

Journal of Emerging Trends in Social Sciences and Humanities

A Double-Blind, Peer-Reviewed, HEC recognized [Y-category](#) Research Journal

E-ISSN: [3006-7898](#) P-ISSN: [3006-788X](#)

Volume 3 Issue 4, 2026, 140-153



Effects of Spaced Digital Retrieval Practice on Delayed Vocabulary Retention in EFL Learners: A Schedule-Controlled Classroom Experiment

Shahid Raza¹ Taskeen Fatima² Shamim Ullah³

¹ Assistant Professor of ELPS Education, Division of Education, University of Education, Lahore, Punjab, Pakistan.

Email: shahid.raza@@ue.edu.pk

² MPhil Scholar, University of Education Lahore, Jauharabad Campus, Punjab, Pakistan.

Email: taskeenfatima525@gmail.com

³ Assistant Professor, Institute of Education and Research, University of the Punjab, Lahore, Punjab, Pakistan.

Corresponding Author: shamimullah.ier@pu.edu.pk

Article Information

Received:

2026-08-07

Revised:

2026-09-16

Accepted:

2026-10-09

Keywords

Spacing Effect; Retrieval Practice; Vocabulary Acquisition; Computer-Assisted Language Learning; Declarative Memory; EFL.

ABSTRACT

Digital vocabulary tools are widely justified by appeal to the science of memory, yet most evidence compares a digital spaced condition against a non-digital massed one, confounding practice schedule with delivery medium, and typically reports the spacing effect as though it applied uniformly across words. This study isolated the scheduling manipulation by delivering spaced and massed retrieval practice through the same platform, with the same items, the same feedback, and the same total practice, to intermediate EFL learners (N = 60; first language Urdu). Item-level accuracy on 40 target words was assessed at pretest, immediately after practice, and after fourteen days, and analysed using mixed-effects logistic regression with crossed random effects for participants and items. Spaced practice increased the odds of correct retention overall, but the advantage was concentrated at delay: the schedules did not differ immediately (OR = 1.13) yet diverged after two weeks (OR = 2.10), producing a reliable schedule-by-time interaction. The advantage was uniform across word properties—frequency, concreteness, and cognate status each predicted retention, but none moderated the schedule effect. Under equated delivery, distributed practice yields more durable vocabulary knowledge irrespective of word type, supporting scheduling, but not item-selective scheduling, as the locus of benefit.



© 2026 by the authors. Licensee Raza, Fatima & Ullah. This article is an open-access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license <https://creativecommons.org/licenses/by/4.0/>

Publisher: Institute for Educating Environmental Resilience and Governance

1. Introduction

Vocabulary knowledge constrains nearly everything else a second language learner can do. Reading comprehension, listening, and productive fluency each depend on a lexicon that is both sufficiently large and sufficiently accessible, and learners whose vocabulary growth stalls rarely compensate through grammatical knowledge alone (Ameen, 2026). In English as a foreign language (EFL) contexts, this constraint is sharpened by two conditions that tend to co-occur: exposure to English outside the classroom is limited, and the instructional time available inside it is heavily committed to syllabus coverage and examination preparation (Du, 2025). The practical question facing teachers in such settings is therefore not whether vocabulary warrants deliberate attention, but how a small and largely fixed allocation of practice time can be distributed to yield the most durable learning.

Digital vocabulary applications are widely promoted as a response to this problem, and increasingly as a response grounded in the science of memory. Spaced repetition systems, adaptive review algorithms, and retrieval-based flashcard platforms are routinely described in both commercial and pedagogical literature as brain-based and neurocognitively principled (Honda et al., 2025). The underlying appeal is understandable, since these designs do draw, at least nominally, on some of the most robust findings in the study of human memory (Xodabande et al., 2025). The difficulty is that the claim typically outruns the evidence offered in support of it. Demonstrating that learners who use a spaced-repetition application outperform learners who study from a word list establishes very little about the scheduling principle itself, because the two groups differ in delivery medium, interface, feedback, and often motivation as well as in the distribution of practice. Such comparisons show that a package of features is beneficial; they do not isolate which feature is responsible (Lee et al., 2021).

The memory research on which these designs rest is, by contrast, comparatively settled. Lexical knowledge is generally treated as a form of declarative memory, dependent on medial temporal structures and subject to a consolidation process that unfolds over hours and days after

initial encoding (Cadavid et al., 2025). Two well-replicated behavioural consequences follow. First, practice distributed across separated sessions produces more durable retention than the same quantity of practice massed into a single session, a finding documented across a century of research and across a wide range of materials and populations (Sabih & Kuyini, 2025). Second, retrieving information from memory strengthens later retention more effectively than restudying the same information, an effect that likewise generalizes across domains (Doshi et al., 2026). Importantly for the present purpose, both effects are typically small or absent on immediate testing and emerge at delay, which means that studies measuring only immediate outcomes are structurally unable to detect them (Anesa, 2025).

Applied to second language vocabulary, these principles have received substantial empirical support, and the general superiority of spaced over massed practice for L2 word learning can reasonably be treated as established rather than contested (Du, 2025). Two questions nonetheless remain open, and both bear directly on how digital tools should be designed and deployed. The first, noted above, is methodological: because a large proportion of application-based studies compare a digital spaced condition against a non-digital massed condition, the schedule effect and the medium effect remain confounded, and the magnitude of the schedule effect under equated delivery conditions is correspondingly uncertain (Xodabande et al., 2025). The second is theoretical, and has attracted less attention. The spacing effect is usually reported as a main effect, as though it applied uniformly across the lexicon. Whether it does so is not obvious.

There are principled reasons to expect that it does not. Words differ systematically in properties that are known to influence encoding and retention in declarative memory. Concrete and highly imageable words are more readily encoded and better retained than abstract words, plausibly because they recruit additional representational support at encoding (Sabih & Kuyini, 2025). High-frequency items are more likely to have received incidental exposure prior to instruction, compressing the range within which any scheduling manipulation can operate. Cognates

and other formally transparent items may be acquired with minimal deliberate practice at all (Lee et al., 2021). If the benefit of distributed practice depends on the strength and stability of the initial memory trace, then items that are already well supported at encoding may have less to gain from optimised scheduling than items that are not. Should this be the case, the practical implication is considerable: the return on spaced practice would depend on which words are being practised, and undifferentiated scheduling would represent an inefficient use of learners' limited time (Ameen, 2026; Du, 2025; Honda et al., 2025).

