

Fostering Higher-Order Thinking in Arabic Learning: The *At-Tafkir Al-Aliy* Android Application for Seventh-Grade Madrasah Students

Ahmad Labib^{a*}, Fatah Syukur^b, Suja'i^b, S.H. Sheik Mohamed^c

^aSekolah Tinggi Islam Kendal; ^{b,c}Universitas Islam Negeri Walisongo Semarang; ^cS.A. College of Arts and Science, Chennai, India

*Corresponding author: ahmadlabib@stik-kendal.ac.id

Article History Received 23 January 2026 Revised 17 August 2026 Accepted 31 August 2026	Abstract Arabic lessons in madrasah classrooms still lean on tasks that ask students to remember and repeat. The national curriculum expects more: it wants higher-order thinking skills (HOTS) built into daily instruction. Materials that actually push students to analyse, evaluate, and create remain hard to find, and so does digital media built for that purpose. This study tests one such medium, an Android-based Arabic learning application called <i>At-Tafkir Al-Aliy</i>, with seventh-grade students at MTsN 2 Kendal. A quasi-experimental nonequivalent control group design was used, involving 72 students from two intact classes. What the study suggests is narrower than a general claim that technology works. Whether a mobile application lifts higher-order performance in Arabic seems to depend far more on how deliberately HOTS tasks are designed into it than on the novelty of the device itself, and any gains have to be read against ordinary classroom conditions and the fact that the two groups were not formed at random. Abstrak Penelitian ini berangkat dari ketegangan yang masih berlangsung dalam pembelajaran bahasa Arab di tingkat madrasah: meskipun kurikulum nasional menuntut integrasi Higher-Order Thinking Skills (HOTS), bahan ajar yang paling banyak digunakan guru masih bertumpu pada tugas kognitif tingkat rendah, sementara media digital yang benar-benar melatih kemampuan analisis, evaluasi, dan penciptaan masih terbatas. Untuk menjawab kesenjangan tersebut, penelitian ini menguji efektivitas aplikasi pembelajaran bahasa Arab berbasis Android berorientasi HOTS, <i>At-Tafkir Al-Aliy</i>, pada siswa kelas VII MTsN 2 Kendal dengan desain kuasi-eksperimental kelompok kontrol non-ekuivalen yang melibatkan 72 siswa dari dua kelas utuh. Dengan kerangka tersebut, penelitian ini menyiratkan bahwa naik-tidaknya kemampuan berpikir tingkat tinggi siswa dalam bahasa Arab lebih ditentukan oleh seberapa sengaja tugas-tugas HOTS dirancang di dalam pengalaman belajar daripada semata oleh kebaruan teknologi, sekaligus menegaskan pentingnya membaca capaian tersebut secara kritis dalam batas kondisi kelas yang sesungguhnya dan penugasan kelompok yang tidak acak. المخلص ينطلق هذا البحث من إشكالية لا تزال قائمة في تعليم اللغة العربية على مستوى المدرسة المتوسطة : فمع أن المنهج الوطني يطالب بإدماج مهارات التفكير العليا (HOTS)، فإن المواد التعليمية التي يعتمد عليها المعلمون في الغالب لا تزال قائمة على المهام المعرفية الدنيا، كما لا تزال الوسائط الرقمية التي تنمي مهارات التحليل والتقويم والإبداع محدودة. ولمعالجة هذه الفجوة، تقيم هذه الدراسة فاعلية تطبيق لتعلم اللغة العربية قائم على نظام أندرويد وموجه نحو مهارات التفكير العليا، وهو تطبيق «التفكير العالي»، لدى طلاب الصف السابع في المدرسة المتوسطة الإسلامية الحكومية 2 كندال، باستخدام تصميم شبه تجريبي بمجموعة ضابطة غير مكافئة شمل 72 طالبًا من فصلين قائمين. وبهذا الإطار تشير الدراسة إلى أن مدى ارتقاء قدرات الطلاب في التفكير العليا باللغة العربية يتوقف على مدى التصميم المتعمد لمهام التفكير العليا داخل تجربة التعلم أكثر مما يتوقف على حداثة التقنية وحدها، كما تؤكد ضرورة قراءة النتائج في ضوء ظروف الصف الواقعية وتوزيع المجموعات غير العشوائي.
---	---

Keywords: Higher-order thinking skills (HOTS); Mobile-assisted language learning; Arabic language learning; Android-based learning application; Instructional effectiveness

