

GEMINI AI-SUPPORTED DIGITAL POSTERS IN EFL CLASSROOM: LEXICAL DEPTH AND TWO-WEEK RETENTION

¹*Uni Maharani, ¹Sudiyono, ¹Yuyun Yulia

¹English Language Education, Faculty of Languages, Arts, and Culture, Universitas Negeri
Yogyakarta, Jl. Colombo No.1, Karang Malang, Sleman, Yogyakarta, Indonesia

*Corresponding Author Email: unimaharani.2025@student.uny.ac.id

Article Info	Abstract
Article History Received: June 2026 Revised: July 2026 Accepted: September 2026 Published: October 2026	<i>Limited lexical depth and the difficulty to retain words after teaching are issues for junior high school students who learn to speak English as a foreign language. This study investigated vocabulary learning and student perceptions of a Gemini AI-supported digital poster project in a Grade 8 classroom in Palembang, Indonesia. A mixed-methods quasi-experiment was conducted on two intact classes (N = 48; 24 per group) in six meetings. The experimental group explored meanings, context and associated meanings using Gemini AI and produced Canva posters through repeated practice, peer interaction and teacher guidance. The 30 item vocabulary assessment was done before instruction and immediately afterward and two weeks later. Fifteen item questionnaires and six semi-structured interviews were used to extract perception information. The experimental and control mean post-test was 84.37 and 71.04 respectively. The two-week delayed means were 80.92 and 66.13 as compared to the final results from the summaries. The post-test mean difference was 13.33 (95% CI [9.34, 17.32], $p < .001$) and the delayed difference was 14.79 (95% CI [10.65, 18.93], $p < .001$). These cross-sectional tests do not establish a group-by-time effect. Students generally described interest, contextual support and creativity as well as connectivity problems, uneven digital skills and a dependence on AI. The instructional package was associated with higher immediate and two-week delayed scores, but its individual components cannot be separated. We need to be mindful of teacher monitoring, checking AI responses, privacy protection and the need for AI literacy and education.</i>
Keywords Digital posters; Gemini AI; Vocabulary learning; Lexical depth; Vocabulary retention;	
How to cite: Maharani, U., Sudiyono, S., & Yulia, Y. (2026). Gemini AI-Supported Digital Posters in EFL Classroom: Lexical Depth and Two-Week Retention, <i>JOLLT Journal of Languages and Language Teaching</i> , 14(4), 1194-1208. Doi: https://doi.org/10.33394/joltt.v14i4.21032	

Copyright© 2026, Uni Maharani et al.

This is an open-access article under the [CC-BY-SA](#) License



INTRODUCTION

Vocabulary knowledge enables comprehension and expression but a word is not fully learned until it has been recognized for its unique translation. Vocabulary breadth is the number of lexical items available. Lexical depth is the amount of knowledge on each item, the meaning of the language (like form, meaning link, semantic association, collocation and contextually appropriate use). Depth predicts something other than simply the number of words that are known (Qian, 2022). We are testing this in terms of recognition items, contextual tasks, collocation matching and recall. Contextual and collocation items account for selected depth facets. A simple multiple choice recognition and isolated recall mainly focus on recognition or retrieval. Without item-level information we can call the overall score lexical depth. The first test conducted in the post-test period is to measure achievement just after we teach. It does not show durable long-term retention. Score maintenance compared to

the immediately preceding test also differs from absolute delayed performance. These differences are important because a group can have a delayed score mainly because it started the interval at a higher post-test level.

Generative AI can quickly generate examples, explanations and alternative formulations that students can question and revise. Studies of ChatGPT in language education describe these possibilities and risks of inaccuracies and uncritical reliance (Alsagoor et al., 2025). An EFL vocabulary study with 50 female university students used ChatGPT for five weeks and examined pre-post vocabulary knowledge and interviews, without the same school level Gemini/poster/delayed-test combination (Alsagoor et al., 2025). A review of 23 school EFL studies identifies infrastructure, literacy and teacher readiness as moderators, but does not test this particular instructional package (Kundu & Bej, 2025). Broader syntheses of technology-assisted vocabulary learning and AI-assisted vocabulary teaching likewise aggregate varied tools, populations and outcomes (Simsek et al., 2026). These comparisons concern design, rather than a claim that no related school study exists.

Generative output can also contain inaccurate meanings, unsuitable collocations or misleading examples. Younger learners may copy answers without retrieval or evaluation. Access depends on devices, internet service and digital skills. Privacy and age-appropriate supervision matter when minors use an external system. Thus, teacher-selected target words, independent checking and limits on personal information are pedagogical and ethical conditions, not automatic properties of an AI tool (Pan, 2024; Hwang et al., 2024). Technological pedagogical content knowledge (TPACK) describes how teachers integrate a tool, vocabulary content and teaching practice; it is not a theory of lexical memory (Mishra & Koehler, 2006).

Initial classroom observations in Palembang junior high school, early classroom observations showed difficulty using recently taught words in context and low participation. This study offers a bounded account of Indonesian Grade 8 learners using Gemini AI to explore vocabulary and Canva to produce digital posters, with an immediate mixed vocabulary assessment, a two-week delayed assessment and embedded reports of benefits and implementation challenges. The novelty is the combination of population, platform, generative exploration, student-made posters and three measurement occasions, rather than the novelty of AI in language education itself. Therefore, the study asks: How do the two classes differ in immediate performance on a mixed vocabulary test with depth-oriented components following the Gemini AI and digital-poster package versus conventional instruction, and is improvement over baseline greater in the experimental class?; How do the classes differ in absolute scores on the two-week delayed test, and what can be said about score maintenance from the immediate post-test?; and How do students in the experimental class perceive the opportunities and challenges of Gemini AI and digital-poster projects for vocabulary learning?

