

Does spaced repetition increase retention of phrasal verbs in adult L2 learners?

Candice Helen Glenday¹,

Claudia Marcela Suescun Giraldo²,

Juan Camilo Suarez Franco³,

Leidys Helena Rivera-Quiroz⁴

ABSTRACT:

This quasi-experimental study investigates whether spaced repetition and chunk-based instruction enhance adult EFL learners' sensitivity to the syntactic constraints governing inseparable phrasal verbs. A total of 83 English undergraduate students at B1 proficiency level were recruited from universities in Brazil and Colombia for the study. The experimental group received a six-week intervention with flashcard activities and Kahoot-based spaced repetition of separable and inseparable phrasal verbs. The control group did not participate in the intervention and followed the standard English curriculum. Participants' performance was assessed using an offline acceptability judgment task (AJT) before the intervention and online reading-time measures administered via PsychoPy after the intervention. However, the two groups were not equivalent at baseline, with the control group demonstrating moderate effect size for pre-test AJT accuracy (Cohen's $d = 0.73$). A mixed-effects logistic regression incorporating pre-test accuracy as a covariate found no significant Group effect at post-test ($p = .216$), indicating that the apparent post-test advantage of the control group was attributable to pre-existing ability differences rather than the instructional intervention. The only robust fixed effect was Condition: the ungrammatical word order (INSEPvop) was significantly harder to evaluate correctly than the grammatical order (INSEPvpo) for all participants regardless of group. These findings highlight the importance of baseline equivalence and covariate-adjusted analyses in quasi-experimental SLA research and identify methodological factors, including task-format differences between assessments and participant attrition—that should be addressed in future investigations of chunk-based instruction for phrasal verb acquisition in adult EFL learners.

KEYWORDS: Phrasal verbs. Chunk-based instruction. Spaced repetition. Acceptability judgment task. Adult EFL learners.

1 Professora da Universidade Estadual Vale do Acaraú. Doutorado pelo Programa de Pós-graduação em Linguística da Universidade Federal da Paraíba (UFPB). E-mail: candice_glenday@uvanet.br. ORCID: <https://orcid.org/0000-0002-7644-5415>

2 Doutoranda pelo Programa Estudios Interdisciplinarios en Psicología da Fundación Universitaria Luis Amigó. Email: claudia.suescungi@amigo.edu.co. ORCID: <http://orcid.org/0000-0002-3254-2416>.

3 Mestrado pelo Programa Intercultural da Universidad de la Rioja. Email: juan.suarezfr@amigo.edu.co. ORCID: <https://orcid.org/0009-0001-0848-1078>.

4 Doutoranda pelo Programa Modelación y Computación Científica da Universidad de Medellín. E-mail: leidys.riveraqu@amigo.edu.co. ORCID: <https://orcid.org/0000-0003-1926-2521>

1. INTRODUCTION

Adult learners constitute a fundamental population in second language learning, as they present several distinctive traits that must be taken into account, such as prior learning experiences (Kolb, 1984), cognitive maturity, and motivation to learn new languages — particularly given their awareness of the increasing importance of multilingual competence. Unlike children and, in some cases, adolescents, adults tend to be more conscious of how they learn and may rely more heavily on metalinguistic awareness. As a result, learning a second language in adulthood is not an easy task, especially with regard to functional elements.

Functional elements—such as determiners, conjunctions, and prepositions—play a crucial role in shaping sentence structure and meaning. While nouns, verbs, and adjectives carry more concrete lexical meaning and are therefore easier to grasp, functional elements are more abstract, which makes them particularly challenging for learners. Among these elements, prepositions are especially difficult for adult second language learners (Abdunazarovna, 2025), especially when they occur in phrasal verbs (Al-Shunnag et al., 2025). To understand phrasal verbs, learners must recognize that meaning emerges from the combination of a verb and a prepositional or adverbial particle, which may or may not be separable from the verb.

Traditional English language instruction often presents learners with lists of phrasal verbs to be memorized, but it rarely emphasizes that some of these constructions function as prepositional verbs and that their interpretation is highly context dependent. Consequently, teaching practices should focus on whole sentence structures and contextualized meaning to foster deeper learning. One instructional approach that has been shown to facilitate the learning of phrasal verbs in adult populations is chunk-based instruction (Lewis, 1993), which is grounded in the view that language does not consist solely of grammatical rules, but rather of multi-word prefabricated units, or chunks (Lewis et al., 1997).

Studies have demonstrated that formulaic sequences, including chunks, contribute significantly to language fluency, particularly in speaking and writing. For example, Wood (2010) and Nergis (2021) found that adult learners exposed to targeted instruction in formulaic sequences showed improvements in fluency and naturalness in speech production. Similarly, Boers and Lindstromberg (2012) reported that explicit chunk-based instruction resulted in

better performance on both immediate and delayed post-tests, suggesting durable learning effects. A growing body of empirical research (Feng and Yang, 2025; Lu and Wang, 2024; Ton et al., 2023; Nergis, 2021) has been investigating the effectiveness of chunk-based instruction for teaching such linguistic features in adult EFL classrooms.

With respect to syntactic complexity and accuracy, chunk-based instruction appears to promote more native-like language production. In a study involving German EFL learners, Peters (2012) demonstrated that familiarity with formulaic sequences led to improved syntactic cohesion and reduced hesitation phenomena in speaking tasks. Moreover, despite the tendency for adult learners to develop more fossilized linguistic systems, they often respond positively to chunk-based instruction. Many perceive it as a pragmatic means for enhancing communicative competence without the need to master every grammatical rule explicitly (Boers and Lindstromberg, 2012). This finding highlights the alignment between adult learning styles—typically strategic and goal-oriented—and the benefits of chunk-based instruction.