Cite this article: Labib, A., Syukur, F., Suja'i, & Sheik Mohamed, S. H. (2026). Fostering higher-order thinking in Arabic learning: The At-Tafkir Al-Aliy Android application for seventh-grade madrasah students. *Alsina : Journal of Arabic Studies*, 8(2), 282–291. <https://doi.org/10.21580/alsina.8.2.31028>

Introduction

Arabic teaching has absorbed digital tools quickly, and the field is hardly alone in that (Chen, 2022; Peng et al., 2021). Phones and tablets now do a large part of the work. Mobile-assisted language learning (MALL) has settled into a familiar routine: lessons students can poke at, not just listen to (Figueiredo, 2023; Tseng et al., 2022). Owning the tool, though, is not the same as using it well. Higher-order thinking at the secondary level depends on the second, not the first (Kosasih et al., 2022; Togaibayeva et al., 2022). The practical upshot is simple enough. Learning media can now be built for a student's own schedule, at whatever hour suits (Aliakbari & Mardani, 2022; Ebrahimi, 2025).

Progress in technology has not cleared away the older problems in Arabic classrooms at the junior secondary level, or Madrasah Tsanawiyah (MTs). These problems persist (Ainin, 2021). Instruction still tends to be teacher-centred. Vocabulary and grammar get memorised, but the thinking behind them is rarely examined, so critical skills stay undeveloped (Desrani et al., 2022; Umam et al., 2026). Few apps have been built for this setting, and students who find Arabic difficult lose interest quickly (Idhan et al., 2026). What is missing, then, is a medium designed to strengthen analytical work in Arabic over the long term rather than a one-off activity (Al amin & Najah, 2023; Rahmadi et al., 2026).

Curriculum documents now expect higher-order thinking skills (HOTS) throughout Arabic instruction (Rahmawati et al., 2024). Classrooms have not caught up. Most of what a teacher can actually put on the desk still belongs to the lower cognitive levels (C1–C3) of Bloom's taxonomy (Setiawan et al., 2021). Ask for analysis, evaluation, creation, and the cupboard is close to bare (Susantini et al., 2024; Widyaningsih et al., 2021). That mismatch is where this study begins (Chavda et al., 2024; Muhayimana et al., 2022).

An Android application is one way in (Al-Abdullatif & Alsubaie, 2022). Image, sound, and text arrive together, and an exercise built well enough can carry a learner past recognition into reasoning about how the language behaves in context (Al Khaldy et al., 2025; Yusuf, 2021), which is where the reports of better motivation and faster command of grammar and vocabulary tend to come from (Hula et al., 2024; Sahrim et al., 2023). The study contributes a documented HOTS-oriented Android application for Arabic, classroom evidence of its effect, and an honest reading of how far that evidence reaches.

Method

Research Design

The study used a quantitative quasi-experimental design of the nonequivalent control group type. Individual random assignment was not possible for administrative reasons, so randomisation took place at the cluster level, which here means whole intact classrooms (Cham et al., 2024). The layout is shown below:

$$\begin{array}{lcl} \text{Experimental Group (Cluster 1):} & O_1 & X \quad O_2 \\ \text{Control Group (Cluster 2):} & O_3 & - \quad O_4 \end{array}$$

Here O1 and O3 stand for the pre-test scores taken before the intervention, and O2 and O4 for the post-test scores taken after it. X marks the instructional treatment, which was delivered through the Android HOTS application (*At-Tafkir Al-Aliy*); the dash marks conventional instruction from printed textbooks.

Participants and Research Context

Data were collected at MTsN 2 Kendal (a state Madrasah Tsanawiyah, that is, an Islamic junior high school), in Central Java, Indonesia, during the first semester of the 2025/2026 academic year, between September and October 2025. All seventh-grade students formed the population ($N = 216$, spread over six classrooms). Cluster random sampling produced two comparable classrooms and a total sample of 72 students: 36 in Class VII-A, the experimental group, and 36 in Class VII-B, the control group. A post-hoc power analysis showed that this sample was large enough (power $> .80$) to detect a medium-to-large effect at $\alpha = .05$. The two groups were demographically similar. Each contained 36 students, 17 male and 19 female, with ages between 12 and 13 years ($M = 12.4$ in the experimental group and $M = 12.5$ in the control group).

Instructional Application and Matrix Framework

The media evaluated in this study was *At-Ta'fikir Al-Aliy* (التفكير العالي), an interactive Android application. Prior to field implementation, the structural, pedagogical, and technical feasibility of the application was evaluated by four expert validators using standardised 20-item rating scales (scored 1 to 5). To operationalise HOTS across the four core Arabic language skills (Istima', Kalam, Qira'ah, Kitabah), the learning materials and evaluation modules were structured according to the cognitive matrix in Table 1.