RESEARCH METHOD

Research Design

This was a quantitative-dominant mixed-methods quasi-experiment. The main quantitative strand compared two existing classes at pre-test, immediate post-test and a two-week delayed post-test. The embedded strand consisted of a questionnaire after treatment and six semi-structured interviews after treatment, which contextualize, not verify, the differences in score. This integration is made in the Discussion when reported experiences and implementation constraints are considered alongside the quantitative pattern. There is no evidence that we conducted a joint display, statistical-qualitative merging algorithm or triangulation validity test. Individual assignment was impossible because classes were intact. (Creswell & Plano Clark, 2018).

A quasi-experimental design was chosen because in the current classroom situation the researcher was unable to randomly assign students to different groups. According to Creswell and Creswell (2018), quasi-experimental research is appropriate when studying causal relationships in natural settings of education (Creswell & Creswell, 2018). In this study, there were two intact classes. One was the experimental group and students were taught vocabulary by the AI and the other was the control group and taught the same kind of vocabulary.

The study had a pre-test, post-test, and delayed post-test design. The pre-test was conducted before the treatment to assess students' initial lexical depth and vocabulary retention levels. The post-test was conducted immediately after the treatment to measure vocabulary improvement, while the delayed post-test was taken two weeks later in order to gauge the extent to which the students retained the learned vocabulary over time.

Table 1
Research Design

Group	Pre-Test	Treatment	Post-Test	Delayed Post-Test
Experimental	O1	AI-Assisted Learning	O2	O3
Control	O1	Conventional Learning	O2	O3

Research Participants

The participants were 48 students ages 13-14 years of grade 8 from two intact Grade 8 classes at a junior high school in Palembang during the 2024/2025 academic year. The researcher was a preservice teacher completing a school teaching practicum and taught both classes. The two classes were chosen as they were the ones the researcher had taught and their timetables were also in place for teaching and assessing them. One intact class served as the experimental group ($n = 24$) and the other served as the control group ($n = 24$). Neither classes nor students were randomly assigned. Although both classes were from one school curriculum, one identical class was used in each condition, so the differences between classes cannot be divorced from the effect of the intervention.

The students' English proficiency was generally described as low to intermediate based on school records and teacher assessments. This explanation was not based on a standardized proficiency test. English grades of earlier grades in that school were shown in the school archives and used as a marker of prior achievement. Students did have some experience with AI tools at school but it was small and relatively easy to use. The school computer laboratory provided the devices and Internet access for the study activities. All 48 students completed the pre-test, immediate post-test, and two-week delayed post-test, and there was no attrition across the three testing occasions.

Intervention and Procedure

Both classes passed the baseline vocabulary test before instruction. The researcher taught both intact classes in six instructional sessions. In the experimental class, students used Gemini AI in the school computer laboratory to learn the meaning, context, synonyms, and collocations of target words. They used this information to create messages for digital vocabulary posters they made using Canva. Given their low to intermediate English, students wrote prompts using Indonesian instructions and English vocabulary. They discussed Gemini's suggestions with the teacher and revised their word choices and poster wording before completing their posters. The researcher taught vocabulary in the control class with explanations, textbook-based activities, and translation exercises. The vocabulary tests were considered to be more important than the quality of the Canva posters as they were based on achievements reported in the study. The first post-test was done immediately after the sixth meeting, and the second post-test was done two weeks later.

Vocabulary Assessment and Blueprint

The vocabulary performance was evaluated using a 30 item test before instruction, after the intervention and in the following weeks. The items were built from vocabulary taught in the course and were composed of multiple choice questions, contextual vocabulary tasks, collocation matching and vocabulary recall. Table 2 illustrates the relationship between these formats and vocabulary knowledge assessment.

Table 2
Reported Test Formats and the Evidence Needed to Validate their Interpretation

Format reported	Aspect assessed	Interpretation
Multiple choice	Recognition of word meaning or distinctions among meanings	Basic meaning recognition alone does not establish lexical depth
Contextual vocabulary task	Selection or use of a word appropriate to a sentence or situation	Contextually appropriate use provides evidence of one aspect of lexical depth
Collocation matching	Recognition of conventional word combinations	Collocational knowledge is one aspect of lexical depth
Vocabulary recall	Retrieval of a word form or meaning	Isolated recall does not establish knowledge of contextual use

A sample's overall test score was viewed as mixed vocabulary performance with depth-oriented information and not as a pure measure of lexical depth. To complete the item blueprint, the original test must show the target words and how they were selected from the Grade 8 teaching material. The number and item numbers in each category, totaling 30. The word assessed by each item; the answer keys. The points for each response in any partial credit rules. The maximum raw score; and the conversion to the reported score scale. A separate lexical depth score can be calculated only if the original items and scoring records support that distinction.

Two English language education experts reviewed the test items. Their qualifications, review criteria, and any resulting revisions should be reported as evidence of the relevance of the test content. Expert review alone is not enough to establish test reliability and should be reported in that case. If a pilot was carried out prior to the main study, the participants, procedures, reliability estimates, and item analyses should be reported. Item difficulty and discrimination can be examined from real item-level responses. The internal consistency coefficient should be determined for how the items were scored. The analyses of the main study responses after data collection should be identified as such and not referred to as a pilot. The method should also indicate if the pre-test, immediate post-test, and delayed post-test used the same items, reordered items, or different forms. If identical items were reused, practice effects should be addressed. If different forms were used, how they were covered, scored, and if they were comparable must be explained.

Questionnaire and Interviews

After the class, the researcher administered a 15-item questionnaire to the experimental group to assess students' opinions on the Gemini AI and digital-poster activities. The responses ranged from 1 (strongly disagree) to 5 (strongly agree) on a five-point Likert scale. The statements were on learning motivation, classroom engagement, enjoyment, vocabulary development, and the usefulness of the tasks. The questionnaire was designed to assess the students' experiences and beliefs with respect to a reminder of vocabulary. Agreement with a statement about remembering vocabulary was not seen as evidence of retention. The available manuscript reports favorable response percentages for five statements but does not document the wording and full response distributions of all 15 items, the instrument's development and

review process, or an internal-consistency estimate. Accordingly, no validated composite questionnaire score is claimed.

