing and therefore cannot determine whether one form of support is more effective than the other. Previous research suggests that glossing and other forms of lexical annotation can facilitate L2 vocabulary learning, although their effects may vary according to gloss type, learner characteristics, exposure conditions, and outcome measures (Boers, 2022; Yanagisawa et al., 2020; Teng, 2020). The present study instead focuses on parenthetical synonym cueing as a specific form of inline semantic support and examines its association with vocabulary acquisition and delayed retention relative to exposure to target words without synonym support. This distinction defines the scope of the present investigation and avoids treating synonym-embedment as a direct alternative to, or superior form of, conventional glossing.

2.4 Limitations of synonym-embedment in vocabulary learning

2.4.1 Reduced desirable difficulty

Immediate semantic access can lower inferencing effort. Desirable-difficulties perspectives suggest that reducing effort can sometimes weaken delayed learning if learners confuse fluency with durable knowledge (Bjork & Bjork, 2011). This raises an empirical question central to the present design: does synonym-embedment mainly inflate short-term performance, or does it support retention? A delayed posttest is therefore essential for interpretation.

2.4.1 Fuzziness and near-synonym constraints

Synonyms are rarely perfectly interchangeable, and near-synonyms may differ in collocation, register, pragmatic constraints, and semantic scope (Read, 2004). If cueing introduces an imprecise mapping, learners may develop representations that support recognition but not accurate use. This boundary condition also matters for measurement: meaning-recognition outcomes capture form–meaning mapping but do not fully address usage precision.

2.4.2 Semantic interference

Semantic clustering research suggests that learning related items together can increase confusion and competition during acquisition (Tinkham, 1993). Although synonym-embedment is not list-based clustering, it places related forms in close proximity. This risk can be mitigated through careful selection (high contextual fit), controlled density, and coherent discourse anchoring rather than abstract presentation.

2.4.3 Cognitive load and density

Cognitive load theory emphasizes managing working memory demands and minimizing split-attention effects (Sweller, 1988). Synonym-embedment may reduce split attention relative to external glosses because the cue is integrated locally; however, excessive cue density or awkward embedding can still impose extraneous load. This supports design controls such as limiting targets per passage and using graded-reader style texts where discourse complexity and lexical load are manageable (Hill, 1977). These principles align directly with the present study's methodology, including controlled density and expert review of plausibility.

2.5 State of evidence and the research gap

The broader evidence base supports reading-based vocabulary learning when learning conditions promote meaning access and repeated encounters (Webb & Nation, 2017; Boers, 2022; Yanagisawa et al., 2020). Exposure frequency and recycling contribute meaningfully to retention trajectories and graded-reader research suggests that without sufficient reinforcement, vocabulary decay is common (Webb, 2017). Meta-analytic syntheses further indicate that incidental learning outcomes are sensitive to methodological decisions, including control of prior knowledge, test format, and learning conditions (Webb et al., 2023).

Despite this strong foundation, there is comparatively less classroom evidence that isolates parenthetical synonym cueing as a systematic, integrated in-text semantic support technique, especially under conditions that strengthen inference quality and interpretability: (a) controlled prior knowledge through screening, (b) tertiary EFL learners in non-immersive contexts, and (c) delayed retention testing to separate short-lived gains from consolidated learning. This gap matters because many tertiary EFL programs rely heavily on print materials and classroom reading, and instructors need low-cost design options that preserve comprehension-first reading while improving learning conditions. The present study addresses this gap by implementing synonym-embedment in graded-reader-derived passages with controlled density, screened unfamiliar targets, repeated classroom exposure, and a delayed posttest to evaluate retention.

2.6 Conceptual model and research questions

2.6 Conceptual model and research questions

The present design aligns with the framework in three ways. First, because synonym-embedment is theoretically expected to influence processing at the point of encounter by providing additional semantic information, vocabulary outcomes are assessed immediately after the instructional sequence. Second, because theoretical perspectives offer different possibilities concerning the relationship between processing demands and durability of learning, the design includes a delayed posttest to examine whether immediate learning is maintained over time. In particular, desirable-difficulty perspectives suggest that reducing inferential effort may not necessarily lead to durable learning (Bjork & Bjork, 2011),

whereas meaning-based processing perspectives suggest that additional semantic information may support subsequent retention (Craik & Lockhart, 1972). Third, because prior knowledge can influence vocabulary learning from reading, target items are screened to establish baseline unfamiliarity and thereby provide a clearer assessment of changes associated with the instructional condition (Wesche & Paribakht, 1996; Webb et al., 2023). These features position the intervention as incidental-with-support: vocabulary is encountered during comprehension-oriented reading, while semantic cues are intentionally embedded to modify the conditions under which target words are processed.

Accordingly, the study addresses two research questions:

1. **RQ1: Immediate vocabulary acquisition.** Does parenthetical synonym cueing enhance learners' immediate vocabulary acquisition compared with exposure to target words without synonym support?
2. **RQ2: Delayed vocabulary retention.** Does parenthetical synonym cueing enhance learners' delayed vocabulary retention compared with exposure to target words without synonym support?

Based on theoretical accounts emphasizing attention to linguistic form, meaning-based processing, semantic elaboration, and lexical representation (Schmidt, 1990; Craik & Lockhart, 1972; Perfetti, 2007), it was expected that learners receiving parenthetical synonym cueing would demonstrate greater immediate vocabulary gains and stronger delayed-posttest performance than learners in the control condition. These theoretical accounts are treated as possible explanations for the expected learning outcomes; the underlying cognitive processes were not directly measured in the present study.

2.7 Transition to the present study

To examine these research questions in a classroom-realistic tertiary EFL context, the present study employed a quasi-experimental pretest–posttest–delayed posttest design. Parenthetical synonym cueing was implemented in reading passages, and vocabulary learning was assessed before the intervention, immediately after it, and two weeks later. The following section describes the participants, materials, procedures, and analytical approach used to investigate the effects of synonym cueing on immediate vocabulary acquisition and delayed retention.

3. Methodology

3.1 Research design

The study employed a quasi-experimental pretest–posttest–delayed posttest design with two instructional conditions. The experimental group (EG) read passages containing synonym-embedment implemented as parenthetical synonym cueing (e.g., *crowded (packed)*), whereas the control group (CG) read the same passages without synonym support. Vocabulary performance was measured at three time points: pretest (Time 1), immediate posttest (Time 2), and delayed posttest (Time 3) administered

two weeks after Time 2. This design was used to examine immediate vocabulary acquisition (RQ1) and longer-term retention (RQ2) under reading-based incidental-with-support conditions.