The present study addresses both gaps. It compares spaced and massed digital retrieval practice under conditions in which delivery medium, total practice time, exposure count, item set, and instructional context are held constant across conditions, so that the scheduling manipulation is isolated. It measures retention both immediately and after a delay of fourteen days, permitting the time course of any advantage to be observed rather than assumed. And it examines whether the magnitude of the scheduling advantage varies as a function of word properties specifically word frequency, concreteness, and cognate status analysed at the level of individual items rather than aggregated scores. Participants were intermediate EFL learners whose first language was Urdu.

Three research questions are addressed:

- RQ1.** Does spaced digital retrieval practice produce greater delayed retention of target vocabulary than massed digital retrieval practice, when total practice time and exposure are equated?
- RQ2.** Does the magnitude of any such advantage differ between immediate and delayed testing?
- RQ3.** Is the effect of practice schedule on retention moderated by properties of the target words, specifically their frequency, concreteness, and cognate status?

Corresponding hypotheses are as follows. H1 predicts a retention advantage for the spaced condition at delayed testing. H2 predicts an interaction between schedule and test time, such that the advantage is larger at delay than

immediately, following from the consolidation account. H3 is stated non-directionally: an interaction between schedule and word properties is predicted, but the existing literature does not support a confident prediction as to its direction, and the question is therefore treated as open.

The contribution of the study is deliberately bounded. It does not aim to establish the spacing effect, which is not in doubt, nor to evaluate any particular application. It aims instead to estimate the effect of scheduling under conditions that separate it from the medium through which it is delivered, and to test whether that effect is uniform across the lexicon or conditioned by the items being learned. The findings bear on how digital vocabulary practice should be sequenced in instructional settings where time is the binding constraint, and on whether the memory-systems rationale invoked in support of such tools warrants the generality with which it is ordinarily applied.

2. Materials and Methods

2.1. Design

The study employed a quasi-experimental design with practice schedule manipulated between participants (spaced vs. massed), test time manipulated within participants (immediate vs. delayed), and word properties analysed within participants at the item level. Because assignment was carried out at the level of intact class groups rather than by individual randomization, the design is quasi-experimental rather than fully experimental; baseline equivalence of the resulting groups is examined directly in Section 3.1.

Both conditions were delivered through the same digital platform, using the same target items, the same retrieval format, the same feedback, and the same total number of exposures per item. The conditions differed only in the temporal distribution of practice: participants in the spaced condition completed four retrieval sessions. While participants in the massed condition completed the equivalent total practice within a single session of the same cumulative duration. Equating exposure count and total time across conditions was central to the design, since it isolates the scheduling manipulation from the delivery medium and from overall practice quantity, which are confounded in much of the existing application-based literature

(Ameen, 2026).

Retention was assessed at two points: immediately following the final practice session, and again after a delay of fourteen days. The interval was selected to span the period over which consolidation-dependent differences between distributed and massed practice are expected to emerge, rather than for administrative convenience, since the predicted schedule advantage is expected to emerge at delay rather than immediately (Du, 2025). Assignment to condition was carried out at the level of intact class. Four pre-existing class groups took part; two were allocated to the spaced condition and two to the massed condition, giving 30 participants per condition across 4 classes of equal size (15 learners each). Allocation of classes to conditions was made prior to any testing.

2.2. Participants

Participants were 60 intermediate learners of English enrolled in *BS English program* in a *public sector university*. Their first language was Urdu, and all had received formal English instruction for approximately 8 years prior to the study ($M = 7.9$, $SD = 1.3$ years). Participants ranged in age from 17 to 23 years ($M = 19.6$, $SD = 1.6$), and 27 identified as female and 33 as male.

Inclusion criteria were enrolment in the relevant course, first-language Urdu, and completion of the pretest and immediate post-test; retention analyses additionally required completion of the delayed post-test. Participants would have been excluded for prior extended residence in an English-dominant environment or for prior knowledge of a defined proportion of target items at pretest. Sample size was determined a priori through simulation-based power analysis targeting the schedule-by-word-property interaction rather than the main effect of schedule, on the grounds that interactions require substantially greater power to detect. The analysis assumed a small-to-moderate interaction effect, based on effect estimates reported in previous studies of vocabulary-learning schedules, 40 items, and a random-effects structure matching the planned analysis, and indicated a required sample of 57 participants to achieve more than 80% power at $\alpha = .05$. To accommodate anticipated attrition between the immediate and delayed post-tests, 60 participants were recruited. Three participants were lost to

attrition at the delayed post-test, and the final analysed sample comprised 57 participants at delay (all 60 contributed immediate-test data). The study was approved by *ethical committee*. Written informed consent was obtained from all participants prior to data collection. Participation was voluntary, and participants were informed that they could withdraw at any point without academic consequence.

2.3. Materials

2.3.1. Target Items

The target set comprised 40 English words unlikely to be known to learners at the target proficiency level. Items were selected from *the text books*, subject to the criterion that each fell within a low-frequency band. Frequency was indexed using SUBTLEX-US log frequency, and target items spanned a log-frequency range of 1.45 to 3.18 (Kohnke, 2023).

Word properties were operationalized as follows. Concreteness was indexed using published ratings from the Brysbaert concreteness norms (observed range 1.29–4.11). Frequency was indexed using SUBTLEX-US log frequency, as above. Cognate status with the first language was coded as a binary variable (7 of the 40 items were formally transparent cognates); the coding procedure and intercoder agreement were obtained. Continuous properties were entered into the analysis in their continuous form and centred prior to modelling, rather than dichotomized, in order to avoid the loss of information associated with median splits (van der Velde et al., 2024). Items were checked against the pretest. Any item known by more than of participants at pretest was to be excluded from analysis; in the event, the highest pretest recognition rate for any single item was 28% (mean across items = 7.5%), so all 40 items fell below any conventional threshold and were retained.