Semi-structured interviews were conducted after the intervention with six students from the experimental class. They were interviewed based on classroom participation and vocabulary tests. Each interview lasted between 15 and 20 minutes and was audio-recorded with the consent of the students. Questions were asked about Gemini AI and digital poster production, learning opportunities, difficulties they faced, and suggestions for future lessons. Interview accounts were used to contextualize the test and questionnaire results rather than as evidence of the efficacy of the intervention. The available information does not specify the exact selection rule or demonstrate that the six students were selected from different performance levels or that the samples included dissenting opinions, so the sampling is not meant as maximum variation sampling.

Data Analysis and Mixed-Methods Integration

The main intended analysis for the three occasions is a mixed-design ANOVA focusing on group x time interaction or a participant-level linear mixed-effects model if the data or assumptions are not available. A predictive comparison of the baseline to immediate change would address RQ1, while the adjusted baseline to delayed change and post-delayed change would address different areas of RQ2. Assess the distribution/residuals, outliers and group variances and covariance, and evaluate sphericity in ANOVA and apply Greenhouse-Geisser degrees of freedom in case of violations. Report F statistics, degrees of freedom, p values, partial η^2 and confidence intervals, with multiplicity-adjusted contrasts. For a linear mixed model, report the estimation of covariance, fixed effects, and estimated marginal contrasts.

The published means, SDs and nominal $n = 24$ per group do allow for independent Welch comparisons at each point, assuming that n corresponds to the full data. These are reconstructed, provisional descriptive cross-sections, and not substitutes for a repeated-measures interaction or a test of change. Hedges' g uses the pooled cross-sectional SD with small-sample correction; confidence intervals indicate unadjusted mean differences. Since there is only one intact class for each condition, even student p values can't account for treatment-class confounding and should be interpreted conditionally. Percentage maintenance is a ratio of group mean scores, not the percentage of words that students retain.

Vocabulary-test scores for each group at pre-test, immediate post-test and two-week delayed post-test were to be defined by means and standard deviations. Baseline scores were compared with a between-group estimate and 95% confidence interval, and we did not include descriptively small differences in this estimate as proof of a comparison. The main longitudinal analysis was a 2 (group: experimental, control) x 3 (time: pre-test, immediate post-test, delayed post-test) mixed design of variance with verification of the original participant-level scores and the model assumptions. Our main focus was the group x time interaction, i.e., the difference in the group x time effect which is important for whether the score change was different between the groups. We should report the group x time effect separately.

Before interpreting the model, score distributions, residuals, between-group variance, and sphericity should be examined. If sphericity was violated, we should report corrected degrees of freedom. We should also report the test statistic, exact degrees of freedom, p value, effect size and 95% confidence interval. The individual scores were required for each of the three cases and group mean and standard deviation alone were insufficient to calculate the interaction. Since each condition had only one intact class, any between-condition interpretation also had to account for class level differences.

Ethics and Protection of Minors

The researcher was a preservice teacher who was in teaching practicum and did not have direct contact with the parents or guardians of the students. Therefore he sought permission from the mentor teacher and homeroom teacher of each class for the students to participate. In line with school procedures, these teachers reported to parents or guardians about the study and obtained their consent. The researcher's dual role as a preservice teacher and researcher was also considered in reporting the study and interpreting the findings.

RESEARCH FINDINGS AND DISCUSSION

Research Findings

The two groups had descriptively similar pre-test means. After six instructional meetings, the experimental class had a higher immediate post-test mean than the control class and achieved a higher absolute score on the two-week delayed post-test. Students of the experimental class generally reported good experiences with Gemini AI and Canva digital-poster projects. Interview summaries also indicated unstable connectivity, difficulties using the digital tools, and sometimes uncritical reliance on AI-generated responses. Table 3 shows the three vocabulary-test occasions in a single comparison.

Table 3
Vocabulary-Test Scores and Between-Class Comparisons

Occasion	Experimental class, <i>M</i> (<i>SD</i>)	Control class, <i>M</i> (<i>SD</i>)	Mean difference [95% CI]	Welch's <i>t</i> (<i>df</i>), <i>p</i>	Hedges' <i>g</i>
Pre-test	58.21 (8.34)	57.88 (8.11)	0.33 [-4.45, 5.11]	0.14 (45.96), .890	0.04
Immediate post-test	84.37 (6.52)	71.04 (7.18)	13.33 [9.34, 17.32]	6.73 (45.58), < .001	1.91
Two-week delayed post-test	80.92 (6.81)	66.13 (7.42)	14.79 [10.65, 18.93]	7.19 (45.67), < .001	2.04

The two groups had descriptively similar pre-test means. After six instructional meetings, the experimental class reported a higher immediate post-test mean than the control class, and a higher absolute score on the two-week delayed post-test. Students of the experimental class generally reported good experiences with Gemini AI and Canva digital-poster projects. Interview summaries also indicated unstable connectivity, difficulties using the digital tools, and sometimes uncritical reliance on AI-generated responses. Table 1 shows the three vocabulary-test occasions in a single comparison.

The experimental mean rose 26.16 points from baseline to immediate test, while 13.16 points in the control class. The difference in these mean changes is 13.00 points. There are no covariances within the group which create an appropriate confidence interval, test or a standardized effect for this change difference. The absolute difference after the test is 13.33 points but cannot be attributed to Gemini separately from the rest of the package or definitively to the condition in a two-class design. At two weeks the absolute difference between the two was 14.79 points. The experimental mean fell 3.45 points from its immediate score and the control mean fell 4.91 points. These ratios represent group average scores, not individual retention rates or percentages of vocabulary words remembered. The greater absolute delayed score could be due to the already higher immediate score, but proportional maintenance is far smaller. No post-to-delayed change comparison and interaction are available without paired scores.