3.2 Participants and setting

Participants were 110 first-year undergraduate EFL students (18–19 years old) enrolled in a BA Applied English program (academic year 2025–2026) at a private university in Andijan, Uzbekistan. Students' primary language was Uzbek, and English served as the medium of instruction and academic communication. To reduce proficiency-related variance, official IELTS certification records were used for screening. From an initial pool of 160 students, 130 had valid IELTS results corresponding to CEFR levels B1–C1. To ensure a relatively homogeneous proficiency level, only students classified within the B1–B2 range were retained for the study, resulting in a final sample of 110 participants. The participants were taught in intact classes and assigned to two equal groups: an experimental group (EG; n = 55) and a control group (CG; n = 55). Because individual random assignment was not feasible, the design is quasi-experimental. Two intact course sections were designated as experimental and control conditions due to timetable constraints. Both groups followed the same syllabus sequence, class duration, and assessment schedule; the only planned instructional difference was the presence versus absence of parenthetical synonym cueing in the reading passages. Because intact classes were used, observations may not be fully independent (students were nested within classes). Analyses therefore treated individuals as the unit of analysis, and findings should be interpreted with this clustering limitation in mind. Future work should replicate the design with multiple classes per condition (cluster randomization) or apply multilevel models when the number of clusters is sufficient.

3.3 Instructional materials and treatment texts

The intervention was conducted over a five-week instructional period consisting of 10 treatment sessions. Each session lasted approximately 60 minutes and focused on one short reading passage adapted from Hill (1977), resulting in a total of 10 passages used during the study. The 10 reading passages contained an average of 260 words per passage. Each passage included 10 target vocabulary items, resulting in 100 target words across the intervention. In terms of readability, the passages had a mean Flesch Reading Ease score of 77.2, indicating that the texts were generally easy and accessible to read. Overall, the passages were estimated to fall within the A2–B1 CEFR range, providing an appropriate level of textual accessibility for the participants while allowing selected higher-level lexical items to be introduced as target vocabulary. The 100 target words represented mid-frequency vocabulary, whereas the 100 anchor-word/synonym cues were generally accessible and mid-frequency. Thus, the anchor words were intended to provide relatively accessible semantic support while the target words supplied the primary vocabulary-learning challenge.

Each passage was administered in a separate instructional session to ensure controlled exposure to the target vocabulary items. Instructional materials consisted of 10 short reading passages adapted from Hill (1977) graded readers, selected for controlled linguistic input appropriate for tertiary EFL learners. Target items were content words (function words excluded).

In the experimental group (EG), the instructional design followed a contextual synonym-embedment approach, in which unfamiliar synonym targets were introduced through familiar lexical anchor words already known to participants. Specifically, participants were first confirmed to know high-frequency lexical items (e.g., crowded, worried, shouted) (see **Appendix-A**). These known words were then used as semantic anchors to introduce less familiar, mid-frequency synonym targets embedded parenthetically in the running text (e.g., crowded (packed), worried (perturbed), shouted (vociferated)) (see **Appendix-A**). This design ensured that learners encountered unfamiliar lexical items in a semantically supported context, facilitating noticing, inferencing, and form–meaning mapping during reading without interrupting comprehension flow.

In the control group (CG), the same target synonym was retained in the corresponding sentence but appeared without the familiar anchor word or parenthetical synonym cue (see **Appendix-B**). For example, the control version read, “The shops and streets were terribly packed.” Thus, both groups encountered and were assessed on the same target synonym items. The critical difference between conditions was the manner in which the target synonym was presented: the EG received the target synonym together with a familiar semantic anchor in parenthetical form, whereas the CG encountered the target synonym without the embedded synonym cue. This design was intended to isolate the added contribution of synonym-embedment rather than simple exposure to the target vocabulary.

To support content validity and instructional suitability, all adapted passages were reviewed by an Applied Linguistics expert for vocabulary appropriateness, clarity, and grammatical and semantic plausibility of synonym-embedment. Minor revisions were made based on feedback. Because the synonym cues were intentionally embedded to support vocabulary learning within reading, the intervention is conceptualized as reading-based learning with in-text semantic support (incidental-with-support learning) rather than purely incidental exposure.

3.4 Instruments

Vocabulary knowledge scale (VKS) screening. To ensure that synonym targets were unfamiliar at baseline and that anchor words were already known, a two-stage screening procedure was conducted using the Vocabulary Knowledge Scale (VKS; Wesche & Paribakht, 1996), complemented by instructional familiarity checks.

1. **Stage 1 (Anchor-word confirmation):**

A pool of 150 candidate content words drawn from the reading passages was screened to identify lexical items already known to participants. Words rated at VKS Level 3–5 (known meaning or used correctly) were retained as familiar anchor words. Approximately 100 items were confirmed as known and selected for use as semantic anchors in the treatment passages.

2. **Stage 2 (Synonym target selection):**

For each confirmed anchor word, semantically related synonym candidates were identified. These synonym candidates were then screened to determine unfamiliarity among participants. Words rated at VKS Level 1–2 (never encountered or seen but meaning unknown) were treated as unfamiliar lexical items. From this process, 100 synonym items were selected as final target vocabulary items for the intervention.

In this study, therefore, a clear distinction was maintained between:

- familiar anchor words (e.g., crowded, worried, shouted), and
- unfamiliar synonym targets (e.g., packed, perturbed, vociferated).

Target synonyms were embedded in the passages only after confirming semantic compatibility, part-of-speech alignment, and contextual fit. Because VKS relies on self-reported knowledge, the screening was treated as a practical classroom-oriented measure of prior familiarity. Future research may strengthen validity by triangulating VKS results with objective pretests or corpus-informed frequency thresholds.

3.4.1 Development, Validation, and Reliability of the Vocabulary Test

Receptive knowledge of the 100 target synonyms was assessed using a researcher-constructed multiple-choice meaning-recognition test (see **Appendix-C**) administered at three measurement points (Time 1, Time 2, and Time 3). Before item construction, a test blueprint in Table 1 was developed to ensure systematic alignment among the target lexical items, the intended meanings conveyed in the reading materials, and the objective of assessing receptive meaning recognition. Each of the 100 target synonyms was represented by one four-option multiple-choice item, yielding a 100-item test. Each item contained one correct meaning and three plausible distractors. The items assessed recognition of word meaning rather than productive knowledge, as receptive vocabulary knowledge provides an important foundation for the subsequent development and use of productive vocabulary knowledge.

Table 1. Test Blueprint for the Vocabulary Meaning-Recognition Test

Component	Specification	No. of items	Percentage
Target lexical items	100 target synonyms selected for the intervention	100	100%
Meaning recognition	Recognition of the intended meaning of each target synonym	100	100%
Item format	Four-option multiple-choice items	100	100%
Correct response	One correct meaning per item	100	—
Distractors	Three plausible incorrect alternatives per item	300	—
Total	Meaning-recognition items	100	100%

Following item construction, the test was reviewed for content validity and item quality. The items were checked against the blueprint to ensure that all target synonyms were represented and that each item measured the intended meaning. Definitions were based on the meanings conveyed by the target words in the reading passages. The wording of the items and alternatives was also reviewed to minimize ambiguity, and the correct answers and distractors were written in grammatically parallel forms and, where possible, comparable in length and complexity. The correct-answer positions were varied across the test to reduce response-position bias. The preliminary test was then subjected to item analysis using pilot data.

