



Enhancing Arabic Vocabulary Acquisition through Mobile-Assisted Language Learning: A Quasi-Experimental Study in an Indonesian Madrasah

Rika Nurmalia^{1*}, Ahmad Rizki Nugrahawan¹, Fatmah Duerasor², Zaimul Hakim³

¹Universitas Muhammadiyah Prof. Dr. Hamka, Indonesia

²Piraya Nawin Khlong Hin Wittaya School, Thailand

³Kolej Islam An-Nabawi, Malaysia

Received, 16 January 2026

Accepted, 15 May 2026

ABSTRACT

Purpose - This study examines the effect of integrating the Kaleela Arabic application into classroom instruction on seventh-grade Arabic vocabulary acquisition in an Indonesian Islamic junior secondary school (madrasah). This study addresses the continuing problem of limited Arabic vocabulary mastery among madrasah learners and the need for empirically examined mobile-assisted language learning (MALL) practices in Arabic education.

Design/Methodology/Approach - A quantitative quasi-experimental design with a nonequivalent control group was employed. The participants were 44 seventh-grade learners at MTs Al-Khairiyah, divided equally into an experimental group ($n = 22$) and a control group ($n = 22$). The experimental group received four weeks of teacher-guided instruction integrating the Kaleela Arabic application, whereas the control group received conventional vocabulary instruction. Data were collected through a 20-item Arabic vocabulary test administered as pre- and posttest. The instrument was validated through expert content review, item validity analysis (Pearson r), and reliability testing (Cronbach's $\alpha = 0.843$). Data analysis included Shapiro-Wilk normality tests, Levene's homogeneity test, pretest equivalence testing, paired sample t -tests within each group, an independent sample t -test for post-test comparison, N-Gain analysis, and Cohen's d -based effect size estimation.

Findings - Pre-test scores confirmed initial group equivalence, $t(42) = 0.468$, $p = 0.642$. A post-test comparison revealed a statistically significant difference between groups, $t(42) = 3.284$, $p = 0.002$, with a large effect size ($d = 0.99$). The experimental group obtained a mean post-test score of 78.18 ($SD = 13.50$) compared with 65.45 ($SD = 12.20$) for the control group. The average N-Gain of the experimental group was 0.57 (moderate category), whereas that of the control group was 0.33 (low-moderate boundary). Variation in individual gains highlights the mediating role of knowledge of prior vocabulary and teacher scaffolding.

Originality/Value - This study provides quasi-experimental evidence on the pedagogically integrated use of Kaleela Arabic for vocabulary acquisition in a madrasah classroom context. It extends MALL research by interpreting learning gains through theoretically grounded mechanisms: repeated exposure, retrieval practice, multimodal input, immediate feedback, and teacher scaffolding, rather than attributing improvement solely to the application.

Paper type - Research paper



Keywords: *Arabic vocabulary acquisition; Mobile-Assisted language learning; Kaleela Arabic; Quasi-experimental design; Multimodal learning.*

***Corresponding author:** Rika Nurmalia, Universitas Muhammadiyah Prof. Dr. Hamka, Indonesia. E-mail: nurmaliar90@gmail.com

1. Introduction

Arabic occupies a distinctive position in Indonesian Islamic education. It is studied not only as a foreign language but also as a medium for accessing the Qur'an, Hadith, classical Islamic scholarship, and contemporary Islamic discourse. In madrasah education, Arabic instruction is expected to support both religious literacy and communicative language competence. However, many students continue to experience difficulty using Arabic beyond controlled classroom tasks, particularly because their lexical repertoire remains limited. For beginning and lower-intermediate learners, vocabulary knowledge provides the foundation from which listening, speaking, reading, and writing skills develop. Without adequate vocabulary, learners cannot decode input, express meaning, or participate in meaningful communication.

Vocabulary acquisition is central to second and foreign language learning because lexical knowledge mediates comprehension, fluency, and accuracy. Nation (2022) argues that learners require substantial vocabulary coverage to process oral and written input effectively. Webb and Nation (2017) further emphasize that vocabulary is learned through repeated encounters, contextual use, retrieval, and gradual strengthening of form-meaning connections. These principles are especially relevant to Arabic learning in Indonesia, where students must navigate unfamiliar phonological patterns, script conventions, root-pattern morphology, and semantic distinctions that differ considerably from Indonesian. In such a context, vocabulary learning cannot rely solely on memorized word lists; learners need repeated, meaningful, multimodal, and interactive exposure to lexical items.

The problem is particularly visible at the Madrasah Tsanawiyah level. Students are often introduced to Arabic vocabulary through teacher explanation, textbook exercises, translation, and memorization. These practices can support initial recognition, but they provide limited opportunities for retrieval, pronunciation practice, multimodal processing, and contextual use. Previous studies in Indonesian Arabic education have reported that low motivation, limited instructional innovation, insufficient engaging media, and monotonous classroom routines contribute to students' difficulty in mastering mufradāt (Hidayat et al., 2025; Khotimah & Nurlaila, 2024; Munir et al., 2023). These findings suggest that the challenge is not simply that students lack vocabulary but that vocabulary learning is often insufficiently supported by instructional design.