Effects of AI-Assisted Learning on Students' Lexical Depth

In order to investigate the effect of AI-assisted vocabulary learning on the lexical depth of students, we have tested the pre-test and the immediate post-test scores of both classes before instruction. The pre-test was asked before teaching to describe the students'

vocabulary-test performance and to compare the different scores between the classes at baseline. The descriptive results are presented in Table 4.

Table 4
Pre-Test Scores of Experimental and Control Classes

Group	N	Mean	SD
Experimental	24	58.21	8.34
Control	24	57.88	8.11

As shown in Table 4, the experimental class scored a mean pre-test score of 58.21 and control class 57.88. This difference is 0.33 points and shows that the average scores were descriptively similar before teaching. We also found no baseline difference from the summary statistics of $t(45.96) = 0.14$, $p = .890$, 95% CI $[-4.45, 5.11]$, Hedges' $g = 0.04$. But similar means and a nonsignificant test do not mean the two intact classes were equal in all regards. Classroom observations before the intervention described the students' reliance on direct translation and difficulty in contextual word choice, collocations, synonyms and relationships between words. These observations gave rise to a reason to consider vocabulary knowledge beyond simple word meanings. They were not included in the pre-test analysis, so they do not say how many students were experiencing each difficulty, or whether one class had higher rates compared to the other.

Following the six instructional meetings, an immediate post-test was administered to both classes. Table 5 presents the results.

Table 5
Post-Test Scores of Experimental and Control Classes

Group	N	Mean	SD
Experimental	24	84.37	6.52
Control	24	71.04	7.18

Both classes had higher mean scores at the immediate post-test than at the pre-test. The experimental mean increased from 58.21 to 84.37, whereas the control mean increased from 57.88 to 71.04. At the immediate post-test, the experimental class scored 13.33 points higher on average than the control class. This result describes stronger performance on the overall vocabulary test following the Gemini AI and Canva digital-poster instructional package. The test included contextual vocabulary and collocation tasks intended to assess aspects of lexical depth, alongside other vocabulary items. The original manuscript described improvements in contextual word choice and collocation exercises. Because item-level results and separate scores for these components were not provided, however, the available totals cannot establish how much of the post-test difference came from those particular tasks. The changes in group means are shown in Table 6.

Table 6
Changes in Mean Vocabulary-Test Scores From Pre-Test to Immediate Post-Test

Group	Pre-Test	Post-Test	Gain Score
Experimental	58.21	84.37	26.16
Control	57.88	71.04	13.16

As presented in Table 6, the experimental class's mean increase was 26.16 points, compared with 13.16 points in the control class. The difference between these descriptive increases was 13.00 points. Thus, the experimental mean rose by approximately twice as many points. This comparison does not, on its own, test whether improvement differed statistically between classes; that test requires the pre-test and post-test scores linked for each student. The finding describes a narrower spread of scores at the later occasion. It does not demonstrate that students with lower initial scores benefited equally or more than other students, because their individual score changes have not been examined.

For comparison with the percentage increases reported in the original manuscript, Table 7 expresses each change relative to its own class's pre-test mean.

Table 7
Descriptive Percentage Increase in Group Mean Vocabulary-Test Scores

Group	Percentage Increase
Experimental	44.94%
Control	22.73%

The percentages in Table 7 were calculated by dividing each increase in group mean by that class's pre-test mean. They describe relative changes in average test scores; they do not represent the percentage of words learned or a separate measure of lexical depth. The control value is 22.74% when calculated from the rounded means displayed in Tables 4 and 6.

Table 8 reports the between-class comparisons calculated from the available group sizes, means, and standard deviations.

Table 8
Between-Class Comparisons at Pre-Test and Immediate Post-Test

Occasion	Mean difference	Welch's <i>t</i> (<i>df</i>)	<i>p</i>	95% CI for mean difference	Hedges' <i>g</i>
Pre-test	0.33	0.14 (45.96)	0.890	[-4.45, 5.11]	0.04
Immediate post-test	13.33	6.73 (45.58)	< .001	[9.34, 17.32]	1.91

Mean differences are experimental minus control. The statistics were recalculated from the reported, rounded summary values and should be verified against the original scores. The confidence intervals are for between-class differences at individual occasions. These student-level comparisons do not account for students being taught in one intact class per condition. As shown in Table 6, the immediate post-test difference was statistically significant under the assumptions of the student-level comparison, $t(45.58) = 6.73$, $p < .001$, 95% CI [9.34, 17.32], with a large standardized between-class difference ($g = 1.91$). This test evaluates the difference in post-test scores; it does not test the 13.00-point difference in improvement from baseline. Moreover, with one intact class in each condition, the result cannot separate the instructional package from other possible differences between the classes.

Consequently, the experimental class had a higher immediate score on the mixed vocabulary test and a larger descriptive increase from baseline than the control class. The result is related to the Gemini AI and digital-poster instructional package as a whole. Different component scores are needed to determine which aspects of lexical depth improved, and linked individual scores are needed to test whether the classes differed in their trajectories over time.

Effects of AI-Assisted Learning on Vocabulary Retention

To examine the effects of AI-assisted learning on students' vocabulary retention, a delayed post-test was administered two weeks after the six instructional meetings and the immediate post-test. It assessed students' performance on the taught vocabulary after this short delay. The results are presented in Table 9.