Following item construction, the quality of the vocabulary test was examined through item-difficulty, item-discrimination, and distractor analyses. Item difficulty was assessed using the proportion-correct index (p), calculated as the number of participants who answered an item correctly divided by the total number of participants responding to that item (Crocker & Algina, 1986; Rezigalla et al., 2024). Higher *p* values indicated easier items, whereas lower values indicated more difficult items. Item discrimination was examined using the point-biserial correlation coefficient (rpb), which represents the relationship between performance on an individual dichotomously scored item and participants' total test scores (Ebel & Frisbie, 1991; Rezigalla et al., 2024). Items with weak or negative discrimination coefficients were reviewed for possible ambiguity, inappropriate distractors, excessive difficulty or easiness, or answer-keying problems. Distractor analysis was conducted to determine whether the three incorrect alternatives functioned as plausible distractors. Following established multiple-choice item-analysis procedures, a distractor selected by fewer than 5% of participants was considered potentially non-functioning and was reviewed for possible revision (Haladyna et al., 2002; Rezigalla et al., 2024). Items and distractors demonstrating problematic performance were reviewed and revised before the final test was administered.

The internal consistency reliability of the 100-item vocabulary test was examined using the Kuder–Richardson Formula 20 (KR-20), which was appropriate because the items were scored dichotomously (0 = incorrect, 1 = correct). KR-20 was used to estimate the extent to which the items consistently measured learners' receptive knowledge of the target synonyms. The test demonstrated high internal consistency, with KR-20 coefficients of .92 in the pilot administration and .91 at the main-study pretest. These coefficients provided evidence of satisfactory internal consistency for the test items.

Vocabulary knowledge was assessed at three measurement points (Time 1, Time 2, and Time 3). One point was awarded for each correct response and zero points for each incorrect response, resulting in total scores ranging from 0 to 100. Higher scores indicated greater receptive knowledge of the target synonyms. To reduce potential item-specific practice effects across the three administrations, parallel forms of the assessment were used, with the wording of the questions modified across administrations while retaining the same target vocabulary and intended meanings.

3.5 Operationalization of word learning

In this study, the primary dependent variable—word learning—was operationalized as learners' receptive meaning recognition of target synonym items introduced during the intervention. Word learning was defined as the extent to which participants could correctly identify the meaning of previously unfamiliar synonym targets encountered in the reading passages. This construct was measured using a researcher-developed multiple-choice vocabulary test administered at three time points (pretest, posttest, and delayed posttest). Each correctly identified item was scored as 1 and incorrect responses as 0, producing a cumulative score ranging from 0 to 100. Therefore, higher scores indicated greater acquisition and retention of the target synonym vocabulary. This operationalization reflects vocabulary learning as form–meaning mapping accuracy under recognition conditions, consistent with receptive vocabulary knowledge frameworks in second language acquisition research.

3.6 Pilot testing

A pilot study with 40 students was conducted to evaluate clarity of instructions, suitability of passage difficulty, and item comprehensibility. Minor adjustments were made to improve readability and testing clarity prior to the main study.

3.7 Data collection procedure

To reduce cross-condition contamination, the two sections met at different times and were asked not to share materials during the intervention period. No synonym-embedded materials were distributed to the control group.

1. **Pretest (Time 1):** All participants completed the vocabulary test to establish baseline comparability.
2. **Intervention phase:**Instruction occurred over six weeks, with three 60-minute sessions per week (10 sessions total). Ten sessions were devoted to the treatment passages (one passage per session). The remaining sessions followed the regular course syllabus and were comparable across groups in time-on-task and instructional content.
    - **EG condition:** Students read synonym-embedded passages containing 10 parenthetically cued target synonyms per passage. Instruction emphasized comprehension and exposure; participants were not instructed to memorize target items, and vocabulary learning was intended to occur during reading with in-text semantic support rather than through explicit vocabulary teaching.
    - **CG condition:** Students read the same passages without synonym support under the same schedule and time constraints. Tasks emphasized comprehension-first reading, and participants were not given explicit memorization instructions.
3. **Immediate posttest (Time 2):** After completion of the treatment sequence, participants completed the posttest.
4. **Delayed posttest (Time 3):** Two weeks after Time 2, participants completed the delayed posttest to assess retention.

3.8 Participant instructions

At the outset of the study, participants in both experimental and control groups were informed that they would engage in regular classroom reading activities followed by comprehension-based tasks. They were instructed to focus on understanding the overall meaning of the passages rather than intentionally memorizing vocabulary items. No explicit instruction for rote vocabulary learning was provided, in order to preserve the reading-based learning condition.

During each instructional session, participants read one adapted passage under teacher supervision. In the experimental group, unfamiliar synonym targets were embedded parenthetically with the familiar anchor words within the text (e.g., crowded (packed)), whereas the control group read identical passages without synonym support. After reading each passage, both the experimental and control groups completed identical comprehension-oriented tasks, including answering comprehension questions and engaging in guided discussion of the passage content. These post-reading activities were kept constant across conditions to ensure that any differences in vocabulary learning outcomes could be attributed solely to the presence or absence of synonym-embedment in the reading texts, rather than differences in task type or instructional exposure. The intervention consisted of 10 instructional sessions over a five-week period. Each session lasted approximately 60 minutes. Ten sessions included treatment passages containing synonym targets, with each passage introducing 10 target items.

Vocabulary assessments were administered at three stages: a pretest before the intervention, an immediate posttest after completion of the treatment phase, and a delayed posttest two weeks later. In each test, participants selected the correct meaning of the target synonym from four options. No feedback on test performance was provided during the study period.

3.9 Data analysis procedure

All analyses were conducted using IBM SPSS Statistics (Version 28). Alpha was set at .05 and two-tailed probabilities were reported unless stated otherwise. Assumptions were evaluated for planned parametric tests, including normality (Shapiro–Wilk and Q–Q plots), homogeneity of variance (Levene's test), and sphericity (Mauchly's test). When assumptions were violated, robust procedures were applied (Welch's t for unequal variances; Greenhouse–Geisser correction for violations of sphericity). Because Box's test indicated unequal covariance matrices for the repeated-measures model, multivariate tests (Pillai's Trace) were also examined as a robustness check alongside the univariate mixed ANOVA. Table 1 summarizes the primary analysis plan aligned to RQ1 (posttest acquisition) and RQ2 (delayed retention), including planned comparisons, assumptions, and effect-size reporting. Because the design includes repeated measures across three time points, the overall Group × Time pattern was also evaluated using a mixed ANOVA framework (Table 3). To complement inferential tests, a retention index was computed to express maintenance of learning gains relative to baseline (Table 2); this index is reported alongside raw means to clarify retention as a proportion of learning rather than only an absolute score difference.