Table 1. HOTS content blueprint and cognitive matrix for Arabic language learning.

Language Skill (Maharah)	Cognitive Domain	HOTS Indicator	Exercise / Item Format in App
Listening (<i>Istima'</i>)	C4 (Analysis)	Analysing implicit contextual information from spoken dialogue.	Audio-based scenario analysis requiring students to infer speakers' backgrounds and motives.
Speaking (<i>Kalam</i>)	C5 (Evaluation)	Assessing structural correctness and contextual appropriateness of dialogue options.	Comparative evaluation of conversational scripts to choose grammatically and contextually optimal responses.
Reading (<i>Qira'ah</i>)	C4 & C5 (Analysis & Evaluation)	Deconstructing paragraph structures and evaluating main arguments/inferences.	Textual analysis tasks where students dissect complex passages, identify main ideas, and critique statements.
Writing (<i>Kitabah</i>)	C6 (Creation)	Synthesising vocabulary, tarkib (grammar rules), and ideas into original contextual sentences/passages.	Interactive sentence reconstruction and paragraph drafting tasks based on situational prompts.

Research Instruments

Data were collected using standardised pre-test and post-test assessments consisting of 25 multiple-choice items with structured rationale options designed to evaluate Bloom's revised cognitive levels: C4 Analysis (10 items), C5 Evaluation (8 items), and C6 Creation/Synthesis (7 items), following Setiawan et al. (2021) and Susantini et al. (2024).

Content and construct validity were established by two senior Arabic language education professors, yielding a Content Validity Index (CVI) of 0.92. With regard to empirical reliability, pilot testing on 30 non-sample students produced a high internal consistency score using Cronbach's alpha ($\alpha = .864$); item discrimination indices ranged from 0.38 to 0.72 and difficulty indices ranged from 0.35 to 0.68, which suggests adequate diagnostic suitability.

Data Analysis Procedures

Statistical analysis was conducted using IBM SPSS Statistics.

Descriptive statistics covering the mean (M), standard deviation (SD), minimum, and maximum scores were computed first. Assumption testing assessed normality using the Kolmogorov-Smirnov test and variance homogeneity using Levene's Test for Equality of Variances. Baseline equivalency was examined through an

independent samples *t* test on the pre-test scores to ensure comparability between the intact classes, while hypothesis testing applied the same test to the post-test scores ($\alpha = .05$, two-tailed). Effect size and confidence intervals were derived by calculating Cohen's *d* and the 95% confidence intervals for mean differences (Evans & Yuan, 2022; Ponce-Renova, 2022). The last step, the normalised gain (N-Gain), used Hake's formula:

$$\text{N-Gain} = \frac{\text{Posttest Score} - \text{Pretest Score}}{\text{Maximum Score} - \text{Pretest Score}}$$

Hake's criteria set the bands: low below 0.30, medium from 0.30 to just under 0.70, high at 0.70 or above.

Results

Expert Validation Analysis

Four experts appraised the application for pedagogical and technical validity before it reached any classroom. Table 2 gives the results by domain.

Table 2. Summary of expert validation assessment.

No	Expert Validation Category	Evaluated Aspects (20 Items Each)	Mean Score (1–5)	Percentage	Validity Category
1	Media Expert	Android Architecture & Audio-Visual Communication	4.87/5.00	97.5%	Very Feasible / Valid
2	Information technology (IT) systems expert	User interface/experience (UI/UX) design, performance, and functional alignment	4.00/5.00	80.0%	Feasible / Valid
3	Subject Matter Expert	Educational Alignment & Arabic Content Accuracy	4.00/5.00	80.0%	Feasible / Valid
4	HOTS Method Expert	Cognitive Alignment (C4–C6), Pedagogy, & Assessment	4.21/5.00	84.5%	Very Feasible / Valid

The four ratings did not agree, and the gap needs explaining. Highest came from the media expert (97.5%). The IT systems and subject matter experts each returned 80.0%. Comments account for the spread. Performance offline drew stricter scrutiny from the IT reviewers, who wanted the application to hold up under weaker conditions. The subject matter reviewers asked for adjustments to Arabic typography and diacritical marks. All four scores remained above the feasibility threshold adopted for this study (> 70.0%).

Prerequisite Assumption Testing

Parametric testing rests on two assumptions, so both were checked before anything else. Normality held for the experimental class post-test, $D(36) = 0.112$, $p = .200$, where *D* is the Kolmogorov-Smirnov statistic. It held for the control class too, $D(36) = 0.124$, $p = .185$.