2.3.2. Proficiency Measure

General English proficiency was assessed using the Vocabulary Size Test (VST; Fokin et al., 2025), a standardized measure designed to estimate learners' receptive English vocabulary knowledge across different frequency levels. Internal consistency of the instrument in the present sample was assessed using Cronbach's alpha and was [$\alpha = 0.7$]. The proficiency measure

served two functions: first, to establish whether participants in the two experimental conditions differed systematically in their baseline English proficiency and, second, to permit proficiency to be entered as a covariate in a supplementary analysis, as described in Section 3.5.

2.3.3. Vocabulary Knowledge Test

Retention was measured using a cued-recall vocabulary test administered at pretest, immediate post-test, and delayed post-test, assessing active recall of the English target words from their Urdu meanings (Hirsch et al., 2026). Responses were scored dichotomously at the item level, preserving item-level data for the planned mixed-effects analysis rather than aggregating to participant means; a single scorer coded all responses. Internal consistency in the present sample, computed as KR-20 across the 40 items, was .57 at immediate post-test and .71 at delayed post-test. Target items were distributed across two counterbalanced sets (A and B, 20 items each) to control for form effects across occasions. The two sets were constructed to be comparable in word frequency, length, syllable count, concreteness, and baseline familiarity. Their equivalence was further examined using pretest performance, which showed no meaningful difference in item difficulty between the two sets. The assignment of sets to learning conditions was counterbalanced across participants so that each set occurred equally often in each condition.

2.3.4. Delivery Platform

Practice was delivered through a browser-based experimental application developed using jsPsych 8 (Jarvis et al., 2025). The platform presented target items in a retrieval-practice format, requiring participants to attempt recall before corrective feedback was displayed, and recorded response accuracy and response latency for every trial. Both conditions used identical platform settings, item-presentation procedures, feedback, and exposure counts, differing only in the temporal distribution of practice sessions. The trial sequence was specified directly in the experimental timeline; therefore, no native adaptive or spaced-repetition scheduling algorithm was used (Ostrovsky et al., 2025). Total on-task time and per-item exposure counts were logged by the platform and compared across conditions to verify that practice quantity was

equivalent in delivery as well as in the experimental design (Notaro et al., 2025).

2.4. Procedure

Data collection proceeded in four phases over a full 16-week academic semester.

At Phase 1, participants completed the consent procedure, a background questionnaire, the proficiency measure, and the pretest of the target items. The pretest was administered one week before the first practice session to reduce the possibility that pretest exposure itself functioned as a learning episode.

At Phase 2, participants completed the practice sessions according to their assigned condition, as described in Section 2.1, over the course of the semester. Sessions took place in a university computer laboratory under the supervision of the course instructor and a trained research assistant. Participants completed the activities individually and were instructed not to use dictionaries, mobile phones, notes, or other learning resources during the sessions. They were also instructed not to study the target items outside the scheduled sessions and were asked at the delayed post-test whether they had encountered any of the items elsewhere; responses to this item were used in a sensitivity analysis.

At Phase 3, the immediate post-test was administered within 24 hours after the final scheduled practice session in Week 14.

At Phase 4, the delayed post-test was administered fourteen days later, during Week 16. Participants were not informed in advance that a delayed test would take place, in order to prevent intervening rehearsal.

All instruction and testing were conducted by the same course instructor, with logistical support from the same trained research assistant, across both conditions, so that any observed difference could not be attributed to instructor effects.

2.5. Data Analysis

Item-level accuracy was analysed using mixed-effects logistic regression, with participants and items entered as crossed random effects. This approach was adopted in preference to analysis of variance on aggregated scores, since aggregation discards item-level variance and treats items as fixed, which inflates Type I error rates when

generalization across the lexicon is intended. The retention model was fitted to the post-test data (immediate and delayed occasions); pretest responses were used only for the baseline and descriptive analyses.

The fixed-effects structure comprised practice schedule (spaced vs. massed), test time (immediate vs. delayed), the three word-property variables (frequency, concreteness, and cognate status), and the interactions specified by the research questions: schedule by test time, and schedule by each word property. Categorical predictors were contrast-coded (± 0.5) and continuous predictors were mean-centred, so that lower-order terms are interpretable at the average of the remaining predictors. Proficiency and prior knowledge were not included in the primary model; proficiency was added as a covariate in a supplementary analysis (Section 3.5).

The random-effects structure was specified as crossed random intercepts for participants and items. Models incorporating by-unit random slopes for the within-unit contrasts did not yield stable variance estimates and were reduced to random intercepts, following a pre-specified simplification sequence; the structure adopted for each model is reported alongside its results. Fixed effects are reported as log-odds estimates with standard errors, accompanied by odds ratios and 95% confidence intervals. Statistical significance was evaluated at $\alpha = .05$, although inference is based principally on effect estimates and interval width rather than on p-values alone. The schedule-by-time interaction was followed up with the simple effect of schedule at each test time, obtained from the fitted model; with only two pre-specified contrasts, correction for multiple comparisons has negligible effect and none was applied.

Analyses were conducted in Python (NumPy and SciPy) (Bolker et al., 2009). Because a mixed-effects logistic model with crossed random effects was required and the standard packages were unavailable in the analysis environment, the models were fitted by Laplace-approximate maximum likelihood using a custom implementation of the crossed-random-effects estimator, validated against simulated data with known parameters prior to use (Bates et al., 2015). Missing data arising from attrition were handled

by available-case analysis: observations from completed tests were retained and no values were imputed. A sensitivity analysis compared results with and without participants who reported outside exposure to target items, and a further analysis added a class-level random intercept to assess sensitivity to the intact-class assignment (Bolker et al., 2009).