TABLE 2 Statistical Analysis Plan for Primary Research Questions

Research Question	Dependent Variable	Comparison	Statistical Test	Assumption Checks	Effect Size & Reporting
RQ1: Immediate acquisition	Vocabulary score (posttest)	EG vs. CG at posttest	Independent-samples t test (Welch when variances unequal)	Independence; outliers (boxplots); normality (Shapiro-Wilk; Q-Q); variance equality (Levene)	Cohen's d; 95% CI for mean difference
RQ1: Immediate acquisition	Vocabulary score	Pretest vs. posttest within EG	Paired-samples t test	Outliers in change scores; normality of difference scores (Shapiro-Wilk; Q-Q)	Cohen's d (paired); 95% CI for mean change
RQ1: Immediate acquisition	Vocabulary score	Pretest vs. posttest within CG	Paired-samples t test	Outliers in change scores; normality of difference scores (Shapiro-Wilk; Q-Q)	Cohen's d (paired); 95% CI for mean change
RQ2: Delayed retention	Vocabulary score (delayed posttest)	EG vs. CG at delayed posttest	Independent-samples t test (Welch when variances unequal)	Independence; outliers (boxplots); normality (Shapiro-Wilk; Q-Q); variance equality (Levene)	Cohen's d; 95% CI for mean difference
RQ2: Delayed retention	Vocabulary score	Posttest vs. delayed posttest within EG	Paired-samples t test	Outliers in change scores; normality of difference scores (Shapiro-Wilk; Q-Q)	Cohen's d (paired); 95% CI for mean change
RQ2: Delayed retention	Vocabulary score	Posttest vs. delayed posttest within CG	Paired-samples t test	Outliers in change scores; normality of difference scores (Shapiro-Wilk; Q-Q)	Cohen's d (paired); 95% CI for mean change
RQ2: Delayed retention	Vocabulary score	Pretest vs. delayed posttest within EG and CG (baseline maintenance)	Paired-samples t test	Outliers in change scores; normality of difference scores (Shapiro-Wilk; Q-Q)	Cohen's d (paired); 95% CI for mean change

Note. EG = experimental group; CG = control group. Unless stated otherwise, α= .05 (two-tailed). Bonferroni corrections are applied to defined families of pairwise comparisons, and 95% confidence intervals accompany all mean comparisons.

TABLE 3 Computation of Retention Index for Research Question 2

Construct	Purpose	Formula	Interpretation
Retention index (gain maintenance)	Quantifies the proportion of immediate learning maintained after a delay	Retention (%) = [(M_delayed - M_pre) / (M_post - M_pre)] x 100	Higher values indicate greater maintenance of learning gains. 100% = full maintenance of posttest gains; 0% = no maintenance above baseline; negative values indicate performance below baseline.

Note. The retention index is computed using group means (and may also be computed at the participant level when individual scores are available). If M_post equals M_pre, the index is undefined and should not be computed.

TABLE 4 Secondary Analysis: Mixed ANOVA Framework

Factor/Effect	Type	Levels	Hypothesis Tested	Effect Size	Notes
Group	Between-subjects	2 (EG, CG)	Main effect of intervention condition	ηp^2	Homogeneity of variances evaluated (Levene) at each time point.
Time	Within-subjects	3 (Pretest, Posttest, Delayed)	Main effect of assessment time	ηp^2	Sphericity evaluated (Mauchly); Greenhouse-Geisser correction applied if violated.
Group x Time	Interaction	2 x 3	Differential change patterns over time between groups	ηp^2	If significant, follow-up simple effects and Bonferroni-adjusted pairwise comparisons are conducted.

Note. Follow-up comparisons are Bonferroni-adjusted. Report F values with degrees of freedom (corrected when applicable), p values, ηp^2, and descriptive statistics (n, M, SD) by group and time.

3.10 Ethical considerations

This study involved standard classroom reading activities and assessments. Participation was voluntary, and all participants provided informed consent prior to the study. Data were treated confidentially and reported in aggregate form. As the research involved routine educational activities with no foreseeable risk to participants, formal ethics committee approval was not required.

4. Results

4.1 Data screening and assumption checks

All cases were retained for analysis (N = 110; EG = 55, CG = 55), with no missing values at Time 1, Time 2, or Time 3. Normality was examined within each group at each time point. Shapiro–Wilk tests indicated departures from normality at several time points (Table 4). Homogeneity of variance was assessed at each time point using Levene's test (Table 5), indicating variance heterogeneity between groups, particularly at posttest and delayed posttest. For the repeated-measures factor, Mauchly's test indicated that sphericity was violated (W = .902, $\chi^2(2) = 11.090$, p = .004); therefore, Greenhouse–Geisser-corrected results are reported for within-subject effects. Box's test of equality of covariance matrices was also significant (Box's M = 115.625, F = 18.690, p < .001); accordingly, multivariate tests (Pillai's Trace) were inspected as a robustness check alongside the univariate mixed ANOVA results (Table 7). Given these diagnostics, Welch's t tests were used for between-group comparisons under unequal variances, and Greenhouse–Geisser corrections were applied for within-subject effects where relevant. Both groups followed the same timetable, reading schedule, and test administration conditions; the only instructional difference was the presence versus absence of parenthetical synonym cueing in the passages.

Table 5 presents Shapiro–Wilk tests of normality by group and time point, and Table 5 reports Levene's tests of equality of variances by time point.

Table 5
Shapiro-Wilk Tests of Normality by Group and Time Point

Time point	Group	df	W	p
Pretest (Time 1)	CG	55	.941	.010
Pretest (Time 1)	EG	55	.952	.029
Posttest (Time 2)	CG	55	.865	< .001
Posttest (Time 2)	EG	55	.919	.001
Delayed posttest (Time 3)	CG	55	.957	.048
Delayed posttest (Time 3)	EG	55	.912	.001

Table 6
Levene's Tests of Equality of Variances by Time Point

Time point	Levene statistic	df1	df2	p
Pretest (Time 1)	5.228	1	108	.024
Posttest (Time 2)	74.547	1	108	< .001
Delayed posttest (Time 3)	63.441	1	108	< .001

Descriptive statistics

Table 6 presents means and standard deviations for both groups across all time points. Pretest means were nearly identical, while posttest and delayed posttest means showed very large separations favoring the experimental group. Given the size of the gains, score distributions were inspected to evaluate potential ceiling effects; variability remained substantial in the experimental group at both posttest and delayed posttest (Table 6), indicating that scores did not compress at the upper end of the scale. Overall, the pattern suggests substantial learning and partial decay in the experimental condition, with near-baseline performance in the control condition at delay.

Table 7
Descriptive Statistics for Vocabulary Scores by Group and Time Point

Time point	Group	n	M	SD	95% CI for M
Pretest (Time 1)	CG	55	18.78	3.86	[17.74, 19.82]
Pretest (Time 1)	EG	55	18.84	2.81	[18.08, 19.60]
Posttest (Time 2)	CG	55	22.27	2.16	[21.69, 22.86]
Posttest (Time 2)	EG	55	78.96	9.34	[76.44, 81.49]
Delayed posttest (Time 3)	CG	55	19.13	3.34	[18.22, 20.03]
Delayed posttest (Time 3)	EG	55	64.38	7.62	[62.32, 66.44]
Note. EG = experimental group; CG = control group.					

Mixed ANOVA: time trends and group × time interaction

To test whether change over time differed by condition, a 2 (Group: EG vs. CG) × 3 (Time: pretest, posttest, delayed posttest) mixed ANOVA was conducted. Greenhouse–Geisser correction was applied for within-subject effects. There was a significant main effect of Time, $F(1.821, 196.641) = 1325.990$, $p < .001$, $\eta p^2 = .925$, and a significant main effect of Group, $F(1, 108) = 2017.179$, $p < .001$, $\eta p^2 = .949$. Importantly, the Time × Group interaction was significant, $F(1.821, 196.641) = 1103.336$, $p < .001$, $\eta p^2 = .911$, indicating differential change over time between groups.