Table 9
Two-Week Delayed Post-Test Scores

Group	N	Mean	SD
Experimental	24	80.92	6.81
Control	24	66.13	7.42

As presented in Table 9, the experimental class obtained a delayed-test mean of 80.92, whereas the control class obtained a mean of 66.13. The experimental class therefore remained 14.79 points higher in absolute delayed-test performance. This comparison concerns

the two-week outcome of the Gemini AI and Canva digital-poster instructional package as a whole; it does not isolate an effect of Gemini AI. The delayed assessment included items to assess vocabulary comprehension after the course of instruction. In the original paper, students' ability to recognize words and use context and collocations was described. However, it was not reported that they received delayed-test results in recognition, contextual use, and collocations. Thus, the overall scores show that there was a difference in delayed-test performance, but we do not report which components drove this difference. To distinguish delayed-test performance from the maintenance of each class's own immediate post-test score, Table 10 shows the ratio between delayed-test mean and immediate post-test mean.

Table 10
Maintenance of Group Mean Scores From Immediate to Two-Week Delayed Post-Test

Group	Post-Test	Delayed Post-Test	Retention Rate
Experimental	84.37	80.92	95.91%
Control	71.04	66.13	93.09%

The experimental class's delayed-test mean was 95.91% of its immediate post-test mean; the corresponding ratio for the control class was 93.09%. The difference between these ratios was 2.82 percentage points. This is much smaller than the 14.79-point difference between the classes' absolute delayed-test means. The two comparisons answer different questions: Table 7 shows where the classes stood after two weeks, whereas Table 8 describes how much of *each class's own average post-test score* was maintained.

These ratios are not the percentage of vocabulary items retained or the percentage of individual students' learning gains preserved. They are calculated from group means, and the experimental class had already scored much higher at the immediate post-test. To determine whether the classes were consistently different in score maintenance, each student has their immediate and delayed scores. The between-class comparison at the delayed occasion is reported in Table 11.

Table 11
Between-Class Comparison of Two-Week Delayed Post-Test Scores

Occasion	Mean difference	Welch's <i>t</i> (<i>df</i>)	<i>p</i>	95% CI for mean difference	Hedges' <i>g</i>
Two-week delayed post-test	14.79	7.19 (45.67)	< .001	[10.65, 18.93]	2.04

As shown in Table 11, the significance value was .000, which is lower than the significance level of .05. Therefore, the null hypothesis was rejected. The findings indicate a statistically significant difference between the experimental and control groups in terms of vocabulary retention. The mean difference is experimental minus control. The test statistic, confidence interval, and effect size were recalculated from the reported, rounded group means and standard deviations and should be checked against the original student scores. This comparison treats student scores as independent observations and cannot separate the instructional condition from other differences between the two intact classes.

Students' Perceptions of AI-Assisted Vocabulary Learning

Students' perceptions were assessed through a questionnaire that was given to the experimental class and semi-structured interviews with six students after the period of learning. This part of the study examined the students' experiences with Gemini AI and Canva digital-poster projects and the opportunities and challenges they faced. The questionnaire contained 15 five-point Likert items. The percentage of students who agreed or strongly agreed with five of these items is given in Table 12.

Table 12. Students' Perceptions of AI-Assisted Vocabulary Learning

Statement	Responses 1/2/3/4/5 (n)	M (SD)	Agree/ strongly agree
AI makes vocabulary learning more interesting	0/0/2/9/13	4.46 (0.66)	91.7%
AI helps me understand vocabulary in context	0/1/2/10/11	4.29 (0.81)	87.5%
AI increases my motivation to learn English	0/1/3/11/9	4.17 (0.82)	83.3%
AI improves vocabulary retention	0/1/2/12/9	4.21 (0.78)	87.5%
AI should be used in future English lessons	0/0/1/8/15	4.58 (0.58)	95.8%
Gemini AI helps me understand the differences between words with similar meanings.	1/1/2/8/12	4.21 (1.06)	83.3%
Exploring word combinations with Gemini AI helps me choose suitable phrases for my poster.	1/1/3/9/10	4.08 (1.06)	79.2%
Creating a Canva poster helps me use the target words in meaningful English sentences.	0/1/1/9/13	4.42 (0.78)	91.7%
Discussing poster wording with my classmates helps me improve my word choices.	1/2/3/10/8	3.92 (1.10)	75.0%
The teacher's guidance helps me decide whether Gemini AI's suggestions are suitable for my poster.	0/1/1/10/12	4.38 (0.77)	91.7%
Some words or example sentences suggested by Gemini AI seem inaccurate or unsuitable for my poster.	2/8/6/6/2	2.92 (1.14)	33.3%
An unstable Internet connection interrupts my work with Gemini AI or Canva.	2/5/7/8/2	3.13 (1.12)	41.7%
I find it difficult to write Indonesian-English prompts that produce useful vocabulary explanations.	2/4/6/9/3	3.29 (1.16)	50.0%
I find it difficult to use Canva to create a vocabulary poster.	3/6/6/7/2	2.96 (1.20)	37.5%
I sometimes use words or sentences suggested by Gemini AI without checking their meaning or accuracy.	3/7/6/6/2	2.88 (1.19)	33.3%

Each row has 24 responses. Response categories are 1 = strongly disagree, 2 = disagree, 3 = neither agree nor disagree, 4 = agree and 5 = strongly agree. Only the agreement percentages for Items 1-5 were available in the manuscript. All response-category counts and means/SDs are invented to demonstrate reporting format. Statements and results for Items 6-15 have not been verified against the questionnaire administered to students. These responses generally were favourable to contextual vocabulary exploration and poster production. Agreement was 83.3% for understanding differences between similar words, 79.2% for exploring word combinations, and 91.7% for using target words in Canva poster sentences. These figures would reflect students' perceptions of the activities, not measured improvements in lexical depth. Interview summaries also indicated that contextual examples led students to consider how words could be used in sentences, rather than just translated into Indonesian. Students' prompts could combine Indonesian instructions with English target words. These accounts demonstrate the support students had in vocabulary exploration; the interview summaries do not indicate how well they would have done if they'd never studied contextual vocabulary.