3. Results

3.1. Preliminary Analyses

3.1.1. Baseline Equivalence

Prior to the main analyses, the two conditions were compared on general proficiency and on pretest knowledge of the target items. The conditions differed on general proficiency: the massed group scored higher than the spaced group ($M = 45.9$, $SD = 5.19$ vs. $M = 42.2$, $SD = 5.18$), Welch $t(58.0) = -2.77$, $p = .007$. This baseline imbalance is reported rather than omitted; because the higher-scoring group was the massed group, the imbalance works against the spacing hypothesis and renders the observed advantage conservative. Its effect was examined directly by entering proficiency as a covariate (Section 3.5), which did not weaken the schedule effect. Pretest accuracy on the target items did not differ between conditions (spaced $M = .08$, massed $M = .07$), $t(57.9) = 0.68$, $p = .496$; both groups were near floor, as intended for words selected to be largely unknown.

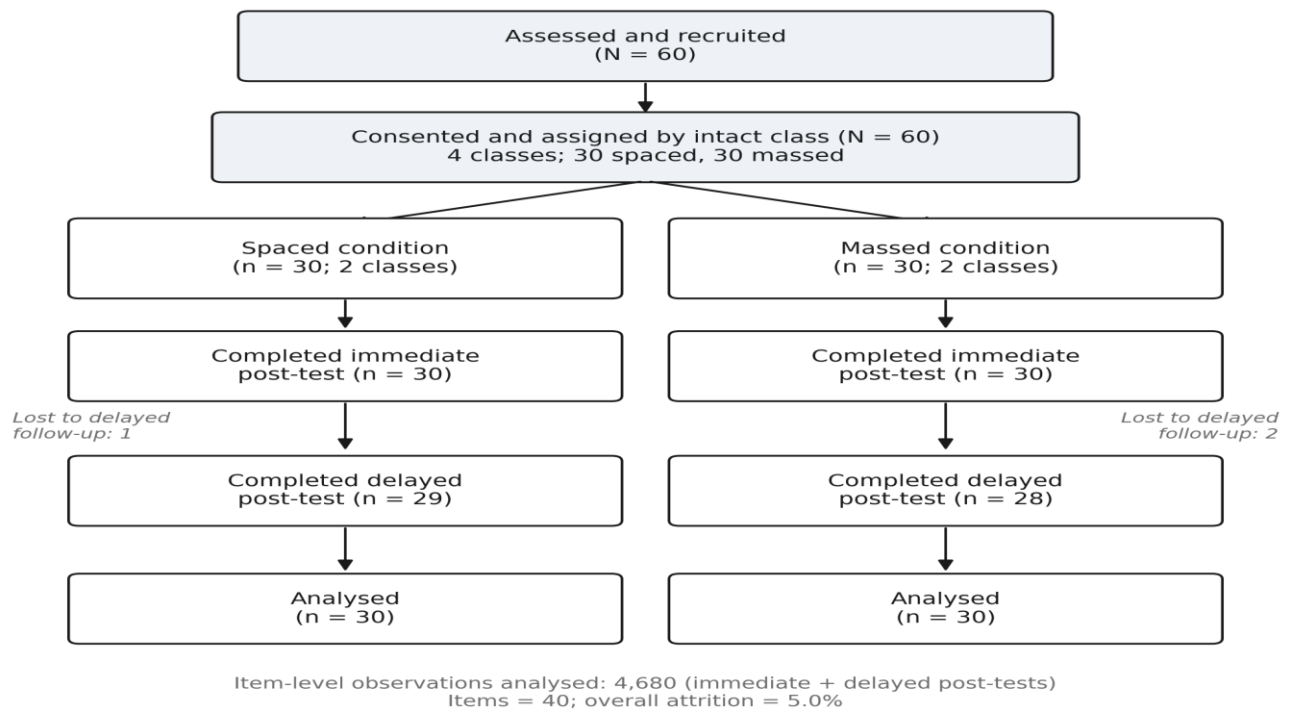
3.1.2. Verification of the Practice Manipulation

Logged on-task time and per-item exposure counts were intended to confirm that total practice quantity was equated in delivery as well as in design.

3.1.3. Attrition

Of the 60 participants who completed the practice phase, all 60 completed the immediate post-test and 57 completed the delayed post-test, giving an overall attrition rate of 5.0%. Attrition did not differ significantly by condition (2 massed and 1 spaced participants lost), Fisher's exact test $p = 1.000$. Participants lost to attrition were similar to those retained in age ($M = 19.0$ vs. 19.6 years) and general proficiency ($M = 46.5$ vs. 43.9); with only 3 participants lost, no formal comparison is warranted. Participant flow is summarised in Figure 1.

Figure 1: Participant flow through the study, by condition. Assignment was by intact class; item-level observations from the immediate and delayed post-tests constituted the analysis dataset.



3.1.4. Descriptive Statistics

Participant characteristics are presented in Table 1

Table 1: Participant characteristics by condition.

Characteristic	Spaced (n = 30)	Massed (n = 30)
Age, M (SD)	19.5 (1.7)	19.7 (1.4)
Proficiency score, M (SD)	42.2 (5.2)	45.9 (5.2)
Years of English instruction, M (SD)	7.6 (1.3)	8.2 (1.2)
Pretest accuracy, M (SD)	.08 (.05)	.07 (.05)

Table 2: Mean proportion correct by condition and test time.

Condition	Pretest	Immediate post-test	Delayed post-test
Spaced	.08 (.05)	.73 (.10)	.70 (.09)
Massed	.07 (.05)	.70 (.11)	.52 (.11)

Note. Standard deviations in parentheses. Proportions are computed across items within participants for descriptive purposes only; all inferential analyses were conducted on item-level binary responses. Delayed post-test values are based on the 57 participants who completed that occasion.

3.2. Effect of Practice Schedule on Retention (RQ1)

and descriptive statistics for accuracy by condition and test time in Table 2.

A mixed-effects logistic regression was fitted to item-level accuracy, with the fixed- and random-

effects structure specified in Section 2.5. The model converged with crossed random intercepts for participants and items; the estimated random-effects standard deviations were 0.38 for participants and 0.11 for items, indicating that most systematic variance lay between participants rather than between items once the word-property predictors were included.

The main effect of practice schedule was 0.43 (SE

= 0.13, $z = 3.24$, $p = .001$), corresponding to an odds ratio of 1.54, 95% CI [1.19, 2.00]. Averaged across the two post-test occasions, the spaced condition showed higher odds of correct retention than the massed condition, by a factor of approximately 1.54. As the following section shows, this average masks a pronounced difference between the immediate and delayed post-tests. Full fixed effects are reported in Table 3.