Because Box's test indicated unequal covariance matrices, Pillai's Trace was examined as a robustness check. The multivariate results corroborated the univariate conclusions, showing a significant effect of Time, Pillai's Trace = .949, $F(2, 107) = 1001.978$, $p < .001$, and a significant Time × Group interaction, Pillai's Trace = .940, $F(2, 107) = 832.939$, $p < .001$. Mixed ANOVA results are summarized in Table 7.

Table 8
Mixed ANOVA Results for Vocabulary Scores (Univariate GG and Multivariate Pillai's Trace)

Effect	Univariate F	df	p	ηp²	Pillai V	Multivariate F	df	*p*	ηp²
Time	1325.990	1.821, 196.641	< .001	.925	.949	1001.978	2, 107	< .001	.949
Time x Group	1103.336	1.821, 196.641	< .001	.911	.940	832.939	2, 107	< .001	.940
Group (between-subjects)	2017.179	1, 108	< .001	.949	-	-	-	—	-

Note. GG = Greenhouse-Geisser correction (Mauchly's W = .902, $\chi^2(2) = 11.090$, p = .004). Pillai's Trace is reported as a robustness check because Box's test indicated unequal covariance matrices (Box's M = 115.625, F = 18.690, p < .001).

Baseline equivalence (pretest)

A Welch independent-samples t test (given variance heterogeneity at Time 1) indicated no significant baseline difference between groups, t(98.789) = 0.085, p = .933, d = 0.02, mean difference = 0.055, 95% CI [− 1.223, 1.332] (see Table 6 for descriptives and Table 8 for test statistics). This supports baseline comparability prior to the intervention.

Research question 1: immediate vocabulary acquisition effects

At immediate posttest, the experimental group scored substantially higher than the control group, t(59.785) = 43.865, p < .001, with a mean difference of 56.691 points and a very large effect size (d = 8.37). The 95% CI for the mean difference was [54.106, 59.276] (Welch CI). Because the assessment targeted the same 100 synonyms used in the passages, these outcomes reflect learning of a tightly specified lexical set; effect sizes should be interpreted in that context.

Research question 2: delayed retention effects

At delayed posttest, the experimental group maintained significantly higher scores than the control group, t(74.043) = 40.311, p < .001, d = 7.69. Retention within the EG was quantified using the retention index: [((M_delayed − M_pre) / (M_post − M_pre)) × 100] = 75.7%. Within-group patterns indicated a significant decline from posttest to delayed posttest in the EG while remaining well above baseline. In contrast, the CG returned to near-baseline levels, with no reliable difference between pretest and delayed posttest (p = 1.000).

Planned comparisons addressing RQ1 and RQ2 are reported in Table 8 (Welch and paired-samples tests). Bonferroni-adjusted pairwise comparisons from the mixed ANOVA are reported in Table 9 to corroborate the planned comparisons under familywise error control.

Because the study did not include online processing measures (e.g., eye-tracking, reading-time indices, or stimulated recall), the present results address outcomes (test performance and retention) rather than directly testing hypothesized processing mechanisms.

Table 9
Between-Group and Key Within-Group Comparisons for RQ1 and RQ2

Comparison	t	df	p	Mean difference	95% CI	Cohen's d
EG vs. CG at pretest	0.085	98.789	.933	0.055	[-1.223, 1.332]	0.02
EG vs. CG at posttest	43.865	59.785	< .001	56.691	[54.106, 59.276]	8.37
EG vs. CG at delayed posttest	40.311	74.043	< .001	45.255	[43.018, 47.491]	7.69
Within EG: Pretest vs. posttest	59.420	54	< .001	60.127	[57.666, 62.588]	8.35
Within EG: Posttest vs. delayed posttest	15.992	54	< .001	14.582	[12.754, 16.410]	2.16
Within CG: Pretest vs. posttest	3.450	54	.002	3.491	[1.030, 5.952]	0.47
Within CG: Posttest vs. delayed posttest	5.429	54	< .001	3.145	[1.984, 4.307]	0.73

Note. Between-group tests report Welch's t (unequal variances). Within-group tests are paired-samples t tests.

Table 10
Bonferroni-Adjusted Pairwise Comparisons from Mixed ANOVA

Group	Comparison	Mean difference (Time I - Time J)	p (Bonferroni)	95% CI	Interpretation
CG	Pretest vs. posttest	-3.491	.002	[-5.952, -1.030]	Small gain
CG	Pretest vs. delayed posttest	-0.345	1.000	[-2.562, 1.871]	No retention
CG	Posttest vs. delayed posttest	3.145	< .001	[1.288, 5.003]	Significant decay
EG	Pretest vs. posttest	-60.127	< .001	[-62.588, -57.666]	Massive gain
EG	Pretest vs. delayed posttest	-45.545	< .001	[-47.762, -43.329]	Strong retention
EG	Posttest vs. delayed posttest	14.582	< .001	[12.724, 16.439]	Decay, but high retention

Note. Negative values indicate higher scores at the later time point (consistent with SPSS direction coding).

Retention and decay indices

To present retention and decay in an interpretable form, Table 11 reports gain, decay, net retention, and percent of gain retained for each group.

Table 11
Retention and Decay Indices by Group

Group	Pre (M)	Post (M)	Delayed (M)	Gain (Post-Pre)	Decay (Post-Delayed)	Net retained (Delayed-Pre)	% of gain retained
EG	18.84	78.96	64.38	60.12	14.58	45.54	75.7%
CG	18.78	22.27	19.13	3.49	3.14	0.35	10.0%

Note. % retained = ((Delayed - Pre) / (Post - Pre)) x 100.

The significant Time × Group interaction is visually represented in Fig. 1, which illustrates the differential trajectories of vocabulary development between the experimental and control groups across the three assessment points.

Alt Text: The x-axis represents three testing times: Pretest (Time 1), Immediate Posttest (Time 2), and Delayed Posttest (Time 3; two weeks later). The y-axis shows vocabulary test scores ranging from 0 to 100 (total correct responses). Two lines are displayed: one for the Experimental group and one for the Control group. Both groups begin with similar mean scores at the pretest. At the immediate posttest, the

Experimental group shows a substantial increase in mean vocabulary scores, while the Control group shows only a slight improvement. At the delayed posttest, the Experimental group's scores decrease slightly but remain higher than pretest levels and higher than the Control group. The Control group shows minimal change across all three time points.

5. Discussion

The present study examined whether synonym-embedment in reading passages improves EFL freshmen's vocabulary acquisition and delayed retention compared with reading the same passages without synonym support. Across analyses, the significant Group × Time interaction (Table 7) and the planned follow-up comparisons (Tables 8–10) show a clear pattern: the experimental group achieved substantially stronger immediate gains and maintained markedly higher performance over a two-week delay.