Baseline Equivalency and Experimental Effectiveness

Before any treatment effect was examined, the two groups were compared on the pre-test. Experimental and control means were close, $M = 57.36$ ($SD = 10.38$) against $M = 58.19$ ($SD = 11.47$), and the difference was not significant, $t(70) = -0.322$, $p = .748$. The two intact groups therefore started from a comparable point.

Post-test scores were examined once the intervention ended. Table 3 shows both groups improving, with the experimental class gaining more, and Table 4 tests whether that gap holds up.

Table 3. Descriptive statistics of pre-test and post-test scores.

Group	Phase	<i>N</i>	Mean Score	Std. Deviation	Minimum	Maximum
Experimental Class	Pre-test	36	57.36	10.38	40.00	85.00
	Post-test	36	87.60	8.24	65.00	100.00

Group	Phase	N	Mean Score	Std. Deviation	Minimum	Maximum
Control Class	Pre-test	36	58.19	11.47	40.00	80.00
	Post-test	36	76.48	10.83	55.00	100.00

Table 4. Independent samples *t*-test inferential analysis (post-test comparison).

Statistical Parameter	Value	Critical Threshold / Format	Statistical Decision & Interpretation
Mean Difference	11.12	–	Experimental class scored higher
Calculated <i>t</i> -value (<i>t</i> _{count})	4.903	<i>t</i> (70) = 4.903	$ t_{\text{count}} > t_{\text{critical}}$
Critical <i>t</i> -value (<i>t</i> _{critical})	1.994	<i>df</i> = 70, α = .05 (two-tailed)	The null hypothesis (<i>H</i> ₀) is rejected
Significance (<i>p</i> -value)	$p < .001$	($p < .001$, non-zero)	Statistically Significant
95% Confidence Interval	[6.60, 15.64]	95% CI of Difference	Difference bounded away from 0
Effect Size (Cohen's <i>d</i>)	1.15	$d \geq 0.80$ (Large)	Substantial practical impact

With $t(70) = 4.903$ and $p < .001$, the null hypothesis was rejected. Performance after the intervention was significantly higher in the experimental group.

Learning Effectiveness via Normalised Gain (N-Gain)

N-Gain was higher in the experimental class than in the control class (Table 5).

Table 5. Results of average N-Gain score calculations. N-Gain is a derived index, so no standard deviation is reported.

Group	Mean N-Gain Score	N-Gain Percentage (%)	Effectiveness Category
Experimental Class	0.681	68.1%	Medium Gain
Control Class	0.399	39.9%	Medium Gain

Specific Cognitive Domain Analysis (HOTS)

Item-level scores were aggregated with item weighting (C4 = 10 items, C5 = 8 items, C6 = 7 items), producing item-weighted pre-test and post-test means of 57.92% and 85.76% that closely parallel the composite means reported in Table 3 (57.36 and 87.60). The relative gain on each HOTS indicator is reported in Table 6.

Table 6. Item-weighted student performance across HOTS cognitive indicators (experimental group). Percentages are item-weighted means; no dispersion statistic is reported.

Cognitive Domain	Pre-test Score (%)	Post-test Score (%)	Absolute Gain (%)	Primary Cognitive Improvement Area
C4 (Analysing)	62%	89%	+27%	Analysing syntax (tarkib) and context in paragraphs.
C5 (Evaluating)	58%	85%	+27%	Detecting and correcting grammatical errors in dialogue.
C6 (Creating)	52%	82%	+30%	Constructing original, contextual Arabic passages.

Android Application At-Taḥkīr Al-ʿAlīy

Figure 1 shows several screens from inside the *At-Taḥkīr Al-ʿAlīy* application.