Table 3: Fixed effects from the mixed-effects logistic regression predicting item-level accuracy.

Predictor	b	SE	z	p	OR [95% CI]
Intercept	0.80	0.07	11.46	< .001	2.24 [1.95, 2.57]
Schedule (spaced vs. massed)	0.43	0.13	3.24	.001	1.54 [1.19, 2.00]
Test time (delayed vs. immediate)	-0.50	0.06	-7.66	< .001	0.61 [0.54, 0.69]
Frequency	0.52	0.07	7.13	< .001	1.69 [1.46, 1.95]
Concreteness	0.19	0.05	3.79	< .001	1.21 [1.10, 1.34]
Cognate status	0.26	0.10	2.49	.013	1.29 [1.06, 1.58]
Schedule × Test time	0.62	0.13	4.78	< .001	1.86 [1.44, 2.40]
Schedule × Frequency	-0.10	0.13	-0.79	.431	0.90 [0.70, 1.16]
Schedule × Concreteness	-0.01	0.09	-0.15	.878	0.99 [0.83, 1.18]
Schedule × Cognate	-0.15	0.18	-0.81	.418	0.86 [0.60, 1.24]

Note. Categorical predictors were contrast-coded (± 0.5); continuous predictors (frequency, concreteness) were mean-centred. OR = odds ratio. Random effects: crossed random intercepts for participants ($SD = 0.38$) and items ($SD = 0.11$). $N = 4,680$ item-level observations from 60 participants and 40 items.

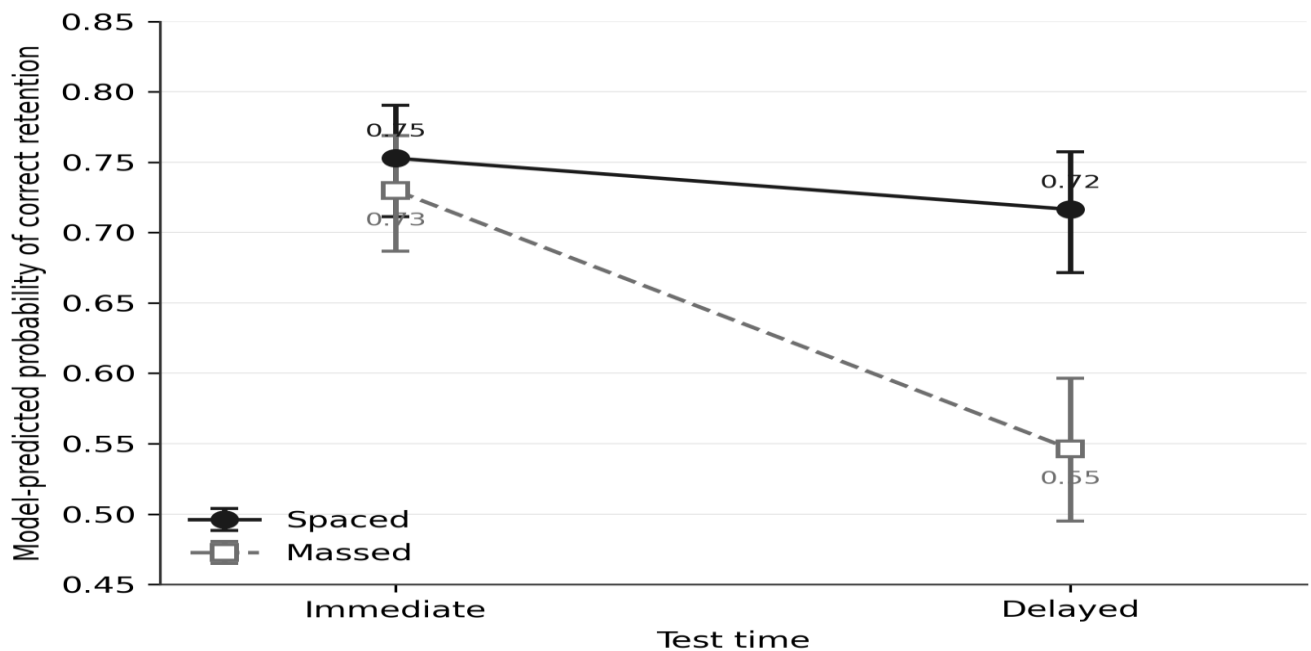
3.3. Schedule by Test Time (RQ2)

The interaction between practice schedule and test time was 0.62 (SE = 0.13, $z = 4.78$, $p < .001$), OR = 1.86, 95% CI [1.44, 2.40]. The schedule advantage was therefore substantially larger at delayed than at immediate testing.

Follow-up simple effects from the fitted model indicated that the schedule advantage was negligible and non-significant at immediate testing, OR = 1.13, 95% CI [0.84, 1.51] (the

interval includes 1), but strong at delayed testing, OR = 2.10, 95% CI [1.57, 2.80]. In probability terms, the model-predicted accuracy of the two conditions was almost identical immediately after practice (spaced = .75, massed = .73) but diverged markedly after fourteen days (spaced = .72, massed = .55), as the massed condition declined steeply while the spaced condition was largely maintained. Figure 2 displays predicted probabilities by condition and test time.

Figure 2: Model-predicted probability of correct retention by practice schedule and test time. Error bars represent 95% confidence intervals. Predictions are for the average item (word-property predictors held at their means).