5.1 Design considerations and strength of inference

Because participants were taught in intact classes, individual random assignment was not feasible and the study remains quasi-experimental. As a result, internal validity threats cannot be fully eliminated, including instructor effects, differential classroom dynamics, contamination across conditions, differential motivation, and Hawthorne effects. In addition, because students were nested within intact classes, observations may not be fully independent. Future work should therefore replicate the design with multiple classes per condition (cluster randomization) or use multilevel models when a sufficient number of clusters is available. At the same time, several features strengthen the credibility of the observed pattern. The groups were equivalent at baseline (pretest), followed the same schedule and passages, and differed only in the presence versus absence of parenthetical synonym cueing. Baseline equivalence reduces (but does not remove) selection bias, particularly for unmeasured variables such as reading habits or strategic behaviour. Taken together, these considerations support a cautious interpretation of the findings. The results suggest that synonym-embedment was associated with greater immediate vocabulary gains and stronger two-week retention than exposure to the target synonyms without parenthetical semantic support. However, because the study used intact classes and one class per condition, the findings should not be interpreted as establishing that synonym-embedment alone caused the observed gains.

5.2 RQ1: Immediate Vocabulary Acquisition

RQ1 asked whether synonym-embedment facilitates immediate vocabulary acquisition compared with exposure to the target synonyms without synonym support. The experimental group's substantially higher posttest performance, combined with near-identical pretest scores, suggests that in-text synonym cueing may have facilitated vocabulary learning during comprehension-oriented reading. More specifically, the findings suggest that embedding unfamiliar target synonyms alongside familiar lexical anchors may facilitate vocabulary learning beyond exposure to the target synonyms alone. This interpretation is consistent with evidence that reading-based vocabulary learning may be strengthened

when unfamiliar items are encountered under conditions that facilitate meaning access and integration (Pellicer-Sánchez, 2016).

5.3 Mechanism interpretation (noticing, mapping, and integration)

The pattern is consistent with the mechanism model developed in the literature review. By providing an immediate semantic link to an already-known lexical concept, synonym cues likely reduced inferential ambiguity and increased the probability that learners would (a) notice the target item, (b) map form to meaning successfully, and (c) integrate the item into a developing semantic network (Craik & Lockhart, 1972; Schmidt, 1990). In this sense, synonym-embedment operates as semantic enrichment embedded within the running text: it supports meaning access while preserving reading continuity, thereby increasing the likelihood of meaningful processing during reading.

5.4 Balanced interpretation and boundary conditions

A plausible counterpoint is that synonym-embedment may reduce inferencing effort and therefore reduce “desirable difficulty” (Bjork & Bjork, 2011). If the treatment only increased processing fluency without strengthening learning, one might expect sharper decay at delay. Instead, the retention profile suggests that reduced inferencing demands did not undermine two-week retention and may have supported the formation of durable form–meaning links. However, synonyms and near-synonyms are rarely fully interchangeable; parenthetical cueing may introduce semantic fuzziness when paired words differ in register, collocation, or pragmatic constraints. Because the outcome measure assessed meaning recognition, the present results cannot determine whether synonym-embedment improved precision of use. Future studies should therefore incorporate outcomes that tap lexical precision and productive control (e.g., collocation appropriateness, register judgments, controlled production) rather than recognition alone.

5.5 RQ2: Two-Week Delayed Retention

RQ2 examined whether synonym-embedment was associated with stronger two-week vocabulary retention than exposure to the target synonyms without synonym support. Retention was evaluated by comparing the two groups at the delayed posttest and by examining changes from the immediate posttest to the delayed posttest. At the two-week delay, the experimental group maintained significantly higher scores than the control group (Tables 9–11). Expressed as gain maintenance, the experimental group retained 75.7% of its posttest gains relative to baseline, indicating substantial maintenance of vocabulary gains over the two-week interval. The within-group trajectories further clarify this pattern. Although the experimental group showed a significant decline from the immediate posttest to the delayed posttest, its performance remained markedly above baseline. In contrast, the control group returned to near-baseline levels, with no significant difference between pretest and delayed posttest performance. This divergence is consistent with the significant Time × Group interaction (Table 8) and is also consistent with the view that vocabulary retention may be influenced by the quality of processing

during lexical encounters and the strength of form–meaning mapping (Craik & Lockhart, 1972; Webb & Nation, 2017).

5.6 Benchmarking and effects size

Incidental vocabulary learning from meaning-focused input is often modest and variable, and delayed outcomes are frequently smaller than immediate outcomes. Meta-analytic benchmarks for reading-only exposure commonly suggest relatively small proportions of unknown items learned, with delayed learning often in the approximate 6%–17% range depending on design and measurement (Webb et al., 2023). Relative to these benchmarks, the present gains and retention are large. Two design features likely contributed: (a) the study used a tightly defined, screened lexical set, and (b) the reading input was intentionally engineered to support inference and elaboration (controlled density, repeated sessions, in-text semantic cues). Accordingly, the outcomes are best interpreted as evidence of what can be achieved through principled materials design rather than as typical gains from unmodified “reading-only” exposure.

5.7 Incidental-with-support and ecological realism

Although learning occurred during comprehension-oriented reading rather than explicit memorization, the intervention is best characterized as incidental-with-support. Parenthetical synonym cueing provided integrated semantic scaffolding at the point of encounter and likely also functioned as a consistent attentional cue in the running text. This framing is consistent with noticing-based accounts (Schmidt, 1990) and input-enhancement perspectives, while emphasizing that salience was paired with semantic access. At the same time, the intervention—while classroom-feasible (printable texts, reading-first orientation)—was optimized through prior-knowledge screening, controlled density, repeated exposure, and systematic embedding of integrated semantic cues. This combination helps explain the magnitude of effects and clarifies the contribution: carefully designed reading materials can substantially reshape retention trajectories in non-immersive tertiary contexts where classroom reading is a major input source.

5.8 Responsible interpretation of effects and measurement alignment

The magnitude of effects should be interpreted in light of the close alignment between instruction and assessment. Effect sizes in L2 research are sensitive to exposure structure, outcome specificity, and the degree to which tests directly target taught items (Plonsky & Oswald, 2014). Here, the dependent measure was a meaning-recognition test aligned to the 100 embedded items and administered repeatedly. Consequently, the strongest inference supported is learning and retention of a defined lexical set under structured exposure conditions. The results do not, by themselves, establish generalized vocabulary growth, transfer to new texts without cues, or productive command of the items. Because the same target set was assessed repeatedly in a similar format across Time 1–Time 3, test–retest familiarity may have contributed to gains; however, the size of the between-group separation suggests repeated testing alone is unlikely to account for the treatment effect. Future work should include

alternate test forms, transfer tasks using new passages, and at least one productive outcome measure to better evaluate generalization and precision.

5.9 Theoretical contribution

Beyond demonstrating an instructional effect, the study contributes to input-based vocabulary learning theory by showing that integrated semantic enrichment embedded within discourse can yield retention trajectories that differ from those typically reported for unassisted meaning-focused reading. The combination of large immediate gains and substantial maintenance is consistent with lexical-quality accounts of durable form–meaning representations (Perfetti, 2007) and with noticing-based accounts emphasizing attention to form during meaningful input (Schmidt, 1990). A testable implication is that synonym-embedment should reduce lookup-like behavior and increase integration-based processing; future studies can examine this through eye-tracking, stimulated recall, or other process-tracing methods.