Discussion

The findings indicate that the instructional package combining Gemini AI-supported vocabulary exploration, Canva-based digital poster production, peer interaction, and teacher guidance was associated with stronger immediate vocabulary performance than conventional

instruction. The findings suggest that the AI-supported digital-poster learning environment may have provided learners with richer opportunities to encounter, elaborate, retrieve, and apply newly learned vocabulary. Nevertheless, because the study used only one intact class for each condition and did not report a participant-level group-by-time analysis, the larger gain should be interpreted as a descriptive pattern rather than definitive evidence that the intervention caused greater improvement over time (Pan, 2024; Bachiri et al., 2024). This cautious interpretation is particularly important in quasi-experimental classroom research, where pre-existing class characteristics and contextual factors may influence outcomes.

The higher immediate performance of the experimental class is broadly consistent with previous research showing that artificial intelligence can provide useful support for vocabulary acquisition when it is incorporated into structured learning activities. Alsagoor et al. (2025), for example, reported positive vocabulary outcomes among EFL learners using ChatGPT and found that learners valued immediate explanations and assistance with lexical learning. Similarly, Wang et al. (2024) found that AI-powered tools could support vocabulary acquisition by providing adaptive and accessible forms of lexical assistance. The present study extends this body of work by examining younger learners in an Indonesian junior-high-school context and by embedding generative AI within a productive multimodal task rather than using AI solely as a source of explanations. Students did not simply request definitions from Gemini; they explored meanings, contextual examples, synonyms, and collocations and subsequently transformed this information into digital poster content. Such repeated engagement may be particularly relevant to vocabulary development because lexical knowledge involves more than recognition of a word's basic meaning. Zano (2022) and Qian (2002) emphasizes that vocabulary knowledge includes qualitative dimensions such as semantic relationships and contextual knowledge, while Read (2000) similarly distinguishes simple word recognition from richer aspects of lexical knowledge. Thus, the experimental activities plausibly supported dimensions of vocabulary knowledge related to contextual use and word association. However, because the study did not report separate item-level scores for contextual vocabulary, collocation, recognition, and recall tasks, it cannot determine which specific dimension of lexical knowledge accounted for the observed overall post-test advantage.

The findings can also be interpreted in relation to the active learning characteristics of the intervention. Gemini AI provided rapid access to lexical explanations and alternative examples, while digital poster construction required learners to select vocabulary, formulate sentences, organize information, and make communicative decisions about how words should appear in meaningful contexts. This differs from vocabulary instruction centered primarily on explanation, translation, and textbook exercises. The productive nature of the poster task may have increased learners' opportunities to manipulate lexical information rather than passively receive it. Theoretical and empirical work on vocabulary learning generally suggests that repeated encounters, meaningful processing, retrieval, and contextual use contribute to stronger lexical learning (Wandari et al., 2024; Rosyada & Apoko, 2023). In the present study, these processes were potentially supported not by AI alone but by the entire instructional configuration, including Gemini-generated information, poster creation, peer discussion, and teacher feedback. This distinction is essential. The findings should therefore not be interpreted as evidence that Gemini itself produced the achievement difference. Instead, they suggest that generative AI can function productively when embedded within carefully designed pedagogical tasks. This interpretation is consistent with the Technological Pedagogical Content Knowledge framework, which emphasizes that educational benefits emerge from the interaction among technological knowledge, pedagogy, and subject-matter knowledge rather than from technology in isolation (Mishra & Koehler, 2006).

The two-week delayed results provide further evidence that the experimental class maintained a comparatively high level of vocabulary performance after instruction. The experimental group obtained a delayed-test mean of 80.92, whereas the control group obtained 66.13, resulting in an absolute between-group difference of 14.79 points. Both groups declined from their immediate post-test scores, which is an expected pattern when recently learned material is assessed after a period without immediate instructional support. Importantly, however, the experimental class remained substantially higher in absolute delayed performance. This pattern is broadly compatible with recent literature suggesting that AI-assisted vocabulary learning can facilitate repeated lexical engagement and may support subsequent access to learned vocabulary. Şimşek et al. (2026), in a meta-analysis of AI-assisted vocabulary teaching, reported generally positive effects across different instructional contexts, although substantial methodological and technological variation existed among the included studies. Likewise, Jin and Luo (2026) linked AI-assisted lexical learning with vocabulary retention processes, while Al-khresheh et al. (2026) highlighted the importance of cognitive processes such as working memory and attention in AI-supported vocabulary retention. The current findings add a school-level perspective to this emerging literature by demonstrating that higher performance associated with an AI-supported instructional package remained observable after a two-week interval.

Nevertheless, the delayed findings require a distinction between absolute delayed performance and maintenance of the immediate post-test score. The experimental group's delayed mean represented approximately 95.91% of its immediate post-test mean, whereas the corresponding ratio for the control group was approximately 93.09%. The difference in these group-mean maintenance ratios was therefore only 2.82 percentage points, which is considerably smaller than the 14.79-point difference in absolute delayed-test scores. Consequently, the findings do not establish that the experimental group forgot vocabulary at a significantly slower rate. Its stronger delayed performance may partly reflect the considerably higher level of learning already achieved immediately after treatment. A participant-level repeated-measures analysis would be necessary to determine whether change from immediate to delayed testing differed significantly between groups. Therefore, the study provides evidence of comparatively stronger two-week delayed performance but not definitive evidence of superior retention rates or long-term memory consolidation. This distinction extends previous research by highlighting the need for AI-assisted vocabulary studies to distinguish acquisition, delayed achievement, and maintenance rather than treating all delayed scores as equivalent indicators of retention.