Mobile-assisted language learning (MALL) offers a promising response to this instructional challenge. Mobile applications can provide portable access to vocabulary input, audio modelling, visual support, quizzes, feedback, and repeated practice. Meta-analytic studies on MALL have shown that mobile platforms support vocabulary learning by increasing practice frequency, learner engagement, and opportunities for self-paced review (Hwang & Fu, 2019; Lin

& Lin, 2019; Mahdi, 2018). Research on app-based learning, including studies on Duolingo, further suggests that mobile applications can support receptive linguistic knowledge when learning activities are structured and connected to meaningful practice (Loewen et al., 2020). However, app-based learning does not automatically produce deeper language acquisition. Its effectiveness depends on how digital affordances are pedagogically integrated into instruction, how teachers scaffold application use, and how learners interact with the target language.

Kaleela Arabic is a mobile application designed to support Arabic learning through vocabulary exercises, audio pronunciation models, dialogues, quizzes, and interactive practice. These features have pedagogical relevance because they align with several well-established principles of vocabulary acquisition: repeated exposure, retrieval practice, multimodal input, immediate feedback, and contextualized use. From a multimedia learning perspective, the combination of textual, visual, and auditory information can help learners process new vocabulary more effectively when it is designed coherently and does not overload working memory (Mayer, 2021). From a gamification perspective, quizzes, scores, and immediate feedback may increase persistence and engagement when aligned with instructional goals (Sailer & Homner, 2020). These theoretical links make Kaleela Arabic a plausible instructional medium for Arabic vocabulary learning, but plausibility must be supported by empirical evidence.

The current literature on digital Arabic learning has examined several applications, including Quizlet, Quizizz, Bamboozle, Rosetta Stone, Al-Kunuz, Takallam, and Duolingo (Aflisia et al., 2022; Kholis & Nadhif, 2023; Naimah & Syaff'i, 2025; Rachmawati et al., 2022). These studies collectively show that technology-supported Arabic instruction can improve engagement and may support language performance. Nevertheless, several research gaps remain. Many studies focus on learner perception rather than experimental evidence. Studies often examine speaking, listening, or general digital media use rather than vocabulary acquisition specifically. Research frequently describes application features without explaining the vocabulary-learning mechanisms that make those features pedagogically meaningful. Furthermore, research specifically examining Kaleela Arabic in classroom-integrated vocabulary instruction at the Madrasah Tsanawiyah level remains limited.

This study addresses this gap by examining the effect of teacher-guided Kaleela Arabic integration on seventh-grade students' Arabic vocabulary acquisition in an Indonesian madrasah. Rather than treating the application as an independent self-study tool, the study situates Kaleela Arabic within structured classroom instruction under teacher supervision. This distinction is important because mobile applications may be more effective when they complement teacher explanation, guided practice, pronunciation support, and classroom interaction. The study contributes to Arabic language education by providing quasi-experimental evidence on how a mobile-assisted application can support vocabulary learning when embedded in pedagogically structured classroom practice.

The study is guided by the following research questions:

RQ1: Is there a statistically significant difference in Arabic vocabulary post-test performance between students taught using teacher-guided Kaleela Arabic integration and students taught through conventional vocabulary instruction?



RQ2: To what extent does the use of Kaleela Arabic improve students' Arabic vocabulary learning as measured by N-Gain scores?

RQ3: How can the observed learning gains be interpreted through the mechanisms of repeated exposure, retrieval practice, multimodal input, immediate feedback, and teacher scaffolding?

2. Literature Review

2.1 Arabic Vocabulary Acquisition

Vocabulary acquisition is not a single act of memorizing word meanings but a gradual process through which learners build knowledge of form, meaning, pronunciation, grammatical behavior, collocation, and contextual use. Nation (2022) distinguishes multiple dimensions of vocabulary knowledge, including spoken form, written form, word parts, meaning, associations, grammatical functions, collocations, and constraints on use. For beginning Arabic learners, this multidimensionality is particularly demanding because the word form itself may be unfamiliar, the orthographic system differs substantially from Latin script, and pronunciation requires attention to phonemes absent in Indonesian.

Research on vocabulary learning consistently emphasizes repetition, retrieval, and context. Repeated exposure allows learners to strengthen memory traces, but exposure alone is often insufficient. Retrieval practice—in which learners actively recall and use target vocabulary—contributes to stronger retention than passive review (Kornell & Bjork, 2008; Roediger & Butler, 2011). Contextualized vocabulary learning helps learners move beyond isolated translation equivalents toward usable lexical knowledge. In Arabic vocabulary learning, this means learners need to hear, read, pronounce, recognize, and use words in meaningful situations rather than merely memorizing bilingual lists. A pedagogically sound vocabulary tool should therefore provide repeated encounters, active recall opportunities, pronunciation models, visual cues, feedback, and contextualized practice.

2.2 Mobile-Assisted Language Learning and Vocabulary Development

Mobile-assisted language learning refers to language learning supported by portable digital devices such as smartphones and tablets. Its educational value lies in mobility, accessibility, personalization, multimodality, and the possibility of frequent short learning sessions. Meta-analytic studies have reported positive effects of mobile devices on language learning outcomes, particularly vocabulary learning, because mobile platforms enable repeated practice and flexible review (Lin & Lin, 2019; Mahdi, 2018). However, the effect of MALL is not uniform. Outcomes depend on learner age, proficiency, instructional support, application design, assessment type, and intervention duration (Shortt et al., 2023; Stockwell, 2022).