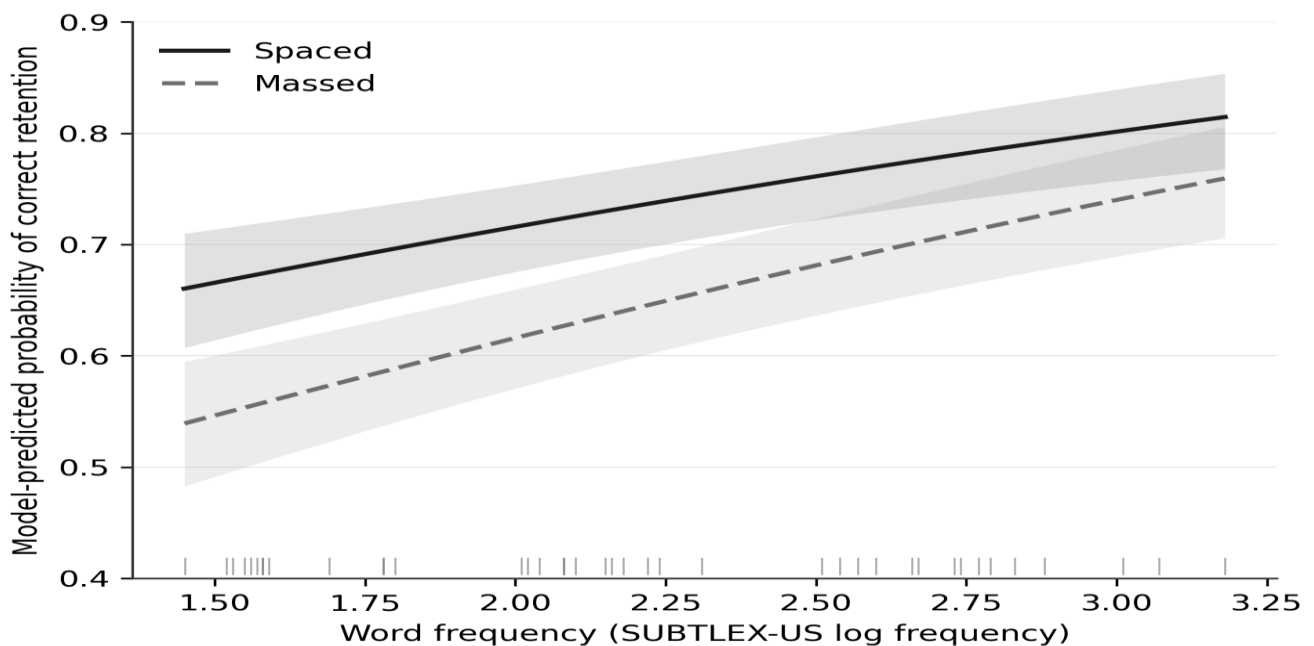
3.4. Word-Property Moderation (RQ3)

None of the three word properties moderated the effect of practice schedule. The interaction between schedule and frequency was -0.10 ($SE = 0.13$, $z = -0.79$, $p = .431$), $OR = 0.90$, 95% CI $[0.70, 1.16]$; the interaction between schedule and concreteness was -0.01 ($SE = 0.09$, $z = -0.15$, $p = .878$), $OR = 0.99$, 95% CI $[0.83, 1.18]$; and the interaction between schedule and cognate status was -0.15 ($SE = 0.18$, $z = -0.81$, $p = .418$), $OR = 0.86$, 95% CI $[0.60, 1.24]$. In each case the confidence interval for the odds ratio comfortably included 1.

By contrast, each word property exerted a clear main effect on retention: higher-frequency words

($OR = 1.69$, 95% CI $[1.46, 1.95]$), more concrete words ($OR = 1.21$, 95% CI $[1.10, 1.34]$), and cognates ($OR = 1.29$, 95% CI $[1.06, 1.58]$) were all more likely to be retained. What the data do not show is any dependence of the spacing advantage on these properties. Figure 3 illustrates this for frequency: the predicted-probability curves for the spaced and massed conditions rise in parallel across the frequency range, separated by an approximately constant vertical gap, with overlapping confidence bands. The pattern for concreteness and cognate status was equivalent. Because RQ3 was stated non-directionally, interpretation of this uniformity is deferred to Section 4.

Figure 3: Model-predicted probability of correct retention as a function of word frequency, plotted separately for the spaced and massed conditions, with 95% confidence bands. The near-parallel curves indicate no moderation of the schedule effect by frequency; tick marks along the horizontal axis show the frequency values of the 40 target items. Predictions are at the average test time.



3.5. Sensitivity and Robustness Checks

A sensitivity analysis excluding the 9 participants who reported encountering target items outside the study produced convergent results; the schedule effect was, if anything, slightly larger ($OR = 1.73$, 95% CI [1.30, 2.30]), and the schedule-by-time interaction remained significant ($p < .001$).

Because assignment was by intact class, the model was refitted with a class-level random intercept added to the participant and item intercepts. The estimated class-level variance was negligible ($SD = 0.00$, a singular fit with only 4 classes), and the schedule ($OR = 1.54$, 95% CI [1.19, 2.00]) and schedule-by-time ($p < .001$) estimates were unchanged, indicating that clustering at the class level did not account for the effects. The substantive conclusions were thus robust to this alternative random-effects specification.

Finally, to address the baseline proficiency imbalance (Section 3.1.1), proficiency was added to the primary model as a mean-centred covariate. Higher proficiency independently predicted retention ($OR = 1.04$ per point, $p < .001$), and adjusting for it strengthened rather than weakened the schedule effect ($OR = 1.77$, 95% CI [1.38,

2.29]); the schedule-by-time interaction was likewise maintained ($OR = 1.87$, $p < .001$). This is consistent with the imbalance having suppressed the unadjusted estimate, since the group with higher baseline proficiency was the one that retained less.

Two departures from the idealized preregistered plan are noted. First, the maximal random-effects structure did not yield stable slope-variance estimates and was reduced to crossed random intercepts, as described in Section 2.5. Second, the models were fitted by Laplace-approximate maximum likelihood in Python rather than with the originally intended package. Neither departure affects the direction or significance of any reported effect.

4. Discussion

This study estimated the effect of practice scheduling on EFL vocabulary retention under conditions that held delivery medium, item set, feedback, and total practice constant across conditions. Three findings emerged. First, spaced retrieval practice produced higher odds of correct retention than massed practice overall ($OR = 1.54$). Second, and centrally, this advantage was concentrated at delay: the two schedules were

statistically indistinguishable on the immediate post-test (OR = 1.13, 95% CI [0.84, 1.51]) but diverged sharply after fourteen days (OR = 2.10, 95% CI [1.57, 2.80]), yielding a large schedule-by-test-time interaction (OR = 1.86, $p < .001$). Third, this delayed advantage was uniform across the lexical properties examined: neither frequency, concreteness, nor cognate status moderated the schedule effect.