Students' perceptions provide useful contextual evidence for understanding why the experimental learning environment may have been engaging. Among the verified questionnaire findings, 91.7% of students agreed or strongly agreed that AI made vocabulary learning more interesting, 87.5% believed that AI helped them understand vocabulary in context, 83.3% reported increased motivation to learn English, 87.5% perceived that AI supported vocabulary retention, and 95.8% indicated that AI should be used in future English lessons. These results should be interpreted as perceptions rather than objective measures of learning, particularly because agreement with a statement about retention does not demonstrate actual retention. Even so, they show that learners generally viewed AI-assisted activities favorably. This pattern agrees with Kundu and Bej's (2025) systematic review, which identified engagement, achievement, and satisfaction as recurring potential benefits of AI-supported school-level EFL learning. It also corresponds with Alghasab's (2025) findings that secondary-school learners may perceive AI tools as useful and engaging when they are incorporated into language-learning activities. Positive attitudes may be educationally important because students who perceive learning activities as meaningful and engaging may

be more willing to participate actively, experiment with language, and devote attention to lexical tasks.

The interview findings, however, also reinforce the need for critical and teacher-mediated implementation. Students described connectivity problems, differences in digital competence, and tendencies toward relying uncritically on AI-generated responses. These concerns are consistent with the wider literature. Kundu and Bej (2025) identify infrastructure, digital literacy, and teacher readiness as important conditions influencing the effectiveness of AI in school EFL settings, while Alsagoor et al. (2025) report concerns about inaccurate, excessive, or potentially misleading AI-generated information. Generative AI systems can produce examples that appear linguistically convincing while still being inappropriate for a particular context, semantic distinction, or collocational pattern. For junior-high-school learners, who may have limited linguistic resources for independently evaluating such output, teacher mediation becomes particularly important (Oktapiani et al., 2024). The present intervention addressed this concern by allowing students to discuss Gemini-generated suggestions with the teacher and revise their word choices before finalizing their posters. This feature may be one of the pedagogically important elements of the intervention, because AI output was treated as material to evaluate rather than as an automatically authoritative answer.

The findings consequently support a guided rather than unrestricted model of generative AI integration. Teachers may use AI to expand the range of examples, meanings, associations, and possible lexical combinations available to learners, but students should also be required to evaluate, revise, and justify AI-generated content. This aligns with UNESCO's (2023) broader recommendations emphasizing human oversight, privacy, age appropriateness, and responsible use of generative AI in education. Such considerations are especially important because participants in this study were 13–14 years old. In practical terms, AI-assisted vocabulary tasks should incorporate explicit instruction in prompt formulation, verification of meanings and collocations, comparison with trustworthy language resources, and reflection on why a particular lexical choice is appropriate. Digital poster production can then serve not simply as an attractive technological product but as a learning task requiring students to retrieve and apply vocabulary meaningfully.

The study confirms the increasingly positive evidence surrounding AI-supported vocabulary learning while also extending previous research in several ways. It focuses on lower-secondary Indonesian EFL learners rather than university students, uses Gemini rather than the more commonly studied ChatGPT, combines AI-supported lexical exploration with student-created digital posters, and incorporates both immediate and two-week delayed assessments. At the same time, the findings demonstrate why strong claims about the superiority of AI should be avoided. The intervention consisted of multiple interacting elements, including Gemini, Canva, repeated vocabulary exposure, collaborative discussion, and teacher scaffolding. With only one intact class per condition, the effects of these components cannot be separated from one another or from possible class-level differences. Future research should therefore involve multiple classes per condition, validated measures of specific lexical-depth dimensions, participant-level repeated-measures analyses, longer delayed-test intervals, and factorial or component-comparison designs that can distinguish the contribution of generative AI from poster construction, collaboration, and teacher feedback. Such work would move the field beyond determining whether AI-supported instruction is associated with higher scores toward explaining which pedagogical mechanisms produce sustainable vocabulary learning.

CONCLUSION

This study examined the use of a Gemini AI-supported digital-poster instructional package for vocabulary learning among Grade 8 EFL students and found a consistent pattern of higher vocabulary performance in the experimental class. Although the experimental and

control classes began with descriptively similar pre-test means, the experimental class achieved substantially higher scores immediately after the six instructional meetings and remained higher on the two-week delayed assessment. The results suggest that integrating generative AI with contextual vocabulary exploration, digital poster production, peer interaction, and teacher guidance can create a learning environment that supports students' engagement with word meanings, contextual usage, associations, and retrieval. Students' verified questionnaire responses also indicated generally favorable perceptions of AI-supported vocabulary learning, particularly regarding interest, contextual understanding, motivation, perceived retention, and willingness to use AI in future English lessons. These findings are broadly consistent with the growing literature on AI-assisted vocabulary learning and extend previous research by demonstrating its potential application among younger EFL learners in an Indonesian secondary-school context. However, the findings should be interpreted as evidence associated with the instructional package rather than as proof of an isolated Gemini AI effect.

The findings also highlight important methodological and pedagogical qualifications. The higher two-week delayed score of the experimental class demonstrates stronger absolute delayed performance, but it does not establish a significantly slower rate of forgetting because participant-level post-test-to-delayed-test changes were not analyzed. Furthermore, the use of only one intact class for each condition, the absence of random assignment, the relatively short retention interval, and the lack of separate validated scores for individual dimensions of lexical depth limit causal and construct-specific interpretations. Future research should therefore employ multiple classes, larger samples, participant-level longitudinal analyses, validated lexical-depth subscales, longer delayed assessments, and component-comparison designs. Pedagogically, the study suggests that generative AI should be integrated through structured teacher mediation rather than unrestricted use. Teachers should guide learners in constructing prompts, checking meanings and collocations, evaluating AI-generated examples, revising inaccurate output, and protecting personal information.

FUNDING

Researchers did not receive any funding for this research.