App-based language learning research is particularly relevant to this study. Loewen et al. (2019) found that mobile application use can support second language development, although gains may differ across receptive and productive abilities. Loewen et al. (2020) further demonstrated that app-based instruction can develop receptive linguistic knowledge and aspects of oral communicative ability when learning is sustained. These findings are important for the present study because the Kaleela Arabic intervention was short and classroom-based. Therefore, the expected outcome should be interpreted as improvement in vocabulary

recognition and comprehension rather than broad communicative mastery. In Arabic learning contexts, mobile technologies are increasingly used to support vocabulary, speaking, listening, and reading activities (Aflisia et al., 2022; Naimah & Syafi'i, 2025), supporting the broader premise that digital tools can enrich Arabic learning when integrated with appropriate pedagogy.

2.3 Multimodal Input, Gamification, and Teacher Scaffolding

The pedagogical potential of Kaleela Arabic can be interpreted through multimodal learning theory. Mayer's (2021) cognitive theory of multimedia learning suggests that learners process information through verbal and visual channels, and that learning improves when words and images are integrated coherently. In vocabulary learning, the combination of written Arabic forms, audio pronunciation, and visual or contextual cues can support encoding and retrieval. For Arabic learners, audio support is particularly important because correct pronunciation and sound discrimination are central to early vocabulary acquisition.

Gamification may also contribute to engagement and persistence. Sailer and Homner (2020) found in a meta-analysis that gamification can positively affect cognitive, motivational, and behavioral learning outcomes when game elements are aligned with instructional goals. In mobile vocabulary learning, quizzes, immediate feedback, progress markers, and challenge structures can encourage learners to repeat practice and monitor their progress. Nevertheless, gamification is not inherently effective; if game elements distract from lexical processing, they may produce engagement without learning. Teacher scaffolding therefore remains an essential mediating factor in classroom-based MALL. Teachers can introduce target vocabulary, connect application activities to curriculum objectives, monitor student progress, correct pronunciation, clarify meaning, and transform individual screen-based practice into interactive classroom learning.

2.4 Research Hypotheses

Based on the theoretical framework reviewed above, this study tests the following hypotheses. H1: Students taught using teacher-guided Kaleela Arabic integration will obtain significantly higher Arabic vocabulary post-test scores than students taught through conventional vocabulary instruction. H2: Students in the experimental group will show a moderate or higher learning gain as measured by N-Gain analysis.

3. Method

3.1 Research Design

This study employed a quantitative quasi-experimental design with a nonequivalent control group. A true experimental design was not feasible because students were already organized into intact classroom groups and the school did not permit random assignment of individual students to different instructional conditions. The study therefore used two existing classes: one as the experimental group and one as the control group. This design is appropriate for educational settings in which researchers examine instructional interventions under real classroom conditions while maintaining group comparability through pre-test administration



and equivalent instructional conditions. The independent variable was the instructional treatment (teacher-guided Kaleela Arabic integration versus conventional instruction), and the dependent variable was students' Arabic vocabulary test performance.

3.2 Participants and Sampling

The population consisted of all seventh-grade students of MTs Al-Khairiyah in the 2025/2026 academic year. The sample was selected through purposive sampling. Class VII.1 (n = 22) was assigned as the experimental group and Class VII.2 (n = 22) as the control group, giving a total sample of 44 students. The two classes were selected because they were taught by the same Arabic teacher, used the same curriculum and textbook, followed comparable learning schedules, and were reported by the school to have relatively similar academic profiles. The pre-test was administered to verify initial group equivalence before treatment commenced. Because the design used intact classes rather than random assignment, findings should be interpreted as evidence of treatment contribution rather than definitive causal proof.

This study was conducted with permission from MTs Al-Khairiyah, and participation followed the school's established academic procedures. Student test scores were used exclusively for research purposes and are reported in aggregated form without any identifiable information. Because the participants were junior secondary students, the researcher ensured that data collection did not compromise students' privacy or normal instructional time. Both groups received Arabic instruction aligned with the same curriculum objectives throughout the study period. The control group was offered access to the Kaleela Arabic application after data collection was complete, ensuring equitable access to educational resources.

3.3 Instrument

The research instrument was a 20-item Arabic vocabulary test administered as pre-test and post-test. Items were designed to measure students' ability to recognize, understand, and apply target Arabic vocabulary in simple multiple-choice contexts. Items were developed based on the Arabic learning objectives for seventh-grade Madrasah Tsanawiyah students and adapted from the textbook used at MTs Al-Khairiyah. Vocabulary domains included self-introduction, daily activities, school environment, classroom objects, and simple communicative expressions.