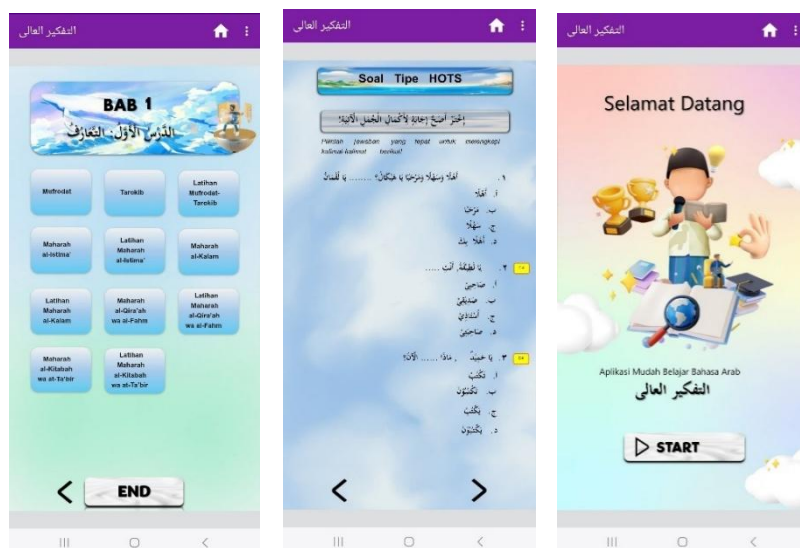


Figure 1. Three main screens of the *At-Taḥkīr Al-ʿAlīy* Android application.

This Android app can be downloaded at the link: https://drive.google.com/file/d/1fx11mL7xfoWZ9KKYCiDhHv5EMjxoVf9/view?usp=drive_link. The same download is available through the quick-response (QR) code in Figure 2.



Figure 2. Quick-response (QR) code of the *At-Taḥkīr Al-ʿAlīy* application.

Discussion

In these seventh-grade classes, something shifted in how students handled Arabic: the experimental class finished the post-test 11.12 points ahead ($p < .001$, $d = 1.15$), a margin too wide for novelty alone to explain, and it points to students being asked to do cognitive work that reaches past memorisation, and doing it.

Digital Transformation in Arabic Pedagogy

Earlier work points the same way, with Mujiono (2023) tying mobile-assisted language learning (MALL) to higher proficiency and Kessler et al. (2025) landing on something similar. Part of the gain here may simply be access: students worked through the analytical exercises when it suited them, not only during class. Instant feedback and multimedia support are what Alkasirah et al. (2026) credit for sustained engagement (see also Fang et al., 2021), and for the Society 5.0 setting Alfurochmatin and Jusoh (2024) push the point further, arguing that static presentation no longer suffices and that Arabic instructional media should move toward environments adaptive enough for independent study (see also Wiboolyasarini & Jinowat, 2024).

Cognitive Elevation from LOTS to HOTS

Rote vocabulary and literal translation have long carried Arabic teaching in madrasahs (Muhayimana et al., 2022; Rahmawati et al., 2024). Chavda et al. (2024, p. 96) argue that higher-order language ability grows only when materials and assessments are matched to Bloom's taxonomy, and what happened here fits that claim (see also Ali et al., 2024). Embedding C4–C6 tasks in the modules turned drill into analytical problem solving and original writing. Silviya (2024) found much the same with sharf (morphology), where mobile test and learning instruments left students able to diagnose and analyse structural patterns on their own.

Pedagogical Framework and Socio-Constructivism

Vygotsky's socio-constructivist framework fits this case. The application works as a digital scaffold, and the graded exercises act as steps through the zone of proximal development, which is roughly where Öztürk and Çakıroğlu (2021) locate the self-regulated study habits built by mobile access combined with structured HOTS activity, and may be where part of the edge held by the experimental group comes from. Nur et al. (2023) came at the question differently, finding that gamified, interactive formative assessment of the kind digital platforms are good at raised motivation, supplied scaffolding when needed, and grew vocabulary alongside contextual understanding.

Alternative Explanations, Boundary Conditions, and Threats to Validity

Not all of this gain need have anything to do with the application: meta-analyses of mobile-assisted language learning do report positive average effects, but the same reviews name a novelty effect and a redundancy effect (Peng et al., 2021), and advantage fades once a technology stops being new or once it only repeats what the class already does. Restraint is warranted on the effect size as well, since small studies tend to yield bigger and more scattered effects than large ones (Evans & Yuan, 2022), significance on its own says little about practical superiority (Ponce-Renova, 2022), and two widely used commercial language apps produced comparable gains from very different designs (Kessler et al., 2025), which hints that engagement and continued use may carry more weight than the medium; motivation, self-efficacy, and the felt relevance of the content drive satisfaction and perceived usefulness, pointing the same way (Togaibayeva et al., 2022).

Whether the findings hold elsewhere is unresolved, because the effect of mobile learning varies with proficiency, treatment length, the skill taught, and the setting (Chen, 2022; Tseng et al., 2022; Mujiono, 2023), and individual and sociocultural factors narrow it further (Figueiredo, 2023). Arabic classrooms add obstacles of their own: low motivation and limited proficiency among students, competency descriptors open to interpretation, and a thin supply of HOTS materials all press on the teacher (Ainin, 2021), whose ability to plan, deliver, and assess higher-order tasks decides whether a HOTS innovation succeeds at all (Kosasih et al., 2022). None of this cancels the present results. It qualifies them: the application appears to have contributed, but how much of the gain it explains, and where it would repeat, this design cannot determine (Cham et al., 2024).