5.10 Pedagogical implications

Synonym-embedment is a low-cost, printable materials-design option that can be implemented without technology and without interrupting reading flow. Based on the mechanism model and observed learning and retention patterns, practical steps for tertiary EFL contexts include: (a) selecting mid-frequency targets relevant to academic reading and confirming unfamiliarity through brief screening; (b) embedding synonyms near-adjacent while maintaining grammatical and pragmatic naturalness; (c) limiting density (about 8–10 targets per short passage) to avoid overload; (d) using consistent parenthetical cueing to support attention without turning reading into overt vocabulary study; (e) preserving comprehension-first tasks to maintain meaning-focused processing; and (f) recycling targets across later readings to support spaced re-encounters associated with stronger retention (Webb et al., 2023). Because near-synonyms may introduce semantic fuzziness, instructors should also check register/collocational compatibility and consider brief follow-up tasks (e.g., “which word fits better here?”) to sharpen boundaries of use.

6. Limitations and Future Directions

Several limitations should be acknowledged. The study was conducted in a single institutional context with a specific age and proficiency range, which may limit generalizability. The intervention examined visual, text-based synonym support; other modalities (audio, digital, multimodal) were not investigated. The delayed posttest interval was two weeks; longer follow-ups are needed to evaluate retention beyond short-term consolidation. The study targeted mid-frequency vocabulary, and learning trajectories may differ for high- or low-frequency items. Finally, the outcome measure assessed receptive meaning recognition and did not capture productive use or transfer to new contexts. Future studies should broaden samples, vary text genres, extend retention intervals, use alternate test forms, and include transfer measures and at least one productive outcome to evaluate flexible and context-appropriate deployment of learned items.

7. Conclusion

The present study examined the association between synonym-embedded reading passages and vocabulary learning and delayed retention among EFL freshmen. Using a quasi-experimental pretest–posttest–delayed posttest design, the study compared learners who encountered target synonyms embedded alongside familiar lexical anchors with learners who received exposure to the target synonyms without synonym support. The findings indicated that synonym embedding was associated with greater immediate vocabulary gains and higher performance on the two-week delayed posttest than the no-support condition. These findings suggest that embedding unfamiliar synonyms alongside familiar lexical anchors in meaningful reading contexts may provide useful semantic support for vocabulary learning and subsequent retention.

Overall, the results support synonym embedding as a promising materials-design approach for reading-based vocabulary instruction. The stronger delayed-posttest performance of the experimental group further suggests that the benefits of synonym support may extend beyond immediate learning to delayed vocabulary retention. Nevertheless, the findings should be interpreted cautiously. Because the study employed a quasi-experimental design and relied on one intact class per condition, the observed between-group differences cannot be attributed exclusively to the synonym-embedding intervention. Accordingly, the present findings provide evidence of an association between synonym embedding and vocabulary learning and delayed retention rather than definitive evidence of a causal effect. Future research should replicate these findings with larger and more diverse samples, multiple classes per condition, random assignment where feasible, and longer delayed-posttest intervals.

References

1. Agustín-Llach, M. P., & Rubio, J. (2024). Navigating the mental lexicon: Network structures, lexical search and lexical retrieval. *Journal of Psycholinguistic Research*, *53(2)*, 21. https://doi.org/10.1007/s10936-024-10059-8
2. Aljasir, N. (2025). Vocabulary learning strategies among Saudi EFL learners: A proficiency-level comparison using think-aloud protocols. *Cogent Education*, *12*(1), 2472480. https://doi.org/10.1080/2331186X.2025.2472480
3. Alsharif, R. (2022). Relationship between Vocabulary Learning Strategies and Vocabulary Size: Evidence from Saudi Female EFL Learners. *International Journal of Education and Literacy Studies*, *10(1)*, 188–197. https://doi.org/10.7575/aiac.ijels.v.10n.1p.188
4. Binhomran, K., & Altalhab, S. (2023). A systematic review of EFL research on the impact of reading comprehension on students' vocabulary retention, and the relation between reading and vocabulary size. *International Journal of Language and Literary Studies*, *5*(3), 146–168. https://doi.org/10.36892/ijlls.v5i3.1395
5. Bjork, E. L., & Bjork, R. A. (2011). Making things hard on yourself, but in a good way: Creating desirable difficulties to enhance learning. In M. A. Gernsbacher, R. W. Pew, L. M. Hough, & J. R.

Pomerantz (Eds.), *Psychology and the real world: Essays illustrating fundamental contributions to society* (pp. 56–64). Worth Publishers.

6. Boers, F. (2022). Glossing and vocabulary learning. *Language Teaching*, *55*(1), 1–23. https://doi.org/10.1017/S0261444821000252
7. Chandy, R., Serrano, R., & Pellicer-Sánchez, A. (2024). 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
8. Chaokongjakra, W. (2023). Is it "important," "significant," or "crucial"? A corpus-based study of English synonyms. *LEARN Journal: Language Education and Acquisition Research Network, 16*(2), 512–532. https://so04.tci-thaijo.org/index.php/LEARN/article/view/266969
9. 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
10. Crocker, L., & Algina, J. (1986). *Introduction to classical and modern test theory*. Holt, Rinehart and Winston, 6277 Sea Harbor Drive, Orlando, FL 32887.
11. Ebel, R. L., & Frisbie, D. A. (1991). *Essentials of educational measurement* (5th ed.). Prentice Hall.
12. Eskandari, Z., Khatin-Zadeh, O., Farsani, D., & Banaruee, H. (2024). The effect of type of task on EFL learners' vocabulary learning. *Frontiers in Psychology*, *15*, 1306306. https://doi.org/10.3389/fpsyg.2024.1306306
13. Georgiou, G., Inoue, T., & Parrila, R. (2023). Are Vocabulary and Word Reading Reciprocally Related? *Scientific Studies of Reading*, *27*(2), 160–168. https://doi.org/10.1080/10888438.2022.2123275
14. Godfroid, A. (2019). *Eye tracking in second language acquisition and bilingualism: A research synthesis and methodological guide*. Routledge.
15. Haladyna, T. M., Downing, S. M., & Rodriguez, M. C. (2002). A review of multiple-choice item-writing guidelines for classroom assessment. *Applied measurement in education*, *15*(3), 309-333. https://doi.org/10.1207/S15324818AME1503_5
16. Hill, L. A. (1977). *Advanced stories for reproduction*. Cambridge University Press.
17. Lestiana, S., Ilma, R., & Saputra, E. (2025). Students' vocabulary mastery and their reading comprehension. *English Community Journal*, *9*(1), 1–9. https://doi.org/10.32502/ecj.v9i1.6757
18. McKeown, M. G. (2019). Effective vocabulary instruction fosters knowing words, using words, and understanding how words work. *Language, Speech, and Hearing Services in Schools, 50*(4), 466–476. https://doi.org/10.1044/2019_LSHSS-VOIA-18-0126
19. Nation, I. S. P. (2001). *Learning vocabulary in another language*. Cambridge University Press.
20. Pellicer-Sánchez, A. (2016). Incidental L2 vocabulary acquisition from and while reading: An eye-tracking study. *Studies in Second Language Acquisition*, *38*(1), 97–130. https://doi.org/10.1017/S0272263115000224