The construct measured in this study is best understood as receptive vocabulary knowledge and contextual vocabulary comprehension. The term vocabulary acquisition is used in this article to refer specifically to improvement in students' tested vocabulary knowledge following instruction, not to claim full productive lexical mastery. Content validity was examined through expert review by specialists in Arabic language education and an Arabic teacher from MTs Al-Khairiyah. The validation process addressed curriculum alignment, item relevance, language clarity, difficulty level, and suitability for seventh-grade learners. Item validity was tested using Pearson product-moment correlation, and internal consistency was examined using Cronbach's Alpha.

3.4 Intervention Procedure

The intervention was conducted over four weeks, with one instructional session per week. Each session lasted approximately 80 minutes. In the experimental class, the Kaleela Arabic application was integrated into face-to-face Arabic vocabulary instruction under teacher supervision. Before the intervention, the teacher received preliminary guidance from the researcher regarding application operation and classroom integration procedures.

Each session followed a structured sequence. First, the teacher introduced learning objectives and target vocabulary. Second, students practised the vocabulary using Kaleela Arabic features such as vocabulary exercises, pronunciation models, dialogue-based activities, and interactive quizzes. Third, students worked individually or in small groups while the teacher monitored progress and provided assistance. Fourth, the teacher facilitated brief classroom reflection, clarified meaning, corrected pronunciation, and connected application practice with textbook content. The control group received conventional instruction using teacher explanation, vocabulary memorization, textbook exercises, and written practice without any application use.

Table 1: *Intervention Design in the Experimental Group*

1	Self-introduction and basic expressions	Audio vocabulary and recognition practice	Guided individual practice and pair repetition	Introduce vocabulary, model pronunciation, monitor use
2	Classroom objects and school vocabulary	Vocabulary exercises and visual/audio input	Small-group vocabulary practice	Clarify meanings and correct pronunciation
3	Daily activities and simple sentences	Dialogue-based practice and quizzes	Guided practice and contextual sentence use	Scaffold sentence construction and feedback
4	Review and consolidation	Interactive quizzes and repeated practice	Individual practice, group review, post-test preparation	Facilitate review and ensure task completion

3.5 Data Analysis

Data were analysed using SPSS. Item validity was examined by correlating each item score with the total score. Reliability was calculated using Cronbach's Alpha. Shapiro-Wilk tests were used to examine normality because each group contained fewer than 50 students. Levene's test was used to examine homogeneity of variance. Pre-test equivalence between the two groups was examined using an independent samples t-test to confirm that any post-test differences could be attributed to the treatment. Within-group improvement was assessed through paired samples t-tests for each group separately. An independent samples t-test was conducted to compare post-test scores between groups. N-Gain analysis was used to measure degree of improvement in the experimental group, categorized as high (> 0.70), moderate



(0.30–0.70), or low (< 0.30) following Hake (1998). Effect size was calculated using Cohen's *d*, with values of 0.20, 0.50, and 0.80 interpreted as small, medium, and large, respectively (Cohen, 1988).

4. Results

4.1 Item Validity and Reliability

The item validity analysis showed that all 20 items obtained *r*-count values exceeding the *r*-table threshold of 0.423 at the 0.05 significance level, indicating that all items were valid for measuring students' Arabic vocabulary knowledge. The reliability analysis produced a Cronbach's Alpha coefficient of 0.843, indicating good internal consistency.

Table 2: *Item Validity Test Results*

1	0.671	0.423	0.002	Valid
2	0.588	0.423	0.005	Valid
3	0.544	0.423	0.011	Valid
4	0.619	0.423	0.004	Valid
5	0.482	0.423	0.028	Valid
6	0.531	0.423	0.014	Valid
7	0.699	0.423	0.001	Valid
8	0.614	0.423	0.004	Valid
9	0.503	0.423	0.021	Valid
10	0.571	0.423	0.008	Valid
11	0.628	0.423	0.003	Valid
12	0.490	0.423	0.025	Valid
13	0.682	0.423	0.002	Valid
14	0.556	0.423	0.010	Valid
15	0.601	0.423	0.005	Valid
16	0.477	0.423	0.031	Valid
17	0.533	0.423	0.013	Valid
18	0.646	0.423	0.003	Valid

19	0.518	0.423	0.018	Valid
20	0.459	0.423	0.045	Valid

Table 3: Reliability Test Result

Cronbach's Alpha	0.843	Good internal consistency
-------------------------	-------	---------------------------

4.2 Assumption Testing

Shapiro-Wilk normality tests showed that all significance values exceeded 0.05, indicating that pre-test and post-test scores in both groups were normally distributed (see Table 4). Levene's test produced a significance value of 0.814, indicating homogeneity of variance across groups (see Table 5). These findings confirm that the parametric assumptions required for independent samples t-tests were met.

Table 4: Normality Test Results (Shapiro-Wilk)

Experimental	Pre-test	0.388	Normal
Experimental	Post-test	0.129	Normal
Control	Pre-test	0.213	Normal
Control	Post-test	0.071	Normal

Table 5: Homogeneity of Variance Test Result (Levene's Test)

Levene's test of homogeneity of variance	0.814	Homogeneous variance
---	-------	----------------------

4.3 Descriptive Statistics and Pre-Test Equivalence

Table 6 presents the descriptive statistics for pre-test and post-test scores of both groups. Pre-test scores were statistically equivalent, $t(42) = 0.468$, $p = 0.642$, confirming that the two groups did not differ significantly in initial vocabulary knowledge before the intervention. This equivalence is important for attributing post-test differences to the instructional treatment



rather than to pre-existing group differences. The experimental group showed a substantially larger mean gain (28.44 points) compared to the control group (17.27 points), indicating that the treatment produced greater vocabulary improvement.