A considerable number of Brazilian and Colombian undergraduate students begin formal English instruction during high school, that is, in late adolescence. Therefore, examining the challenges these learners face with phrasal verbs — a particularly complex linguistic feature as there are no equivalents in Portuguese or Spanish — can yield valuable insights into their learning processes. This focus allows for a better understanding of how age, language exposure, and educational experience influence language acquisition, thereby contributing to the development of teaching strategies tailored to adult learners' needs. Accordingly, this quasi-experimental study investigates whether spaced repetition of separable and inseparable phrasal verbs enhances adult EFL learners' sensitivity to the syntactic constraints governing these constructions, as reflected in acceptability judgments and reading-time measures. By combining offline accuracy and online processing data, the study aims to clarify to what extent the sensitivity of adult learners to syntactic violations in phrasal verbs would be affected by chunk-based instruction.

Spaced repetition has emerged as a significant methodology in ESL teaching, broadly defined as a technique that leverages the principles of cognitive psychology to optimize the retention of information over time, aligning seamlessly with the dynamic and non-linear nature of language acquisition (Arvanitis, 2019). A prevalent trend in the literature is the unanimous acknowledgment of its role as a potent facilitator of vocabulary retention, with studies

demonstrating that systematic spacing of learning intervals not only enhances learners' ability to recall a broader range of vocabulary but also fosters a deeper understanding and integration of words into their linguistic repertoire (Mortazavi et al., 2021). Beyond vocabulary, the spaced delivery of grammatical concepts has been shown to optimize cognitive processing, contributing to a more robust and sustained grasp of grammatical rules and overall language competence (Choffin et al., 2019). Researchers consistently assert that these benefits extend holistically to overall language proficiency, systematically reinforcing various linguistic elements among ESL learners (Kondratjew and Kahrens, 2019). Furthermore, the personalized nature of spaced repetition accommodates diverse learner preferences — whether visual, auditory, or kinesthetic — making it a universal tool for heterogeneous learner populations (Belyaeva et al., 2019), while its effectiveness has been demonstrated across varied age groups, cultural backgrounds, and educational settings (Suzuki, 2023). The growing recognition of its efficacy has also prompted a discernible trend toward technological integration, with mobile applications, online platforms, and specialized classroom software increasingly incorporating spaced repetition algorithms to provide flexible, personalized, and continuous learning experiences (Lei et al., 2022; Su e Zou et al., 2022; Tabari, 2026). Nevertheless, conflicts remain in the literature regarding optimal spacing intervals, applicability across proficiency levels, and methodological variations between laboratory and real-world settings (Altiner, 2019; Hautasaari et al., 2019), underscoring the compelling need for further research to uncover the full potential of spaced repetition as a foundational instructional approach in ESL education (Godwin-Jones, 2018).

1. METHOD

This research was conducted within an empirical-analytical framework and used a quasi-experimental design. The population consisted of students enrolled at Universidad Católica Luis Amigó (Medellín) and Universidade Estadual Vale do Acaraú (Sobral). A total of 83 adult learners enrolled in B1-level English courses were selected as the final study sample. Participants were divided into two groups: an experimental group that received explicit

instruction with spaced repetition of phrasal verbs, and a control group that did not receive any additional instruction and followed the standard English curriculum.

B1-level students were intentionally selected based on both pedagogical and research-related considerations. At this proficiency level, learners are able to communicate using more complex syntactic structures, making them suitable for examining the impact of chunk-based instruction on syntactic development. Additionally, B1 learners are in a transitional stage between basic and more independent language use, during which the incorporation of chunks may enhance both fluency and grammatical accuracy.

Participant distribution was designed to ensure balance in terms of both group size and linguistic proficiency. This arrangement helped ensure that any observed differences in syntactic development could be attributed to the instructional treatment rather than to extraneous variables.

1.1. PARTICIPANTS

The research was carried out from January to May 2025. A total of 83 undergraduate students were recruited to participate in the first stage of the study. After analysis of the preliminary data, 25 students were excluded for not meeting the inclusion criteria. During the intervention stage, 10 students assigned to the experimental group failed to participate in the minimum of 75% of the sessions and were not included in the posttest stage. The final number of participants in both experimental and control groups consisted of 37 students (dropouts in posttest stage $n = 11$), that is, 20 students in the experimental group and 17 students in the control group.

The experimental group consisted of 7 male (35%) and 13 female (65%) students. The control group consisted of 17 students, of which 9 were male (53%) and 8 were female (47%) students. The final sample consisted of 20 Colombian students and 17 Brazilian students. No significant difference was found between the number of male and female participants in each group, indicating gender matching after the intervention (Chi-square p -value = 0.27). The mean age of the participants was 23.2 years. All participants gave their consent prior to their participation. The research project was submitted and approved by the Research Ethics Committee of the Universidad Católica Luis Amigó, Medellín, Colombia, under protocol

number 50518 and the Research Ethics Committee of the Universidade Estadual Vale do Acaraú under protocol number 6,812,860.

2. INSTRUMENTS AND PROCEDURE

The present study was conducted in three stages. In the first stage, the screening stage, four tasks were administered to assess participants' prior knowledge of vocabulary and syntax. The second stage consisted of the instructional intervention, and the third stage involved the administration of posttests. This quasi-experimental study was designed and implemented using Google Forms and PsychoPy (Peirce et al., 2019) for data collection.

During the first screening stage, baseline data were collected using four instruments: the Vocabulary Size Test (VST), the Word Associates Test (WAT), a syntactic knowledge test, and a pre-intervention acceptability judgment task. The VST (Laufer and Nation, 1999) measures breadth of vocabulary knowledge and consists of 100 items. To estimate the number of word families known by each participant, the total number of correct responses was multiplied by 200, yielding a maximum score of 20,000-word families. The test includes items ranging from high-frequency to low-frequency vocabulary.