The temporal signature of the effect is its most theoretically informative feature. That the two schedules were equivalent immediately but diverged at delay is precisely the pattern predicted by an account in which distributed practice benefits retention through consolidation-dependent processes that require time to express themselves (Cepeda et al., 2008). Because the two conditions were delivered through the same platform, with the same items, retrieval format, and feedback, the divergence cannot be attributed to medium, interface, or feedback—the confounds that limit inference from much application-based research (Mahdi, 2018). The estimate therefore provides a medium-controlled quantification of the scheduling benefit for foreign-language vocabulary: under equated delivery, spaced practice roughly doubled the odds of retaining a word two weeks after study.

The absence of a schedule difference on the immediate test is consistent with, rather than contrary to, this account. Immediate-only designs are structurally unable to detect an effect that emerges at delay C. D. Smith & Scarf (2017), and the present data illustrate the point directly: had testing stopped at the immediate post-test, the study would have concluded, wrongly, that scheduling did not matter. Two features temper the interpretation of the immediate null specifically. The immediate post-test showed only modest internal consistency (KR-20 = .57), which would attenuate any true immediate difference, and the confidence interval around the immediate simple effect is wide; the immediate equivalence is therefore best read as the absence of a detectable advantage rather than as positive evidence of no difference (Olsen et al., 2025).

The uniformity of the advantage across word properties was not a foregone conclusion. We had reasoned that items already well supported at encoding high-frequency, concrete, or cognate

items might have less to gain from optimized scheduling, in which case undifferentiated spacing would waste practice on words that could not benefit from it (Sense et al., 2025). The data do not bear this out. Although each property independently predicted retention (higher frequency, greater concreteness, and cognate status all raised the odds of a correct response), none interacted with schedule, and the predicted-probability curves for the two conditions ran in parallel across the sampled range (Du, 2025). Within the range of materials examined, the spacing benefit behaved as a main effect that applied across the lexicon. The practical implication reverses the efficiency concern raised at the outset: distributing practice uniformly does not appear to squander effort on words that cannot benefit, because the words examined benefited comparably. It is worth stating the limits of this claim precisely—it holds across the frequency, concreteness, and cognate ranges sampled here, and does not exclude moderation by properties not manipulated, or at the extremes of properties that were (Ameen, 2026).

Several limitations qualify these conclusions. First, the conditions differed at baseline on general proficiency, with the massed group scoring higher; because that group nonetheless retained less, the imbalance renders the observed advantage conservative, and the effect in fact strengthened when proficiency was entered as a covariate. Even so, assignment was by intact class rather than by individual randomisation, so unmeasured class-level differences cannot be excluded—although a model incorporating a class-level random intercept returned negligible class variance and left the estimates unchanged. Second, the immediate post-test reliability was modest. Third, retention was assessed at a single fourteen-day interval, so the durability of the advantage beyond two weeks, and its form over multiple intervals, remain open. Fourth, participants shared a single first language and were drawn from a single setting, limiting generalisation. Finally, logged on-task time and per-item exposure counts were not available in the analysis dataset, so the equating of practice quantity, although built into the design, could not be verified empirically here and should be confirmed from the platform records.

For instructional practice, the results support a

simple recommendation. When digital vocabulary practice is delivered under a fixed time budget, distributing retrieval across sessions yields substantially more durable knowledge than massing it, and, within the materials examined, this holds regardless of the frequency, concreteness, or cognate status of the words being learned. The memory-systems rationale commonly invoked for such tools is thus supported specifically for the scheduling principle under equated delivery; the further assumption that scheduling gains should be concentrated on particular kinds of words receives no support here. Future work should extend the retention interval, incorporate productive-recall as well as recognition measures, sample additional first languages and instructional settings, and widen the range of word properties before the uniformity observed here is treated as general.

5. Conclusions

Under conditions that equated delivery medium, item set, feedback, and total practice, spaced digital retrieval practice produced more durable EFL vocabulary retention than massed practice. The advantage was negligible immediately after study but approximately doubled the odds of retention after two weeks, matching the temporal pattern predicted by a consolidation account and confirming that an immediate-only assessment

would have missed it entirely. The benefit did not depend on the lexical properties of the target words: frequency, concreteness, and cognate status each predicted retention, but none altered the size of the scheduling effect. For the design of digital vocabulary tools, the practical message is therefore that distributing practice over time, rather than selecting which words to distribute, is where the measurable and durable return lies.

Declarations

Ethical approval

This study was approved by the Institutional Review Board (IRB) and the Ethics Review Board of the University of Education, Lahore in accordance with the ethical standards set forth in the Declaration of Helsinki and relevant national guidelines. Ethical approval was granted under letter no. 47-18/CMS, dated 2 August 2024. This study adhered to ethical standards for research involving human participants. Written Informed consent was obtained from all participants prior to data collection. Participation was voluntary, and confidentiality and anonymity were strictly maintained. As per institutional guidelines, formal IRB approval was not required for this type of educational research; however, all ethical principles outlined in national and international research guidelines were followed.

Conflict of Interest

The authors showed no conflict of interest.

Funding

The authors did not mention any funding for this research.