Table 6: Descriptive Statistics for Pre-Test and Post-Test Scores

Pre-test mean	49.74	48.18
Pre-test SD	11.23	10.87
Post-test mean	78.18	65.45
Post-test SD	13.50	12.20
Mean gain	28.44	17.27
N-Gain score	0.57 (Moderate)	0.33 (Low-Moderate)
Pre-test t-test	t(42) = 0.468, p = .642 (not significant)	—

Note. Scores are presented on a 0–100 scale. Pre-test t-test confirms group equivalence before intervention.

4.4 Treatment Effect: Post-Test Comparison and Effect Size

The independent samples t-test revealed a statistically significant difference in post-test scores between the experimental and control groups, $t(42) = 3.284$, $p = 0.002$ (see Table 7). This result supports H1, indicating that students who received teacher-guided Kaleela Arabic integration performed significantly better on the post-test than those who received conventional instruction. The effect size was large (Cohen's $d = 0.99$), indicating that the treatment had a practically meaningful impact on vocabulary performance beyond what would be expected by chance. A 95% confidence interval for the mean difference ranged from 4.72 to 20.74 points.

Table 7: Independent Samples T-Test Result: Post-Test Comparison

Post-test scores	3.284	42	0.002	12.73	0.99 (large)	[4.72, 20.74]
-------------------------	-------	----	-------	-------	--------------	---------------

Note. Cohen's $d = 0.99$ indicates a large effect (Cohen, 1988). CI = confidence interval for the mean difference.

Although this result supports the effectiveness of the intervention, the interpretation should remain cautious because the study used intact classroom groups rather than random

assignment. The result indicates that the treatment contributed to higher vocabulary performance under the conditions of this study; it does not constitute proof that the application will be equally effective in all instructional contexts.

4.5 Within-Group Improvement: Paired Samples T-Tests

Paired samples t-tests were conducted within each group to examine pre-to-post-test improvement. The experimental group showed a statistically significant improvement, $t(21) = 9.91, p < 0.001$, demonstrating that the Kaleela Arabic integration produced substantial within-group gains. The control group also showed a statistically significant improvement, $t(21) = 5.44, p < 0.001$, consistent with the fact that conventional instruction also supported some vocabulary learning. However, the magnitude of improvement was considerably greater in the experimental group, as reflected in the higher mean gain and N-Gain score (see Table 6).

4.6 N-Gain Analysis

The N-Gain analysis showed that the experimental group obtained an average N-Gain score of 0.57 (56.59%), falling in the moderate category. Six students reached the high category, 13 students the moderate category, and 3 students the low category. The lowest individual N-Gain was 0.20 and the highest was 1.00. The control group's mean N-Gain was 0.33 (33.00%), placing it at the low-to-moderate boundary. This distribution indicates that the Kaleela Arabic intervention produced substantially stronger gains than conventional instruction, although individual variation within the experimental group points to differential responsiveness to mobile-assisted practice.

Table 8: N-Gain Score Distribution in the Experimental and Control Groups

High	> 0.70	6 students	1 student
Moderate	0.30-0.70	13 students	9 students
Low	< 0.30	3 students	12 students
Mean N-Gain	—	0.57 (Moderate)	0.33 (Low-Moderate)

4.7 Discussion

The findings indicate that teacher-guided integration of Kaleela Arabic contributed positively to students' Arabic vocabulary learning. The significant post-test difference, large



effect size ($d = 0.99$), and moderate N-Gain suggest that the experimental treatment produced learning benefits substantially exceeding those achieved through conventional vocabulary instruction. The pre-test equivalence test confirms that these differences reflect treatment effects rather than pre-existing ability differences. However, the effect should be interpreted as meaningful but contextually bounded. The data support the claim that Kaleela Arabic was beneficial in this setting under the conditions described; they do not justify a universal claim that the application is superior across all Arabic learning contexts.

The first pedagogical explanation is repeated exposure. Students in the experimental group encountered target vocabulary through teacher explanation, application exercises, audio models, quizzes, and classroom review. This repeated contact strengthened form-meaning connections. Vocabulary research emphasizes that learners require multiple encounters with lexical items before those items become stable and retrievable (Nation, 2022; Webb & Nation, 2017). Conventional instruction typically presents vocabulary once and then moves to memorization and written exercises. Application-supported learning, by contrast, provides repeated encounters within a short period, making practice more frequent, varied, and multimodally rich.

The second explanation is retrieval practice. Kaleela Arabic activities required students to recall vocabulary, select meanings, respond to audio prompts, and repeat pronunciation. Retrieval strengthens memory more effectively than passive review because learners must actively reconstruct lexical knowledge (Roediger & Butler, 2011). The moderate N-Gain score suggests that this active processing supported learning, although the four-session intervention was probably too short to produce high gains for all learners. The three students in the low N-Gain category likely required more instructional time, more teacher support, or more foundational vocabulary preparation before they could benefit fully from application-based practice.