The WAT (Read, 1993, 2000) assesses depth of lexical knowledge and consists of 40 items. Each item presents an adjective or a noun followed by eight words, from which participants are required to select four semantically related options. The maximum possible score on the WAT is 160 points.

The syntactic knowledge test consisted of 45 multiple-choice items adapted from TOEFL materials, targeting basic grammatical knowledge. Of these items, 20 items required participants to identify the correct syntactic word order. The maximum attainable score on this test was 45 points.

The acceptability judgment task (AJT) included 40 sentences, yielding a maximum score of 40 points. This task was designed to assess participants' prior knowledge of phrasal verbs and served as baseline data for comparison with post-intervention performance. Examples of the experimental items are shown below. All four instruments in the first stage were administered in person in a laboratory setting using Google Forms.

To be eligible to participate in the second stage, the first criterion was the VST score. Participants were excluded from the study if their scores were below 20 points or above 90 points in the VST. The second criterion was the score in the syntactic knowledge test. Participants were excluded from the study if their scores were below 12 points or above 42 points in the syntactic knowledge test. A total of 25 participants were excluded in the first stage. To be placed in the experimental group in the second stage, participants had to score between 15 and 27 points in the pre-intervention acceptability judgment task.

In the second stage, 20 participants in the experimental group received a six-week instructional intervention consisting of two complementary instructional strategies. During the first three weeks, students interacted with a set of flashcards featuring 20 separable (transitive) and 20 inseparable (intransitive) phrasal verbs. The phrasal verbs were selected in accordance with the guidelines of B1 CEFR-level. Each flashcard included a contextualized sentence and a corresponding image related to the phrasal verb. During the following three weeks, the groups of students participated in Kahoot-based activities in which they practiced the same phrasal verbs in novel sentence contexts different from those used in the flashcards. Throughout the instructional sessions, all phrasal verbs were presented randomly to prevent the likelihood of order effects. Both stages of the intervention were delivered in biweekly 30-minute sessions at the university.

In the third stage, following completion of the intervention, participants in the experimental and control groups completed a 1-week delayed post-intervention acceptability judgment task administered in person via PsychoPy software (version 2024.2.4). The purpose of the acceptability judgment task was to compare post-intervention acceptability ratings with pretest performance. The task assessed learners' sensitivity to syntactic violations involving inseparable phrasal verbs. The task consisted of 50 sentences, including 20 experimental items and 30 distractors. The experimental items were divided into three conditions: grammatical inseparable verb–particle–object sentences (INSEPvpo), grammatical separable verb–particle–object sentences (Controlsep) and ungrammatical inseparable verb–object–particle sentences (INSEPvop). Each full sentence was displayed in the center of the screen. At the end of each sentence, participants were asked to judge if the sentence was grammatical or ungrammatical by pressing keys Y (yes) or N (no). Although there was no time limit on this task, the participants were encouraged to respond to each item as quickly and accurately as possible,

since both reaction time and accuracy would be recorded. The experimental session was preceded by detailed instructions about the task that were presented in English, and a few practice questions were provided. The entire session took approximately 20 min for each participant. All sentences in the INSEP condition were controlled for length and contained 10 words (47–48 characters). The sentences in the Control condition were separable phrasal verbs, and the trials were presented with the verb followed by the preposition or the verb followed by the object. Examples of experimental stimuli are shown below:

(Controlsep) Kate put the shelf up in her bedroom. // Kate put up the self in her bedroom.

(INSEPVpo) Mac suddenly broke up with Patrica before they left home.

(INSEPVop) Mac suddenly broke with Patrica up before they left home (violation).

The stimuli were randomized so that each participant was exposed to only one condition per phrasal verb. It was expected that the reading times would be faster for inseparable verb–particle–object constructions and slower for the verb–object–particle violation sentences. It was also expected that the acceptability ratings would be higher for grammatical inseparable verb–particle–object constructions and lower for verb–object–particle violation sentences when compared with the pretest results. The reading times for the conditions were recorded and analyzed in relation to accuracy rates.

3. DATA ANALYSIS AND RESULTS

Prior to analysis, normality of all continuous variables was assessed using the Shapiro-Wilk test. Pre-test AJT scores did not significantly deviate from normality ($W = 0.94$, $p = .054$), supporting the use of Student's t for the pre-test group comparison (Table 2). Post-test AJT scores for the INSEPVpo condition showed a significant departure from normality ($W = 0.94$, $p = .047$), and non-parametric tests (Mann-Whitney U) were used for that posttest comparison. Gain score difference variables were normally distributed (all $W_s \geq 0.96$, $p_s \geq .166$); however, given the small sample size ($n = 20$ and $n = 17$), Wilcoxon signed-rank and Mann-Whitney U tests were used as conservative non-parametric alternatives and are reported alongside parametric results. Data were analyzed using R statistical software (4.5.2 version).

Mixed-effects logistic regression models were fitted using the lme4 package (Bates et al., 2015), with p-values obtained via lmerTest (Kuznetsova et al., 2017). Estimated marginal means and pairwise contrasts were computed with emmeans (Lenth, 2024). Marginal and conditional R² values were calculated following Nakagawa and Schielzeth (2013). Non-parametric tests and effect sizes were computed using base R functions (R Core Team, 2024).

No statistically significant differences were found between the experimental and control groups on the control variables, including WAT scores (4-word and partial-word measures), VST scores, and syntactic knowledge scores (Table 1), indicating baseline comparability across general language proficiency measures.

Table 1: Baseline values between groups.