Internal validity is exposed on several fronts. Teachers come first, since both classes had certified Arabic teachers working from standardised lesson plans but delivery and enthusiasm differ, and the design cannot dismiss the effect. Novelty is another: motivation in the experimental group could have risen simply because the devices were new, a Hawthorne-type effect that the design cannot strip away from the treatment. Classroom dynamics complete the list, because intact classes carry unobserved group processes that could have added small confounding variance, though the baseline comparison found no significant difference to begin with ($p = .748$).

Conclusion

An Android-based Arabic environment built around higher-order thinking did appear to move learners past memorisation and literal translation. Where that shows most clearly is not in any single score but in the kind of work students were asked to do: analysing how the language is built, judging whether a choice fits the context, and writing original Arabic of their own. The value of the study, then, is not a verdict on one product. It is evidence that HOTS can be designed into mobile Arabic learning on purpose, and defended.

The design, however, will not support an unqualified claim of superiority. The two classes already existed and were not formed by random assignment, the intervention ran only six weeks, and the outcomes came mainly from structured written tests. Any of those could put part of the difference down to context or to the way learning was measured rather than to the application. A stronger study would follow learners for longer

and in more than one madrasah, set test scores beside performance-based and qualitative evidence, and watch closely how teachers and students actually handle HOTS tasks in class.

Acknowledgments

Our thanks go to the Arabic language teachers and seventh-grade students of MTsN 2 Kendal, and to the expert validators, whose assessment of the media and test instruments shaped the final version.

Conflict of Interest

The authors declare no competing interests and report no financial interests or personal relationships that could be seen to have influenced this work.

Declaration of AI Use

AI-assisted tools were used only for language editing. The authors reviewed, verified, and approved all content, analyses, and interpretations, and accept full responsibility for the originality, accuracy, and integrity of the published work; research data were not generated by AI, and neither was the substantive scholarly content of the article.

References

- Ainin, M. (2021). Challenges in applying higher order thinking skills assessment in the teaching of Arabic in Indonesia. *KnE Social Sciences*, 5(3), 413–420. <https://doi.org/10.18502/kss.v5i3.8564>
- Al-Abdullatif, A. M., & Alsubaie, M. A. (2022). Using digital learning platforms for teaching Arabic literacy: A post-pandemic mobile learning scenario in Saudi Arabia. *Sustainability*, 14(19), Article 11868. <https://doi.org/10.3390/su141911868>
- Al amin, Y., & Najah, Z. (2023). Developmental research: An Android-based Arabic language learning application as a means to improve middle school students' learning outcomes. *Language, Technology, and Social Media*, 1(1), 52–64. <https://doi.org/10.70211/ltsm.v1i1.26>
- Al Khaldy, M. A., Shaheen, A., Al-Shaikh, A., Alzyadat, W., & Alhroob, A. (2025). Android application for children to learn: Arabic alphabets and numbers. *International Journal of Interactive Mobile Technologies (IJIM)*, 19(5), 115–127. <https://doi.org/10.3991/ijim.v19i05.52695>
- Alfurochmatin, A., & Jusoh, A. M. B. M. (2024). Tren pengembangan media pembelajaran bahasa Arab di Indonesia era 5.0. *Alsina: Journal of Arabic Studies*, 6(1), 1–26. <https://doi.org/10.21580/alsina.6.1.21278>
- Ali, M. M., Alaa, A. M., & Shahnaz, A. (2024). The impact of mobile learning in English language classrooms in Pakistan. *Asian-Pacific Journal of Second and Foreign Language Education*, 9(1), Article 62. <https://doi.org/10.1186/s40862-024-00274-0>
- Aliakbari, M., & Mardani, M. (2022). Mobile-assisted language learning and its effects on learners' speaking development. *Education Research International*, 2022, 1–14. <https://doi.org/10.1155/2022/9043326>
- Alkasirah, N. A. M., Mohamad, M., Sanmugam, M., & Ramdas, G. (2026). Development of mobile application-based digital drama for Arabic language to improve students' vocabulary. *Journal of Education and Learning (EduLearn)*, 20(2), 1240–1249. <https://doi.org/10.11591/edulearn.v20i2.23655>
- Cham, H., Lee, H., & Migunov, I. (2024). Quasi-experimental designs for causal inference: An overview. *Asia Pacific Education Review*, 25(3), 611–627. <https://doi.org/10.1007/s12564-024-09981-2>
- Chavda, M., Patel, H., & Bhatt, H. (2024). Quality education through writing: Aligning learning objectives in learning materials and question papers using Bloom's taxonomy. *Quality Assurance in Education*, 32(1), 96–110. <https://doi.org/10.1108/QAE-03-2023-0045>