The third explanation is multimodal input. Arabic vocabulary learning involves written forms, pronunciation, and meaning simultaneously. For Indonesian learners, the phonological and orthographic distance between Arabic and Indonesian creates cognitive difficulty that visual-only or text-based instruction often fails to address. Kaleela Arabic provided integrated audio-visual support, allowing students to process vocabulary through multiple channels. Mayer's (2021) cognitive theory of multimedia learning predicts that coherent integration of verbal and visual information supports encoding and retrieval, which appears consistent with the gains observed in this study.

The fourth explanation is immediate feedback and gamified practice. Interactive quizzes and prompt-response cycles allowed learners to identify errors quickly and repeat items until they improved. Sailer and Homner's (2020) meta-analysis found that gamification positively affects cognitive and motivational outcomes when game elements are aligned with

instructional goals. In this study, quizzes and feedback did not function as entertainment; they created repeated opportunities for vocabulary recognition and error correction. Nevertheless, the variation in N-Gain scores confirms that gamified practice alone is insufficient. Students with limited digital familiarity or weaker initial vocabulary knowledge still required direct teacher scaffolding.

The fifth and arguably most theoretically important explanation is teacher-guided integration. The application was not used as a stand-alone self-study tool. The teacher introduced vocabulary, supervised application use, corrected pronunciation, clarified meaning, and connected digital exercises to classroom objectives. This finding supports a balanced view of MALL: mobile applications can enrich instruction, but their effectiveness is mediated by pedagogy. Loewen et al. (2019, 2020) demonstrated that app-based learning can support language development, but outcomes depend on how learners engage with the application and what dimensions of language are assessed. In the present study, teacher guidance transformed application use into structured vocabulary learning, which is consistent with Bower's (2019) technology-mediated learning framework.

The large effect size ($d = 0.99$) is noteworthy and deserves careful interpretation. Al-khresheh (2024) reported comparable effect sizes in a quasi-experimental EFL study, and Mahdi (2018) reported large average effect sizes in a meta-analysis of mobile vocabulary learning. In this study, the large Cohen's d may reflect the relatively controlled nature of the classroom context, the alignment between the application features and the test content, and the teacher scaffolding that reinforced application-based practice. Future studies with larger and more diverse samples should examine whether comparable effect sizes are reproducible across different madrasah contexts.

The findings align with recent Arabic digital learning studies. Aflisia et al. (2022) emphasize the importance of technological adaptation in Arabic education, while Naimah and Syafi'i (2025) show that interactive media can improve Arabic learning performance. Kholis and Nadhif Aflisia (2023) report that Quizlet-supported vocabulary learning improves motivation and practice. The present study extends this literature by providing quasi-experimental evidence with effect size reporting for Kaleela Arabic in a madrasah classroom and by explaining the pedagogical mechanisms that account for the observed gains.

Several limitations must be acknowledged. The intervention lasted only four sessions, which is short for lasting vocabulary acquisition. The sample of 44 students from one madrasah limits generalizability. The 20-item multiple-choice instrument primarily measures receptive vocabulary knowledge rather than productive lexical mastery. The study did not include a delayed post-test, so retention over time cannot be assessed. Motivation, engagement, and digital literacy were not measured with validated instruments. These limitations define the scope of the findings; the strongest conclusion is that teacher-guided Kaleela Arabic integration



produced statistically significant, moderately positive, and practically meaningful improvement in tested Arabic vocabulary knowledge under the conditions of this study.

4.8 Limitations and Directions for Future Research

Several limitations should be considered when interpreting the findings. First, the four-session intervention is short for vocabulary acquisition research. A delayed post-test administered two to four weeks after the intervention would provide stronger evidence of retention. Future studies should include both immediate and delayed post-tests to distinguish short-term performance gains from durable vocabulary learning.

Second, the 20-item multiple-choice instrument is appropriate for receptive vocabulary recognition but cannot capture productive lexical mastery. Future studies should combine receptive tests with productive tasks, oral vocabulary production, sentence completion, or picture description to provide a richer picture of Arabic vocabulary development.

Third, the absence of validated measures of motivation, engagement, and digital literacy means that affective claims remain interpretive. Future research could incorporate learner engagement questionnaires, application usage logs, and classroom observation rubrics to explain individual differences in N-Gain scores.

Fourth, the current analysis relies on independent samples t-tests for group comparison. A stronger analytical approach for nonequivalent control group designs would use ANCOVA with post-test score as the dependent variable and pre-test score as covariate, which provides a more precise estimate of the treatment effect by statistically controlling for pre-test differences. Mixed-effects modelling would be appropriate if future research includes repeated measurements across sessions. Future studies should implement these stronger designs to improve causal inference.

Finally, the study was conducted in one madrasah with 44 students, limiting generalizability. Future research should replicate the study across multiple schools, compare urban and rural madrasah contexts, and examine whether Kaleela Arabic is differentially effective for learners of varying proficiency levels. Comparative studies could also examine Kaleela Arabic alongside other applications such as Duolingo, Quizlet, or teacher-designed vocabulary platforms.