Variable	Group	Mean	SD	Median	IQR	p-value
WAT score 4 words/40	Experimental	3.4	4.5	1.5	3.7	0.85
	Control	3.1	3.0	2.0	3.0	
WAT score partial words/160	Experimental	49.3	19.4	42.5	25.2	0.45
	Control	46.2	18.6	40.0	23.0	
VST score/100	Experimental	43.7	18.4	40.0	30.2	102
	Control	54.1	19.5	52.0	36.0	
Syntactic knowledge score	Experimental	18.3	7.2	15.0	2.5	0.48
	Control	19.5	7.2	18.0	5.0	

P-value for the Mann-Whitney U test.

However, baseline performance on the acceptability judgment task (AJT) was not equivalent between groups. Prior to the intervention, the control group demonstrated significantly higher acceptability rates than the experimental group (Table 2), $t = 2.239$, $p = .032$, with a moderate effect size (Cohen's $d = 0.73$, 95% CI [0.06, 1.40]). This baseline non-equivalence is an important methodological consideration and it is addressed directly in the primary analysis through the inclusion of pre-test score as a covariate.

Table 2: Difference in acceptability rates in the acceptability judgment task between the experimental and control groups before the intervention.

Group	Mean	SD	Mean diff.	p-value	Effect size	CI (95%)
Experimental	64.8	16.7	-11.4	.032*	0.73	0.06; 1.40
Control	76.2	14.3				

*Student's t-test. * Statistical difference.*

3.1. PRIMARY ANALYSIS: MIXED-EFFECTS LOGISTIC REGRESSION WITH PRE-TEST COVARIATE

Given the significant baseline difference between groups in AJT performance, the primary inferential analysis used a mixed-effects logistic regression model that included each participant's mean pre-test accuracy as a centered continuous covariate. This approach controls for pre-existing differences in phrasal verb knowledge and allows the Group effect to be interpreted after removing initial ability differences. The model included Group (Control, Experimental), Condition (INSEPvpo, INSEPvop), their interaction, and centered pre-test score as fixed effects, with random intercepts for Participant and Item.

The model fit the data significantly better than a null model. The fixed effects accounted for 7.3% of the variance in accuracy (marginal $R^2 = .073$; conditional $R^2 = .059$). Key results are shown in Table 3.

Table 3: Mixed-effects logistic regression — posttest accuracy (pre-test score as covariate).

Predictor	β	SE	z	p	OR	95% CI (OR)	Sig.
Intercept	1053	224	4.69	< .001	2.87	[1.85, 4.45]	***
Group (Experimental)	-.375	303	-1.24	.216	0.69	[0.380, 1.245]	ns
Condition (INSEPvop)	-.530	242	-2.19	.029	0.59	[0.366, 0.946]	*
Pre-Test Score (covariate)	878	766	1.15	.252	2.405	[0.536, 10.79]	ns
Group $\times$ Condition	-.397	319	-1.24	.214	0.672	[0.359, 1.257]	ns

Note. β = log-odds coefficient. OR = odds ratio. 95% CI for OR. Pre-Test Score centered at grand mean. Random intercept variance: Participant = 0.281; Item = 0.000 (boundary singular).

After controlling for pre-test accuracy, the Group effect was not statistically significant ($\beta = -0.375$, $z = -1.24$, $p = .216$, OR = 0.69). This indicates that once baseline differences are accounted for, the experimental and control groups did not differ significantly in overall post-test accuracy on the grammatical condition. The only robust fixed effect was Condition: the ungrammatical condition INSEPvop was significantly harder to evaluate correctly than the grammatical condition INSEPvpo for all participants, regardless of group ($\beta = -0.530$, $z = -2.19$, $p = .029$, OR = 0.59). The Group $\times$ Condition interaction was not significant ($p = .214$),

indicating that the intervention did not differentially affect performance on one condition relative to the other.

Pre-test score was in the expected direction—higher baseline accuracy predicted higher posttest accuracy (OR = 2.41)—but did not reach significance ($p = .252$), likely reflecting limited variance in pre-test scores and the relatively small sample size.

Estimated marginal means from the model provide the clearest picture of group performance at posttest after covariate adjustment (Table 4). The Control group showed higher estimated accuracy in both conditions, and the group difference on the ungrammatical condition (INSEPvop) was significant even after Bonferroni correction (OR = 2.16, $p = .008$), while the difference on the grammatical condition (INSEPvpo) was not ($p = .216$).

Table 4: Estimated marginal means by Group $\times$ Condition (posttest).

Group	Condition	Prob.	SE	95% CI (lower)	95% CI (upper)
Control	INSEPvpo (grammatical)	741	43	649	816
Experimental	INSEPvpo (grammatical)	663	44	573	743
Control	INSEPvop (ungrammatical)	628	50	527	719
Experimental	INSEPvop (ungrammatical)	438	47	349	532

Note. Probabilities estimated from the mixed-effects model at the grand mean of pre-test score. Confidence intervals back-transformed from the logit scale.

3.2. GAIN SCORE ANALYSIS

To provide an additional perspective on pre-to-post change, gain scores were computed by subtracting each participant's mean pre-test accuracy from their mean posttest accuracy at the Participant $\times$ Condition level. Descriptives are reported in Table 5.

Table 5: Pre-test, post-test, and gain score accuracy by group and condition.