- Chen, M.-L. (2022). The impact of mobile learning on the effectiveness of English teaching and learning: A meta-analysis. *IEEE Access*, 10, 38324–38334. <https://doi.org/10.1109/ACCESS.2022.3165017>
- Desrani, A., Arifa, Z., & Susanto. (2022). The impact of discovery learning model based on high order thinking skills (HOTS) in learning Arabic on students' analytical thinking skills. *Taqdir*, 7(2), 213–228. <https://doi.org/10.19109/taqdir.v7i2.9910>
- Ebrahimi, M. (2025). Ubiquitous learning: The effect of LingAR application on EFL learners' language achievement and the realization of their motivation towards mobile learning. *Interactive Learning Environments*, 33(8), 4699–4717. <https://doi.org/10.1080/10494820.2022.2041044>
- Evans, D. K., & Yuan, F. (2022). How big are effect sizes in international education studies? *Educational Evaluation and Policy Analysis*, 44(3), 532–540. <https://doi.org/10.3102/01623737221079646>
- Fang, W.-C., Yeh, H.-C., Luo, B.-R., & Chen, N.-S. (2021). Effects of mobile-supported task-based language teaching on EFL students' linguistic achievement and conversational interaction. *ReCALL*, 33(1), 71–87. <https://doi.org/10.1017/S0958344020000208>
- Figueiredo, S. (2023). The effect of mobile-assisted learning in real language attainment: A systematic review. *Journal of Computer Assisted Learning*, 39(4), 1083–1102. <https://doi.org/10.1111/jcal.12811>
- Hula, I. R. N., Loni, J., & Sarif, S. (2024). Development of Android-based learning media using iSpring Suite to improve Arabic vocabulary mastery among students. *Edulab: Majalah Ilmiah Laboratorium Pendidikan*, 9(2), 222–239. <https://doi.org/10.14421/edulab.2024.92.04>
- Idhan, M., Jabir, M., Ruslin, R., Asmawi, M. N., Arafah, B., Ubadah, Ubay, & Fatima. (2026). Teachers' experiences with mobile apps for Arabic language learning: Insights from Islamic schools in Palu, Indonesia. *Journal of Language Teaching and Research*, 17(1), 338–350. <https://doi.org/10.17507/jltr.1701.34>
- Kessler, M., Loewen, S., & Gönülal, T. (2025). Mobile-assisted language learning with Babbel and Duolingo: Comparing L2 learning gains and user experience. *Computer Assisted Language Learning*, 38(4), 690–714. <https://doi.org/10.1080/09588221.2023.2215294>
- Kosasih, A., Supriyadi, T., Firmansyah, M. I., & Rahminawati, N. (2022). Higher-order thinking skills in primary school: Teachers' perceptions of Islamic education. *Journal of Ethnic and Cultural Studies*, 9(1), 56–76. <https://doi.org/10.29333/ejecs/994>
- Muhayimana, T., Kwizera, L., & Nyirahabimana, M. R. (2022). Using Bloom's taxonomy to evaluate the cognitive levels of primary leaving English exam questions in Rwandan schools. *Curriculum Perspectives*, 42(1), 51–63. <https://doi.org/10.1007/s41297-021-00156-2>
- Mujiono. (2023). Mobile-assisted language learning intervention and its effect on English language proficiency of EFL learners: A meta-analysis. *Theory and Practice in Language Studies*, 13(5), 1124–1135. <https://doi.org/10.17507/tpsls.1305.05>
- Nur, A. M., Mudrofin, M., & Bukhori, N. F. (2023). Taṭwīr adawāt al-taqwīm al-takwīnī al-qā'imah 'alā Kahoot li-tarqīyat itqān al-mufradāt fī ta'lim al-lughah al-'arabīyah bil-madrasah al-miḥanīyah. *Alsina: Journal of Arabic Studies*, 5(2), 257–274. <https://doi.org/10.21580/alsina.5.2.22933>
- Öztürk, M., & Çakıroğlu, Ü. (2021). Flipped learning design in EFL classrooms: Implementing self-regulated learning strategies to develop language skills. *Smart Learning Environments*, 8(1), 1–20. <https://doi.org/10.1186/s40561-021-00146-x>
- Peng, H., Jager, S., & Lowie, W. M. (2021). Narrative review and meta-analysis of MALL research on L2 skills. *ReCALL*, 33(3), 278–295. <https://doi.org/10.1017/S0958344020000221>
- Ponce-Renova, H. F. (2022). Comparing effect sizes and their confidence intervals: A primer on equivalence testing in educational research. *Journal of New Approaches in Educational Research*, 11(2), 209–225. <https://doi.org/10.7821/naer.2022.7.930>
- Rahmadi, L., Effendi, M. J., Pratama, S., & Fransiska, W. (2026). Enhancing student engagement in Arabic