5. Conclusion

This study examined the effect of teacher-guided Kaleela Arabic integration on seventh-grade students' Arabic vocabulary acquisition in an Indonesian madrasah. Pre-test equivalence testing confirmed that the two groups were comparable before the intervention. Post-test comparison revealed a statistically significant group difference, $t(42) = 3.284$, $p = 0.002$, with a large effect size ($d = 0.99$). The experimental group achieved a mean post-test score of 78.18 (SD = 13.50) and a mean N-Gain of 0.57, compared to 65.45 (SD = 12.20) and 0.33 for the control

group. These findings indicate that Kaleela Arabic can support Arabic vocabulary learning when integrated into structured classroom instruction rather than used as an isolated self-study tool.

The findings carry practical implications for Arabic teachers in madrasah contexts. Mobile-assisted vocabulary learning should be planned as part of the lesson architecture rather than appended as a supplementary activity: teachers need to select which vocabulary items will be practised through the application, which features will be used, how much time will be allocated, and how application-based practice will be connected to classroom interaction. Application use should be accompanied by explicit scaffolding, because beginning Arabic learners may struggle with script recognition, pronunciation, and morphological patterns, and teacher modelling and feedback are central in transforming application affordances into structured learning experiences. Teachers should also differentiate support for students who show low learning gains, as learners with weaker prior vocabulary knowledge may need pre-teaching, additional repetition, or peer assistance before using the application effectively. At the institutional level, madrasah administrators should consider providing infrastructure support, device access, clear acceptable-use policies, and teacher professional development to ensure that mobile integration is sustainable and pedagogically purposeful.

This study contributes empirical evidence that a mobile-assisted Arabic learning application, when pedagogically guided by a teacher, can meaningfully support vocabulary learning in a madrasah classroom. Because the intervention was short, the sample limited, and the instrument primarily assessed receptive vocabulary knowledge, the findings should not be generalized beyond the study context without further evidence. Future research should involve larger and more diverse samples, longer treatment periods, delayed post-tests, productive vocabulary tasks, validated engagement measures, and ANCOVA or mixed-effects models to provide stronger causal inference. Despite these limitations, the results support the broader argument that mobile-assisted language learning is most effective when technological affordances are aligned with curriculum objectives and classroom pedagogy—a principle with relevance beyond the Indonesian madrasah setting examined here.

References

- Aflisia, N., Hendriyanto, H., & Suhartini, A. (2022). Arabic language development in boarding schools at the industry era 4.0: Potentials and challenges. *Alsuna: Journal of Arabic and English Language*, 5(1), 45–61. <https://doi.org/10.31538/alsuna.v5i1.2040>
- Agustina, N., Mukhlis, A., & Syamsuddin, A. (2025). Systematic review of innovations in Arabic language curriculum development in madrasahs. *Alsuna: Journal of Arabic and English Language*, 8(2). <https://e-journal.uac.ac.id/index.php/alsuna/article/view/7471>
- Al-khresheh, M. H. (2024). The role of presentation-based activities in enhancing speaking proficiency among Saudi EFL students: A quasi-experimental study. *Acta Psychologica*, 243, Article 104159. <https://doi.org/10.1016/j.actpsy.2024.104159>
- Almelhes, S. S. (2024). Enhancing Arabic language acquisition: Effective strategies for addressing non-native learners' challenges. *Education Sciences*, 14(10), Article 1116. <https://doi.org/10.3390/educsci14101116>



- Annisa, M. N., Arista, D., La Udin, Y., & Wargadinata, W. (2023). Pemerolehan bahasa Arab sebagai bahasa kedua: Kajian psikolinguistik. *Ajamiy: Jurnal Bahasa dan Sastra Arab*, 12(2), 468–484. <https://doi.org/10.31314/ajamiy.12.2.468-484.2023>
- Bower, M. (2019). Technology-mediated learning theory. *British Journal of Educational Technology*, 50(3), 1035–1048. <https://doi.org/10.1111/bjet.12771>
- Chai, C. S., Jong, M. S. Y., Yin, H. B., Chen, M., & Zhou, W. (2019). Validating and modelling teachers' technological pedagogical content knowledge for integrative science, technology, engineering and mathematics education. *Educational Technology & Society*, 22(3), 61–73.
- Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Lawrence Erlbaum Associates.
- Fatahillah, M., & Hafidhoh, H. (2025). Students' perception of the use of Arabic language learning applications (Al-Kunuz and Rosetta Stone) in online learning in Islamic universities. *Jurnal Pendidikan Islam*, 6(11), 1852–1861.
- Hake, R. R. (1998). Interactive-engagement versus traditional methods: A six-thousand-student survey of mechanics test data for introductory physics courses. *American Journal of Physics*, 66(1), 64–74. <https://doi.org/10.1119/1.18809>
- Hidayat, R., Siyam, F. F., Wahyudi, M., & Avilya, A. (2025). Effectiveness of the station rotation blended learning model in mastering mufradat. *Arabiyatuna: Jurnal Bahasa Arab*, 9(1). <https://doi.org/10.29240/jba.v9i1.12431>
- Hwang, G.-J., & Fu, Q.-K. (2019). Trends in the research design and application of mobile language learning: A review of 2007–2016 publications in selected SSCI journals. *Interactive Learning Environments*, 27(4), 567–581. <https://doi.org/10.1080/10494820.2018.1486861>
- Kholis, M. N., & Nadhif, M. F. (2023). The effectiveness of Quizlet.com in Arabic vocabulary learning: Students' perception and acceptance of technology. *Jurnal Ilmiah Bahasa Arab dan Pendidikannya*, 7(1), 1–13.
- Khotimah, H., & Nurlaila, N. (2024). Investigating challenges in Arabic vocabulary mastery among students at MTs Nahdlatul Ulama Maesan Bondowoso. *Guru: Jurnal Pendidikan dan Pembelajaran*, 3(2), 105–114.
- Khuluq, M., Hasanah, M., & Imamah, N. (2024). Mobile-assisted language learning apps: The analysis of Duolingo's content using ACTFL standardization. *Al-Ta'rib: Jurnal Ilmiah Program Studi Pendidikan Bahasa Arab*, 12(2), 229–246.
- Kohnke, L., & Moorhouse, B. L. (2021). Using Kahoot! to gamify learning in the language classroom. *RELC Journal*, 52(3), 769–775. <https://doi.org/10.1177/00336882211040270>
- Kornell, N., & Bjork, R. A. (2008). Learning concepts and categories: Is spacing the 'enemy of induction'? *Psychological Science*, 19(6), 585–592. <https://doi.org/10.1111/j.1467-9280.2008.02127.x>