21. Pellicer-Sanchez, A., Siyanova-Chanturia, A., & Parente, F. (2022). The effect of frequency of exposure on the processing and learning of collocations: A comparison of first and second language readers' eye movements. *Applied Psycholinguistics, 43*(3), 727-756. https://doi.org/10.1017/S014271642200011X
22. Perfetti, C. A. (2007). Reading ability: Lexical quality to comprehension. *Scientific Studies of Reading*, *11(4)*, 357–383. https://doi.org/10.1080/10888430701530730
23. Plonsky, L., & Oswald, F. L. (2014). How big is “big”? Interpreting effect sizes in L2 research. *Language Learning*, *64*(4), 878–912. https://doi.org/10.1111/lang.12079
24. Read, J. (2004). Plumbing the depths: How should the construct of vocabulary knowledge be defined? In *Vocabulary* (pp. 70-86). Routledge.
25. Rezigalla, A.A., Eleragi, A.M.E.S.A., Elhussein, A.B. *et al.* Item analysis: the impact of distractor efficiency on the difficulty index and discrimination power of multiple-choice items. *BMC Med Educ* 24, 445 (2024). https://doi.org/10.1186/s12909-024-05433-y
26. Schmidt, R. (1990). The role of consciousness in second language learning. *Applied Linguistics*, *11*(2*)*, 129–158. https://doi.org/10.1093/applin/11.2.129
27. Shahiwala, S., Rahul, D. R., & Baker, J. R. (2024). Incidental vocabulary learning: A scientometric review. *Research Methods in Applied Linguistics*, *3*(3), 100160. https://doi.org/10.1016/j.rmal.2024.100160
28. Sweller, J. (1988). Cognitive load during problem solving: Effects on learning. *Cognitive Science*, *12(2)*, 257–285. https://doi.org/10.1207/s15516709cog1202_4
29. Teng, F. (2020). Retention of new words learned incidentally from reading: Word exposure frequency, L1 marginal glosses, and their combination. *Language Teaching Research*, *24(6)*, 785–812.
30. Tinkham, T. (1993). The effect of semantic clustering on the learning of second language vocabulary. *System*, *21(3)*, 371–380. https://doi.org/10.1016/0346-251X(93)90040-3
31. Van den Broek, G. S., Wesseling, E., Huijssen, L., Lettink, M., & Van Gog, T. (2022). Vocabulary learning during reading: Benefits of contextual inferences versus retrieval opportunities. *Cognitive Science*, *46(4)*, e13135. https://doi.org/10.1111/cogs.13135
32. Wagner, R. K., & Meros, D. (2010). Vocabulary and reading comprehension: Direct, indirect, and reciprocal influences. *Focus on Exceptional Children*, *2010*, 1G1–245473138.
33. Webb, S., & Nation, I. S. P. (2017). *How vocabulary is learned*. Oxford University Press.
34. 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
35. Wesche, M., & Paribakht, T. S. (1996). Assessing second language vocabulary knowledge: Depth versus breadth. *The Canadian Modern Language Review*, *53(1)*, 13–40. https://doi.org/10.3138/cmlr.53.1.13

36. Yanagisawa, A., Webb, S., & Uchihara, T. (2020). How do different forms of glossing contribute to L2 vocabulary learning from reading? a meta-regression analysis. *Studies in Second Language Acquisition*, *42(2)*, 411–438. https://doi.org/10.1017/S0272263119000688
37. Zeng, Y., Kuo, L. J., Chen, L., Lin, J. A., & Shen, H. (2025). Vocabulary instruction for English learners: A systematic review connecting theories, research, and practices. *Education Sciences*, *15(3)*, 262. https://doi.org/10.3390/educsci15030262

Figures

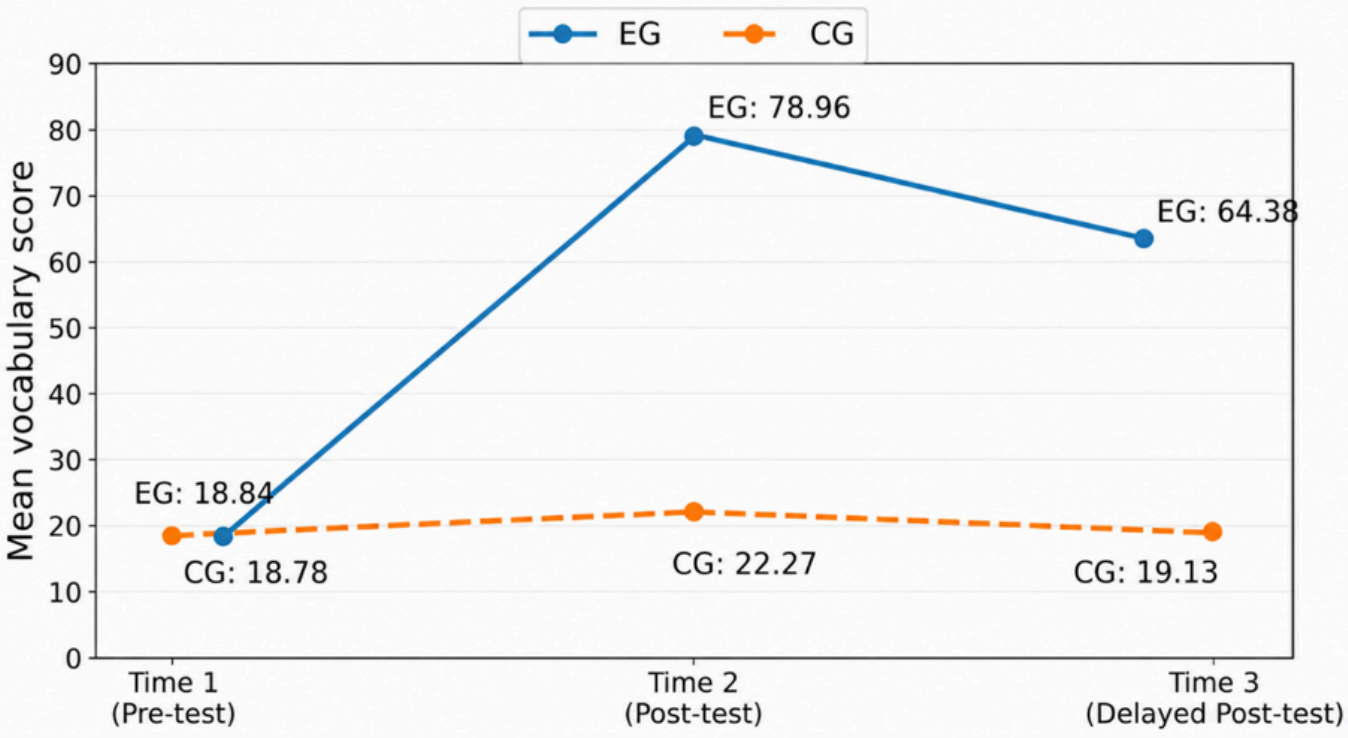


Figure 1

Mean Vocabulary Scores Across Time by Group (Experimental vs. Control)

Note. Scores represent total correct responses on the 0–100 vocabulary test at pretest (Time 1), immediate posttest (Time 2), and delayed posttest (Time 3; two weeks later).

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- Appendices.docx