Group	Condition	Pre M	Post M	Gain M	Gain SD
Control	INSEPvop	700	629	-71	346
Control	INSEPvpo	805	735	-71	205
Experimental	INSEPvop	590	430	-160	359
Experimental	INSEPvpo	770	645	-125	197

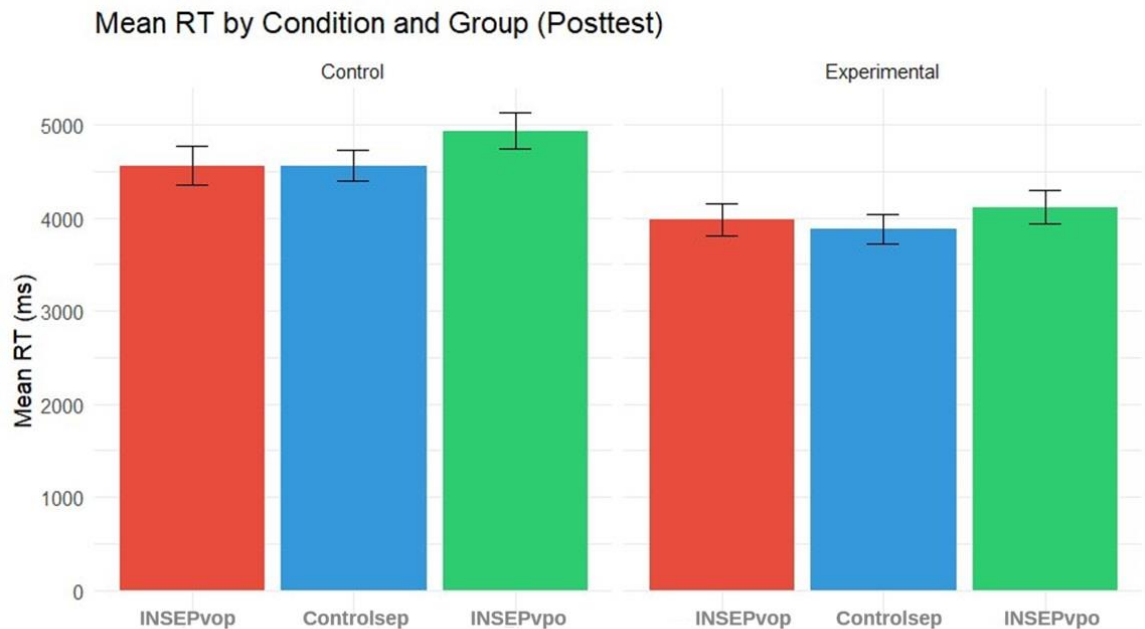
Note. M = mean proportion correct. Negative gain scores indicate a decline in accuracy from pre- to posttest.

Gain scores were negative in all four group–condition combinations, indicating that both groups declined in accuracy from pre- to post-test. Mann-Whitney U tests comparing gain scores between groups revealed no significant differences for INSEPvop ($U = 148.5$, $p = .521$, $r = .126$) or INSEPvpo ($U = 141.5$, $p = .390$, $r = .168$), indicating that the groups did not differ significantly in the magnitude of pre-to-post change. Wilcoxon signed-rank tests against zero showed that the only statistically significant change was a decline in the Experimental group on the grammatical condition INSEPvpo ($W = 24.0$, $p = .007$, $M = -0.125$). No group showed any significant improvement from pre- to posttest.

3.3. READING TIMES DURING THE ACCEPTABILITY TASK

Sentence reading times were analyzed using a linear mixed-effects model with Condition (INSEPvop, INSEPvpo, ControlVnp), Group (Control, Experimental), their interaction, and centered pre-test accuracy as fixed effects, and random intercepts for Participant. Reading times were log-transformed prior to analysis and outliers were removed (RTs < 100 ms and RTs exceeding 2.5 SD above each participant's mean; 1.4% of trials removed). A model comparison indicated that the simplified model without interaction terms fit the data better than the full model ($\chi^2(2) = 2.11$, $p = .35$) and was therefore retained as the final model. The analysis revealed a significant main effect of Group ($\beta = -0.195$, $SE = 0.074$, $t = -2.62$, $p = .013$), with the Experimental group reading sentences approximately 180 ms faster overall than the Control group across all conditions. Neither the main effect of Condition nor the Group $\times$ Condition interaction reached significance (all $ps > .08$), indicating that the intervention did not selectively facilitate processing of grammatical over ungrammatical word orders. Pre-test accuracy was not a significant predictor of reading time ($p = .31$). Figure 1 shows the mean reading times by condition and group.

Figure 1: Mean reading times (ms) by condition and group at posttest. Error bars represent standard errors. Red = INSEPvop (ungrammatical); Blue = Controlsep (grammatical control); Green = INSEPvpo (grammatical inseparable).



A second linear mixed-effects model was fitted to log-transformed answer times. The model revealed no significant main effects of Group or Condition and no Group $\times$ Condition interaction (all $ps \geq .13$), indicating that post-sentence response latencies did not differ across experimental conditions or groups.

For reference, unadjusted accuracy data (not controlling for pre-test baseline) showed a significant main effect of Group ($z = -6.75$, $p < .001$), Condition ($z = -3.95$, $p < .001$), and a non-significant interaction ($p = .84$), as reported in Table 6. However, given the significant baseline non-equivalence between groups, these unadjusted figures should be interpreted with caution.

Table 6: Fixed effects — unadjusted accuracy model (log-odds scale).

Effect	β	SE	z	p	Note
Intercept	1.24	0.07	18.97	< .001	
Group	-0.58	0.09	-6.75	< .001	Control > Exp.
Condition	-0.20	0.05	-3.95	< .001	INSEPvpo > INSEPvop
Group $\times$ Condition	-0.01	0.07	-0.20	.84	ns

Note. Unadjusted model does not include pre-test covariate. Reported for reference only; see Table 3 for the primary covariate-adjusted analysis.

4. DISCUSSION

The present study investigated the effects of spaced repetition and chunk-based instruction on adult EFL learners' processing and grammatical sensitivity to inseparable phrasal verbs. Building on previous research highlighting the pedagogical value of formulaic language for adult learners (Feng and Yang, 2025; Lu and Wang, 2024; Ton et al., 2023), this study sought to determine whether targeted instruction would modulate learners' sensitivity to syntactic constraints associated with phrasal verbs, as well as their online sentence processing. The findings, however, do not support the conclusion that the intervention produced significant learning gains, and a careful reanalysis incorporating baseline differences between groups substantially revises the interpretation of the posttest data.