- Lin, J.-J., & Lin, H. (2019). Mobile-assisted ESL/EFL vocabulary learning: A systematic review and meta-analysis. *Computer Assisted Language Learning*, 32(8), 878–919. <https://doi.org/10.1080/09588221.2018.1541359>
- Loewen, S., Crowther, D., Isbell, D. R., Kim, K. M., Maloney, J., Miller, Z. F., & Rawal, H. (2019). Mobile-assisted language learning: A Duolingo case study. *ReCALL*, 31(3), 293–311. <https://doi.org/10.1017/S0958344019000065>
- Loewen, S., Isbell, D. R., & Sporn, Z. (2020). The effectiveness of app-based language instruction for developing receptive linguistic knowledge and oral communicative ability. *Foreign Language Annals*, 53(2), 209–233. <https://doi.org/10.1111/flan.12454>
- Mahdi, H. S. (2018). Effectiveness of mobile devices on vocabulary learning: A meta-analysis. *Journal of Educational Computing Research*, 56(1), 134–154. <https://doi.org/10.1177/0735633117698826>
- Mayer, R. E. (2021). *Multimedia learning* (3rd ed.). Cambridge University Press. <https://doi.org/10.1017/9781108894333>
- Metruk, R. (2021). The use of smartphone English language learning apps in the process of learning English: Slovak EFL students' perspectives. *Sustainability*, 13(15), Article 8205. <https://doi.org/10.3390/su13158205>
- Munir, D. R., Luthfi, T., & Junita, R. E. (2023). The use of flash card media to increase Arabic vocabulary mastery. *Jurnal Pendidikan Bahasa Arab*, 2(3), 16–35.
- Naimah, I., & Syafi'i, S. (2025). The effectiveness of Bamboozle interactive media in improving the speaking skills of students MAN 1 Lamongan. *Alsuna: Journal of Arabic and English Language*, 8(2), 265–276. <https://doi.org/10.31538/alsuna.v8i2.8665>
- Nation, I. S. P. (2022). *Learning vocabulary in another language* (3rd ed.). Cambridge University Press.
- Rachmawati, M., Rizki, A., & Hidayati, N. (2022). Quizizz application as Arabic online learning evaluation tool in university level. *Jurnal Pendidikan dan Kebudayaan*, 5, 87–100.
- Rahimi, M., & Miri, S. S. (2019). Technology use and acceptance among pre-service teachers of English as a foreign language. *Computer Assisted Language Learning*, 32(5–6), 587–612. <https://doi.org/10.1080/09588221.2018.1531037>
- Roediger, H. L., & Butler, A. C. (2011). The critical role of retrieval practice in long-term retention. *Trends in Cognitive Sciences*, 15(1), 20–27. <https://doi.org/10.1016/j.tics.2010.09.003>
- Sailer, M., & Homner, L. (2020). The gamification of learning: A meta-analysis. *Educational Psychology Review*, 32, 77–112. <https://doi.org/10.1007/s10648-019-09498-w>
- Shortt, M., Tilak, S., Kuznetcova, I., Martens, B., & Akinkuolie, B. (2023). Gamification in mobile-assisted language learning: A systematic review of Duolingo literature from public release of 2012 to early 2020. *Computer Assisted Language Learning*, 36(3), 517–554. <https://doi.org/10.1080/09588221.2021.1933540>



Stockwell, G. (2022). *Mobile assisted language learning: Concepts, contexts and challenges*. Cambridge University Press.

Webb, S., & Nation, P. (2017). *How vocabulary is learned*. Oxford University Press.

Zou, D., Huang, Y., & Xie, H. (2021). Digital game-based vocabulary learning: Where are we and where are we going? *Computer Assisted Language Learning*, 34(5-6), 751-777. <https://doi.org/10.1080/09588221.2019.1640745>