- language learning: An Android-based interactive application developed using MDLC. *Ingénierie des Systèmes d'Information*, 31(3), 1001–1010. <https://doi.org/10.18280/isi.310329>
- Rahmawati, Febriani, S. R., & Yasmadi. (2024). Assessment of textbook published by Ministry of Religion Republic Indonesia: HOTS, MOTS, or LOTS? *Ijaz Arabi Journal of Arabic Learning*, 7(3), 1051–1063. <https://doi.org/10.18860/ijazarabi.v7i3.27762>
- Sahrim, M., Mohamad Soad, N. F. A., & Asbulah, L. H. (2023). Augmented reality technology in learning Arabic vocabulary from the perception of university students. *International Journal of Interactive Mobile Technologies (iJIM)*, 17(21), 79–96. <https://doi.org/10.3991/ijim.v17i21.41277>
- Setiawan, J., Sudrajat, A., Aman, & Kumalasari, D. (2021). Development of higher order thinking skill assessment instruments in learning Indonesian history. *International Journal of Evaluation and Research in Education (IJERE)*, 10(2), 545–552. <https://doi.org/10.11591/ijere.v10i2.20796>
- Silviya, A. (2024). Taṭwīr 'adawāt ikhtibār ta'lim al-ṣarf al-mu'tamidah 'alā niẓām 'andrūyd: Dirāsah taṭbīqīyah 'alā ṭalabah ma'had Dār al-Najāh al-'Islāmī. *Alsina: Journal of Arabic Studies*, 6(1), 63–80. <https://doi.org/10.21580/alsina.6.1.23364>
- Susantini, E., Sari, Y. M., Marzuqi, M. I., & Asteria, P. V. (2024). EduQuestioning mobile learning application: A catalyst for developing HOTS-based assessment questions referring to revised Bloom's taxonomy. *Research and Practice in Technology Enhanced Learning*, 20, Article 022. <https://doi.org/10.58459/rptel.2025.20022>
- Togaibayeva, A., Ramazanova, D., Yessengulova, M., Yergazina, A., Nurlin, A., & Shokanov, R. (2022). Effect of mobile learning on students' satisfaction, perceived usefulness, and academic performance when learning a foreign language. *Frontiers in Education*, 7, Article 946102. <https://doi.org/10.3389/educ.2022.946102>
- Tseng, W.-T., Chen, S., Wang, S.-P., Cheng, H.-F., Yang, P.-S., & Gao, X. A. (2022). The effects of MALL on L2 pronunciation learning: A meta-analysis. *Journal of Educational Computing Research*, 60(5), 1220–1252. <https://doi.org/10.1177/07356331211058662>
- Umam, M., Alfian, M., Huda, I. S., & Maziyah, L. (2026). Analysis of internal quality and cognitive distribution in Grade VII MTs Arabic textbook. *Tatsqifiy: Jurnal Pendidikan Bahasa Arab*, 7(2), 169–178. <https://doi.org/10.30997/tjpba.v7i2.25144>
- Wiboolyasarini, W., & Jinowat, N. (2024). Revolutionising dual language learning for young learners with the RILCA app: A mobile-assisted approach. *Journal of Computers in Education*, 11(2), 575–593. <https://doi.org/10.1007/s40692-023-00271-1>
- Widyaningsih, S. W., Yusuf, I., Prasetyo, Z. K., & Istiyono, E. (2021). The development of the HOTS test of physics based on modern test theory: Question modeling through e-learning of Moodle LMS. *International Journal of Instruction*, 14(4), 51–68. <https://doi.org/10.29333/iji.2021.1444a>
- Yusuf, K. (2021). The role of mobile applications in learning Arabic-Indonesian translation. *Journal of Physics: Conference Series*, 1933(1), Article 012024. <https://doi.org/10.1088/1742-6596/1933/1/012024>