The primary analysis—a mixed-effects logistic regression including pre-test accuracy as a centered covariate—found no significant Group effect after controlling for baseline differences ($\beta = -0.375$, $p = .216$). This indicates that the apparent posttest advantage of the Control group, observed in unadjusted analyses, is largely attributable to the groups' non-equivalent starting points rather than to the instructional intervention. The only robust effect that survived covariate adjustment was Condition: the ungrammatical word order (INSEPvop) was significantly harder to evaluate correctly than the grammatical order (INSEPvpo) for all participants regardless of group ($OR = 0.59$, $p = .029$). This is a purely linguistic finding, reflecting the intrinsic asymmetry between grammatical and ungrammatical sentence types in acceptability tasks, and is not attributable to instruction.

The Group $\times$ Condition interaction was not significant ($p = .214$), indicating that chunk-based instruction did not differentially enhance performance on either condition relative to the control group once baseline accuracy was taken into account. Critically, the gain score analysis revealed that both groups declined in accuracy from pre- to posttest across all conditions, and the groups did not differ significantly in the magnitude of this decline. The only statistically significant pre-to-post change in the entire dataset was a decline in the Experimental group on the grammatical condition INSEPvpo ($p = .007$), which is in the opposite direction to the expected intervention effect.

These results stand in contrast to previous findings suggesting that chunk-based instruction facilitates learners' sensitivity to construction-level constraints by promoting holistic representations of verb-particle combinations (Lewis, 1993; Peters, 2012). Several methodological factors may account for this discrepancy. First, the significant baseline non-equivalence between groups (Cohen's $d = 0.73$ in favor of Controls on pre-test AJT performance) substantially complicates the attribution of posttest differences to the intervention. Although random intercepts for participants were included in the mixed-effects model, the quasi-experimental design with non-random group assignment limits causal inference. Second, the six-week instructional period, though ecologically plausible, may have been insufficient to produce durable representational changes of the kind detectable in an acceptability judgment task administered via a computer-based platform (PsychoPy) that differed from the Google Forms pre-test format. Task-format differences between pre- and posttest may have introduced unfamiliarity effects that obscured genuine learning. Third, participant attrition reduced the final sample to 37 participants, substantially limiting statistical power for detecting small to medium effects.

With respect to online processing, the linear mixed-effects model on sentence reading times revealed a significant main effect of Group ($\beta = -0.195$, $SE = 0.074$, $t = -2.62$, $p = .013$), with the Experimental group reading sentences approximately 180 ms faster overall than the Control group across all conditions. However, no significant main effect of Condition or Group $\times$ Condition interaction was observed (all $ps > .08$), indicating that this speed advantage was general rather than specific to grammatical or ungrammatical word orders. Whilst previous work has suggested that instructional gains may first emerge at the level of explicit judgments before affecting early-stage processing measures (Lu and Wang, 2024), the present results do not provide evidence for condition-specific gains even at the level of offline acceptability judgments after controlling for baseline performance. The absence of condition-specific reading-time effects therefore cannot be interpreted as evidence of incipient representational change for inseparable verb-particle constructions. It is notable, however, that the general processing advantage of the Experimental group suggests that the intervention might have produced broader facilitation of sentence processing fluency, even in the absence of targeted grammaticality sensitivity.

It is notable that the pre-test score covariate was directionally consistent with expectations—higher baseline accuracy predicted better post-test performance ($OR = 2.41$)—but did not reach significance ($p = .252$). This likely reflects the restricted range of pre-test scores in both groups following the eligibility screening and the limited sample size, rather than the absence of a true baseline-to-outcome relationship.

The present findings highlight the importance of establishing rigorous baseline equivalence in quasi-experimental studies of instructional effects. The use of covariate-adjusted analyses—particularly mixed-effects models that account for participant-level clustering—is essential when random assignment to conditions is not feasible. Future research should use matched-group or crossover designs, or incorporate pre-test scores as covariates from the outset, to allow more confident attribution of post-test differences to the instructional manipulation.

At the same time, the decline in accuracy observed in both groups across all conditions warrants further investigation. Possible explanations include differences in task format between the Google Forms pre-test and the PsychoPy post-test, participant fatigue or reduced engagement during the delayed posttest, the effects of a one-week gap allowing partial forgetting, or ceiling effects in the Control group that may have paradoxically reduced their posttest scores relative to a high pre-test baseline. Disentangling these factors would require additional experimental control, such as the inclusion of a pure retest group receiving neither instruction nor alternative activities.

In conclusion, the present study does not provide evidence that the chunk-based spaced repetition intervention significantly enhanced adult EFL learners' sensitivity or accuracy on an acceptability judgment task for inseparable phrasal verbs, once pre-existing group differences in baseline performance are accounted for. The results underscore the methodological challenges inherent in quasi-experimental SLA research and highlight the need for covariate-adjusted analyses, stricter group equivalence, and larger samples in future studies. Despite the null pattern of instructional effects, the study makes a substantive empirical contribution by demonstrating the importance of baseline controls in this research area and by identifying potential confounds—including task-format differences and pre-test non-equivalence—that should be addressed in subsequent investigations.

4.1. LIMITATIONS

Several limitations should be acknowledged. First, the final sample size was substantially reduced due to participant attrition, resulting in a relatively small number of observations at the post-intervention stage. Although mixed-effects modelling mitigates some concerns related to reduced power, future studies with larger and more stable samples would allow for more fine-grained analyses, including the inclusion of random slopes and more complex model structures.