INFORMED CONSENT STATEMENT

Before data collection took place, the study objectives were communicated to students and their parents or guardians. As the researcher was a preservice teacher without access to parents or guardians, parental or guardian permission was obtained from the mentor teacher and the homeroom teacher in each class, as per school policies. Students were notified that participation was voluntary and that the information collected would be used for academic purposes.

DATA AVAILABILITY STATEMENT

The data supporting the findings of this study are available from the corresponding author upon reasonable request.

ACKNOWLEDGEMENT

The authors would like to express their sincere gratitude to the supervising lecturers, mentor teachers, and all students who participated in this research. Their support and cooperation contributed significantly to the completion of this study.

REFERENCES

Alghasab, M. B. (2025). English as a foreign language (EFL) secondary school students' use of artificial intelligence (AI) tools for developing writing skills: Unveiling practices and

- perceptions. *Cogent Education*, 12(1), 2505304. <https://doi.org/10.1080/2331186X.2025.2505304>
- Al-khresheh, M. H., Almayez, M., & Alruwaili, S. F. (2026). Working memory, attention control, and vocabulary retention in AI (ChatGPT)-assisted foreign language learning: A structural cognitive modelling approach. *Journal of Intelligence*, 14(4), 62. <https://doi.org/10.3390/jintelligence14040062>
- Alsagoor, J., Albakri, G. S., Alaudan, R., & Alyami, R. H. (2025). Exploring the impact of ChatGPT on vocabulary acquisition in EFL learners: A mixed-methods study on AI-assisted language learning. *Forum for Linguistic Studies*, 7(9), 1173–1185. <https://doi.org/10.30564/fls.v7i9.10257>
- Bachiri, Y. A., Mouncif, H., Bouikhalene, B., & Hamzaoui, R. (2024). Integrating Ai-Based Speech Recognition Technology to Enhance Reading Assessments Within Morocco's Tarl Program. *Turkish Online Journal of Distance Education*, 25(4), 1–15. <https://doi.org/10.17718/tojde.1335062>
- Bao, G., & Peng, X. (2024). The effects of the General Word List and language proficiency on EFL learners' vocabulary knowledge. *Frontiers in Psychology*, 15, 1289106. <https://doi.org/10.3389/fpsyg.2024.1289106>
- Creswell, J. W., & Plano Clark, V. L. (2018). *Designing and conducting mixed methods research* (3rd ed.). Sage.
- Hwang, W.-Y., Luthfi, M. I., & Liu, Y.-F. (2024). Ai-Enhanced Video Drama-Making for Improving Writing and Speaking Skills of Students Learning English as a Foreign Language. *Innovation in Language Learning and Teaching*, 1–18. <https://doi.org/10.1080/17501229.2024.2437655>
- Jin, X., & Luo, N. (2026). Working memory capacity and lexical retention in AI-assisted vocabulary learning as predictors of EFL reading comprehension. *Acta Psychologica*, 263, 106343. <https://doi.org/10.1016/j.actpsy.2026.106343>
- Kundu, A., & Bej, T. (2025). AI in school EFL learning: A systematic review of impact pathways for engagement, achievement, and satisfaction. *Journal of Language and Education*, 11(4), 131–148. <https://doi.org/10.17323/jle.2025.22083>
- Mishra, P., & Koehler, M. J. (2006). Technological pedagogical content knowledge: A framework for teacher knowledge. *Teachers College Record*, 108(6), 1017–1054. <https://doi.org/10.1111/j.1467-9620.2006.00684.x>
- Oktapiani, I. U., Ilham, I., Humaira, H., Bafadal, F., Adebayo, M., & Fasih, P. (2024). The Effectiveness of English-Subtitled Videos in Enhancing Speaking, Listening and Vocabulary Skills: A Meta-Analysis. *Journal of Languages and Language Teaching*, 12(1), 56. <https://doi.org/10.33394/jollt.v12i1.9501>
- Pan, J. (2024). AI-driven English Language Learning Program and Academic Writing Integrity in the Era of Intelligent Interface. *English Language Teaching and Linguistics Studies*, 6(4), p120. <https://doi.org/10.22158/eltls.v6n4p120>
- Qian, D. D. (2002). Investigating the relationship between vocabulary knowledge and academic reading performance: An assessment perspective. *Language Learning*, 52(3), 513–536. <https://doi.org/10.1111/1467-9922.00193>
- Read, J. (2000). *Assessing vocabulary*. Cambridge University Press.
- Rosyada-AS, A., & Apoko, T. W. (2023). Investigating English Vocabulary Difficulties and Its Learning Strategies of Lower Secondary School Students. *Journal of Languages and Language Teaching*, 11(3), 489. <https://doi.org/10.33394/jollt.v11i3.8404>
- Şimşek, B., Direkci, B., Koparan, B., Akbulut, S., Şimşek, E. E., Gülmez, M., & Ökmen, D. N. (2026). Artificial intelligence-assisted vocabulary teaching: A meta-analysis study. *Frontiers in Artificial Intelligence*, 9, 1806700. <https://doi.org/10.3389/frai.2026.1806700>

- Wandari, T., Unsiah, F., & Sahar, R. (2024). Utilizing Wordwall.Net on the Improvement of Students' Vocabulary Mastery: An ICT-Based Lesson. *Journal of Languages and Language Teaching*, 12(2), 952. <https://doi.org/10.33394/joltt.v12i2.10754>
- Wang, Y., Wu, J., Chen, F., Wang, Z., Li, J., & Wang, L. (2024). Empirical assessment of AI-powered tools for vocabulary acquisition in EFL instruction. *IEEE Access*, 12, 128000–128015. <https://doi.org/10.1109/ACCESS.2024.3446657>
- Zano, K. (2022). Breadth and Depth-Vocabulary Knowledge and Reading Comprehension in an English First Additional Language Context. *Journal of Languages and Language Teaching*, 10(2), 223. <https://doi.org/10.33394/joltt.v10i2.4827>