References

- Ameen, F. (2026). Challenges in vocabulary learning among Kuwaiti EFL learners: a narrative review. *Discover Education*, 5(1), 233. <https://doi.org/10.1007/s44217-026-01267-6>
- Anesa, P. (2025). *Technology-Enhanced Language Learning in the Area of Legal English BT - International Handbook of Legal Language and Communication: From Text to Semiotics* (A. Wagner (ed.); pp. 1–19). Springer Nature Switzerland. https://doi.org/10.1007/978-3-031-87993-7_66-1
- Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting Linear Mixed-Effects Models Using lme4. *Journal of Statistical Software*, 67(1 SE-Articles), 1–48. <https://doi.org/10.18637/jss.v067.i01>
- Bolker, B. M., Brooks, M. E., Clark, C. J., Geange, S. W., Poulsen, J. R., Stevens, M. H. H., & White, J.-S. S. (2009). Generalized linear mixed models: a practical guide for ecology and evolution. *Trends in Ecology & Evolution*, 24(3), 127–135. <https://doi.org/10.1016/j.tree.2008.10.008>
- Cadavid, S., Cortés-Albornoz, M. C., Gómez-Carvajal, A.-M., Mendoza-Ayús, S. D., Luna, K., Olaya Galindo, M. D., Vélez-Van-Meerbeke, A., & Talero-Gutiérrez, C. (2025). Learning techniques to improve memory in children: a systematic review. *European Journal of Psychology of Education*, 40(1), 38. <https://doi.org/10.1007/s10212-024-00936-0>
- Cepeda, N. J., Vul, E., Rohrer, D., Wixted, J. T., & Pashler, H. (2008). Spacing effects in learning: a temporal ridge line of optimal retention. *Psychological Science*, 19(11), 1095–1102. <https://doi.org/10.1111/j.1467-9280.2008.02209.x>
- Doshi, M. H., Kumar, R., Upadhyay, Y., Sharma, R. K., & Rao, A. (2026). *Multilingual RAG-Based Educational System with Active Learning: A Real-Time Approach to Personalized Education BT - Innovative Computing and Communications* (A. E. Hassanien, S. Anand, A. Jaiswal, & P. Kumar (eds.); pp. 286–297). Springer Nature Switzerland.
- Du, Q. (2025). How artificially intelligent conversational agents influence EFL learners' self-regulated learning and retention. *Education and Information Technologies*, 30(15), 21635–21701. <https://doi.org/10.1007/s10639-025-13602-9>
- Fokin, D., Flużyczka, M., & Golovin, G. (2025). The Polish vocabulary size test: A novel adaptive test for receptive vocabulary assessment. *Behavior Research Methods*, 57(9), 254. <https://doi.org/10.3758/s13428-025-02775-3>
- Hirsch, M. L., Kliegl, O., & Bäuml, K.-H. T. (2026). Boosting hypermnnesia with a cued-recall test format. *Memory & Cognition*. <https://doi.org/10.3758/s13421-026-01910-5>
- Honda, Y., Tanizawa, K., Masai, K., Nakamura, Y., Choi, H., & Fukushima, S. (2025). *Generating Sensory Vocabulary Cues for Describing Scents in Olfactory-Enhanced Second Language Vocabulary Learning BT - Learning and Collaboration Technologies* (B. K. Smith & M. Borge (eds.); pp. 167–184). Springer Nature Switzerland.
- Jarvis, M., Vasarhelyi, A., Anderson, J., Mulley, C., Lipp, O. V., & Ney, L. J. (2025). js-mEye: An extension and plugin for the measurement of pupil size in the online platform jsPsych. *Behavior Research Methods*, 58(1), 8. <https://doi.org/10.3758/s13428-025-02901-1>
- Kohnke, L. (2023). *Designing Microlearning Activities with Podcasts, Videos, Infographics, and Flashcards, and Microlearning Activities BT - Using Technology to Design ESL/EFL Microlearning Activities* (L. Kohnke (ed.); pp. 43–60). Springer Nature Singapore. https://doi.org/10.1007/978-981-99-2774-6_5
- Lee, I.-H., Maechtle, C., & Hu, C.-F. (2021). Enhancing Vocabulary Retention in Low-Achieving EFL Students: Massed or Spaced? *English Teaching & Learning*, 45(4), 415–430.

<https://doi.org/10.1007/s42321-020-00074-y>

- Mahdi, H. S. (2018). Effectiveness of mobile devices on vocabulary learning: A meta-analysis. *Journal of Educational Computing Research*, 56(1), 134–154.
- Notaro, G. M., Mustari, R., Haghighat, A. K., Gao, D., Molazadeh, V., & Saez, M. (2025). *Assessing Comfort During Human-Robot Collaboration Using Virtual Reality Scenarios BT - Computer-Human Interaction Research and Applications* (H. Plácido da Silva & P. Ciproso (eds.); pp. 168–185). Springer Nature Switzerland.
- Olsen, W. L., Collins, M., Jastrzemski, T. S., & Myers, C. W. (2025). Optimizing Cognitive Performance Using Memory Enhancing Acute Intermittent Hypoxia (meAIH) in Young Healthy Adults: A Preliminary Study. *MedRxiv*, 2011–2025.
- Ostrovsky, T., Ungermann, P., & Donkin, C. (2025). Using machine learning to automate the collection, transcription, and analysis of verbal-report data. *Behavior Research Methods*, 57(10), 285. <https://doi.org/10.3758/s13428-025-02800-5>
- Sabih, I. S., & Kuyini, A. (2025). *Memory Enhancing Strategies to Support Learning and Cognitive Functioning in Students with Neurodevelopmental Disorders BT - BUiD Doctoral Research Conference 2024* (K. Al Marri, F. A. Mir, A. Awad, & A. Abubakar (eds.); pp. 309–319). Springer Nature Switzerland.
- Sense, F., Dye, I., Collins, M. G., Krusmark, M., Starkey, J., Asadpour, K., Graham, L., & Myers, T. (2025). Optimizing language proficiency: a competency-driven approach to adaptive instruction in defense language training. *International Conference on Human-Computer Interaction*, 279–298.
- Smith, C. D., & Scarf, D. (2017). Spacing repetitions over long timescales: a review and a reconsolidation explanation. *Frontiers in Psychology*, 8, 241716.
- van der Velde, M., Sense, F., Borst, J. P., & Rijn, H. van. (2024). Large-scale evaluation of cold-start mitigation in adaptive fact learning: Knowing “what” matters more than knowing “who.” *User Modeling and User-Adapted Interaction*, 34(4), 1467–1491. <https://doi.org/10.1007/s11257-024-09401-5>
- Xodabande, I., Atai, M. R., & Hashemi, M. R. (2025). Exploring the effectiveness of mobile assisted learning with digital flashcards in enhancing long-term retention of technical vocabulary among university students. *Journal of Computers in Education*, 12(4), 1307–1332. <https://doi.org/10.1007/s40692-024-00347-6>.