Second, and most critically, baseline differences between groups in acceptability judgment performance complicate the interpretation of instructional effects. The groups were not equivalent at pre-test, which means that unadjusted comparisons of posttest scores confound the intervention effect with pre-existing ability differences. The covariate-adjusted analysis reported here addresses this concern statistically, but future research should aim for stricter baseline equivalence through matched-group or crossover designs.

Third, the pre-test and posttest were administered in different formats: the pre-test used Google Forms in a laboratory setting, while the posttest was delivered via PsychoPy. Differences in interface familiarity, response modality, and task context may have introduced construct-irrelevant variance that contributed to the observed pre-to-post decline in both groups. Future studies should ensure format consistency across assessment points.

Fourth, the instructional intervention was relatively short in duration and focused primarily on receptive and judgment-based tasks. The absence of accuracy gains may therefore reflect limited opportunities for proceduralization and automatization of phrasal verb knowledge. Longer interventions, increased spacing intervals, or the inclusion of productive practice may be necessary to observe reliable learning effects.

Finally, the study focused exclusively on inseparable phrasal verbs and on acceptability judgments as the primary outcome measure. Future research should extend this line of inquiry to separable constructions, different verb–particle combinations, and additional tasks tapping into comprehension and production.

5. CONCLUSION

This study examined the effects of spaced repetition and chunk-based instruction on adult EFL learners' sensitivity and processing of inseparable phrasal verbs. After controlling for significant baseline differences between groups using a mixed-effects logistic regression with pre-test accuracy as a covariate, no significant Group effect was found at posttest. Both groups declined in accuracy from pre- to post-test across all conditions, and the groups did not differ significantly in the magnitude of pre-to-post change. The only robust finding was a Condition effect: the ungrammatical word order (INSEPvop) was harder for all participants to evaluate correctly, regardless of group assignment. No reading-time differences were observed between groups or conditions.

These findings highlight the methodological importance of establishing baseline equivalence and employing covariate-adjusted analyses in quasi-experimental SLA research. They also draw attention to potential confounds arising from task-format differences between pre- and posttest assessments. Future research with larger samples, randomized designs, and format-consistent assessments will be necessary to draw firmer conclusions about the effectiveness of chunk-based spaced repetition for phrasal verb acquisition in adult EFL learners.

Despite the null instructional effect, the study contributes to the empirical literature by demonstrating the value of pre-test covariate adjustment in this research area and by identifying factors—including baseline non-equivalence, task-format differences, and sample attrition—that should be addressed in future investigations of chunk-based instruction in adult L2 contexts.

¿La repetición espaciada aumenta la retención de verbos frasales en estudiantes adultos de L2?

RESUMEN:

Este estudio cuasi-experimental investiga si la repetición espaciada y la instrucción basada en fragmentos mejoran la sensibilidad de los estudiantes adultos de EFL a las restricciones sintácticas que rigen los verbos frasales inseparables. Un total de 83 estudiantes universitarios de inglés con nivel de competencia B1 fueron reclutados de universidades de Brasil y Colombia para el estudio. El grupo experimental recibió una intervención de seis semanas con actividades de flashcards y repetición espaciada basada en Kahoot de verbos frasales separables e inseparables. El grupo de control no participó en la intervención y siguió el currículo estándar de inglés. El desempeño de los participantes se evaluó utilizando una tarea de juicio de aceptabilidad fuera de línea (AJT) antes de la intervención y medidas de tiempo de lectura en línea administradas a través de PsychoPy después de la intervención. Sin embargo, los dos grupos no fueron equivalentes al inicio, con el grupo de control demostrando un tamaño de efecto moderado para la precisión de AJT antes de la prueba (d de Cohen = 0.73). Una regresión logística de efectos mixtos, que incorporó la precisión pretest como covariable, no encontró un efecto significativo de grupo en el posttest ($p = 0,216$), lo que indica que la aparente ventaja del grupo control en el posttest se atribuyó en gran medida a diferencias de capacidad preexistentes, más que a la intervención instruccional. El único efecto fijo robusto fue la condición: el orden de palabras agramatical (INSEPvop) fue significativamente más difícil de evaluar correctamente que el orden gramatical (INSEPvpo) para todos los participantes, independientemente del grupo. Estos hallazgos resaltan la importancia de la equivalencia inicial y los análisis ajustados por covariables en la investigación cuasiexperimental de la SLA, e identifican factores metodológicos —incluidas las diferencias en el formato de las tareas entre las evaluaciones y la deserción de los participantes— que deberían abordarse en futuras investigaciones sobre la instrucción basada en fragmentos para la adquisición de verbos frasales en estudiantes adultos de inglés como lengua extranjera (EFL).

PALABRAS CLAVE: Verbos frasales. Instrucción basada en fragmentos. Repetición espaciada. Tarea de juicio de aceptabilidad. Estudiantes adultos de EFL.

REFERENCES:

- ABDUNAZAROVNA, M. L. Prepositions in English: a persistent challenge for language learners. **Global Science Review**, v. 3, no. 6, p. 473-480, 2025.
- AL-SHUNNAG, M.A.; AHMED, W.K.; ALWREIKA, E.A., ALHAMMAIDEH, A.; ALQARYOUTI, M.H. A deeper understanding of the syntactic behavior of phrasal verbs. **International Journal of Innovative Research and Scientific Studies**, v. 8, n. 4, p. 167-175, 2025.
- ARVANITIS, P. Self-paced language learning using online platforms. In: DRESSMAN, M.; SADLER, R. W. (ed.). **The handbook of informal language learning**. John Wiley and Sons Ltd, 2019. p. 117–138. DOI: 10.1002/9781119472384.ch8.
- BATES, D.; MÄCHLER, M.; BOLKER, B.; WALKER, S. Fitting linear mixed-effects models using lme4. **Journal of Statistical Software**, v. 67, n. 1, p. 1–48, 2015. DOI: 10.18637/jss.v067.i01.
- BELYAEVA, I. G.; SAMORODOVA, E. A.; VORON, O. V.; ZAKIROVA, E. S. Analysis of innovative methods' effectiveness in teaching foreign languages for special purposes used for the formation of future specialists' professional competencies. **Education Sciences**, v. 9, n. 3, p. 171, 2019. DOI: 10.3390/educsci9030171.
- BOERS, F.; LINDSTROMBERG, S. Experimental and intervention studies on formulaic sequences in a second language. **Annual Review of Applied Linguistics**, v. 32, p. 83-110, 2012. DOI: 10.1017/S0267190512000050.
- CHOFFIN, B.; POPINEAU, F.; BOURDA, Y.; VIE, J. J. DAS3H: Modeling student learning and forgetting for optimally scheduling distributed practice of skills. In: TWELFTH INTERNATIONAL CONFERENCE ON EDUCATIONAL DATA MINING, 2019, p. 29–38.
- DEKEYSER, R. (Ed.). **Practice in a second language: perspectives from applied linguistics and cognitive psychology**. Cambridge: Cambridge University Press, 2007.
- FENG, Y.; YANG, M. Effects of practice types on the acquisition of English phrasal verbs. **Languages**, v. 10, n. 9, p. 214, 2025.
- KOLB, D. A. **Experiential learning: experience as the source of learning and development**. Englewood Cliffs: Prentice Hall, 1984.
- KONDRATJEV, H.; KAHRENS, M. Leveraging experiential learning training through spaced learning. **Journal of Work-Applied Management**, v. 11, n. 1, p. 30–52, 2019. DOI: 10.1108/JWAM-05-2018-0011.
- KUZNETSOVA, A.; BROCKHOFF, P. B.; CHRISTENSEN, R. H. B. lmerTest package: Tests in linear mixed effects models. **Journal of Statistical Software**, v. 82, n. 13, p. 1–26, 2017. DOI: 10.18637/jss.v082.i13.

LAUFER, B.; NATION, P. A vocabulary-size test of controlled productive ability. **Language Testing**, v. 16, n. 1, p. 33-51, 1999. DOI: 10.1177/026553229901600103.

LEI, X.; FATHI, J.; NOORBAKHS, S.; RAHIMI, M. The impact of mobile-assisted language learning on English as a foreign language learners' vocabulary learning attitudes and self-regulatory capacity. **Frontiers in Psychology**, v. 13, p. 872922, 2022. DOI: 10.3389/fpsyg.2022.872922.

LENTH, R. V. **emmeans: Estimated marginal means, aka least-squares means**. R package version 1.10.0. 2024. Disponível em: <https://CRAN.R-project.org/package=emmeans>. Acesso em: 08 maio 2026.

LEWIS, M. **The lexical approach**: the state of ELT and a way forward. Hove: Language Teaching Publications, 1993.

LEWIS, M. et al. **Implementing the lexical approach**: putting theory into practice. Hove: Language Teaching Publications, 1997.

LU, X.; WANG, J. What does the processing of chunks by learners of Chinese tell us? An acceptability judgment investigation. **International Review of Applied Linguistics in Language Teaching**, v. 62, n. 2, p. 571-598, 2024. DOI: 10.1515/iral-2021-0026.

MORTAZAVI, M.; NASUTION, M. K.; ABDOLAHZADEH, F.; BEHROOZI, M.; DAVARPANAH, A. Sustainable learning environment by mobile-assisted language learning methods on the improvement of productive and receptive foreign language skills: a comparative study for Asian universities. **Sustainability**, v. 13, n. 11, p. 6328, 2021. DOI: 10.3390/su13116328.

NAKAGAWA, S.; SCHIELZETH, H. A general and simple method for obtaining R^2 from generalized linear mixed-effects models. **Methods in Ecology and Evolution**, v. 4, n. 2, p. 133–142, 2013. DOI: 10.1111/j.2041-210x.2012.00261.x.

NERGIS, A. Can explicit instruction of formulaic sequences enhance L2 oral fluency? **Lingua**, v. 255, p. 103072, 2021. DOI: 10.1016/j.lingua.2021.103072.

PEIRCE, J. W. et al. **PsychoPy2**: experiments in behavior made easy. Behavior Research Methods, 2019. DOI: 10.3758/s13428-018-01193-y.

PETERS, E. Learning German formulaic sequences: the effect of two attention-drawing techniques. **The Language Learning Journal**, v. 40, n. 1, p. 65-79, 2012. DOI: 10.1080/09571736.2012.658224.

R CORE TEAM. **R: A Language and Environment for Statistical Computing**. R Foundation for Statistical Computing, Vienna, Austria, 2024. <https://www.R-project.org/>.

READ, J. **Assessing vocabulary**. Cambridge: Cambridge University Press, 2000.

_____. The development of a new measure of L2 vocabulary knowledge. **Language Testing**, v. 10, n. 3, p. 355-371, 1993. DOI: 10.1177/026553229301000308.

SUZUKI, Y. (ed.). **Practice and automatization in second language research**: perspectives from skill acquisition theory and cognitive psychology. Taylor and Francis, 2023.

TABARI, M. A.; LEE, H.; HANZAWA, K. Effects of massed and spaced task repetitions on L2 writing task performance and task emotions. **Reading and Writing**, v. 39, p. 29–67, 2026. DOI: 10.1007/s11145-024-10628-2.

TON, N. M.; TRAN, N. T.; DANG, N. T. Promoting Vietnamese learners' knowledge of English phrasal verbs: an action research study. **World Journal of English Language**, v. 13, n. 8, 2023. DOI: 10.5430/wjel.v13n8p108.

WOOD, D. **Formulaic language and second language speech fluency**. London: Bloomsbury Publishing, 2010.