



Developing and Evaluating an Arabic Speaking Curriculum in an Indonesian Islamic Boarding School: A 4-D Model-Based Quasi-Experimental Study

تطوير منهج لتعليم مهارة الكلام وتقويم فاعليته في معهد إسلامي داخلي بإندونيسيا: دراسة قائمة على نموذج الأبعاد الأربعة بتصميم شبه تجريبي

Gamil Taher Ahmed^{1*}, Dindin Solahudin², Isop Syafei³, Dedih Wahyudin⁴

¹Al-Saeed University, Yemen

^{2,3,4}Postgraduate Program in Islamic Education, UIN Sunan Gunung Djati Bandung, Indonesia

*Corresponding E-mail: gameelazzani8080@gmail.com

ABSTRACT

Arabic curricula for non-native learners may include all four language skills while providing insufficient structure for sustained oral production in classroom implementation. This study developed and evaluated speaking-focused instructional units for first-year middle-school students at Al-Wadi Al-Mubarak Institute, Bogor, Indonesia. Research and development followed the define, design, and develop stages of the Thiagarajan 4-D model and was combined with a nonrandomized quasi-experimental pretest-posttest control-group design. The available field sample comprised 32 students, with 16 in each group; no documented a priori power analysis was available. Data were collected through interviews, observation, curriculum analysis, expert-validation forms, a speaking test, and an analytic oral-performance rubric. The surviving study records do not document independent multiple rating or an inter-rater reliability coefficient; this measurement limitation is therefore acknowledged. The product comprised speaking units for Part One of Arabic Between Your Hands in a student book and teacher guide, integrating communicative language teaching, task-based learning, cooperative work, role-play, formative assessment, self-assessment, and portfolios. The experimental-group mean increased from 9.75 to 17.19, compared with 9.38 to 10.63 in the control group. The adjusted posttest difference was 6.272 points (95% CI [5.688, 6.856]); ANCOVA yielded $F(1, 29) = 481.853$, $p < .001$, partial $\eta^2 = .943$. This effect size describes variance associated with group membership after adjustment within this sample, not the percentage of improvement caused by the curriculum. The study offers a methodological and contextual contribution by operationalizing needs analysis, the 4-D process, communicative tasks, and performance assessment within a boarding-school curriculum. Findings remain limited by the small, single-institution, nonrandomized sample and undocumented inter-rater reliability.

Keywords: Arabic curriculum development; speaking skills; communicative language teaching; task-based learning; Thiagarajan 4-D model

مستخلص البحث

يواجه تعليم العربية للناطقين بغيرها فجوة بين اشتغال المنهج على المهارات الأربع وبين محدودية الوقت والتنظيم المخصصين للإنتاج الشفهي في التطبيق الصفّي. هدفت هذه الدراسة إلى تطوير وحدات تعليمية مركزة على مهارة الكلام وتقويم فاعليتها لدى طلاب الصف الأول المتوسط في معهد الوادي المبارك ببوجور. استخدمت الدراسة منهج البحث والتطوير وفق مراحل التعريف والتصميم والتطوير من نموذج ثياغاراجان رباعي الأبعاد، مع تصميم شبه تجريبي غير عشوائي ذي مجموعتين واختبارين قبلي وبعدي. بلغ عدد المشاركين المتاحين للتطبيق ٣٢ طالباً، بواقع ١٦ طالباً في كل مجموعة. ولم يتوافر في سجلات الدراسة تحليل قدرة إحصائية قبلي موثق. جمعت البيانات بالمقابلات والملاحظة وتحليل المنهج وتحكيم الخبراء واختبار الكلام وبطاقة تقدير تحليلية للأداء الشفهي. ولا توثق السجلات المتاحة تعدد المقومين المستقلين أو معامل اتفاق بينهم؛ ولذلك صرح بهذه المحدودية القياسية. وتمثل المنتج في وحدات الكلام في الجزء الأول من كتاب «العربية بين يديك»، في صورة كتاب للطالب ودليل للمعلم، مع دمج المدخل التواصل والتعلم القائم على المهام والعمل التعاوني ولعب الأدوار والتقويم التكويني والذاتي وملف الإنجاز. ارتفع متوسط

Citation:

Ahmed, G. T., Solahudin, D., Syafei, I., & Wahyudin, D. (2026). Developing and Evaluating an Arabic Speaking Curriculum in an Indonesian Islamic Boarding School: A 4-D Model-Based Quasi-Experimental Study. *Al-Muyassar: Journal of Arabic Education*, 5(3), 1302-1315.

التجريبية من ٩,٧٥ إلى ١٧,١٩، مقابل ارتفاع الضابطة من ٩,٣٨ إلى ١٠,٦٣. وبلغ الفرق المعدل ٦,٢٧٢ درجات بفاصل ثقة ٩٥٪ بين ٥,٦٨٨ و ٦,٨٥٦، وأظهر تحليل التباين المصاحب أثرا دالا للمجموعة (قيمة ف = ٤٨١,٨٥٣، ومستوى الدلالة أقل من ٠,٠٠١، ومربع إيتا الجزئي = ٠,٩٤٣). وتمثل الدراسة إسهما منهجيا وسياقيا في تشغيل تحليل الاحتياجات ونموذج الأبعاد الأربعة والمهام التواصلية والتقويم الأدائي داخل منهج معهد إسلامي، لا ابتكارا نظريا مستقلا. وتفسر النتائج في حدود العينة الصغيرة غير العشوائية والمؤسسة الواحدة وعدم توثيق ثبات الاتفاق بين المقومين.

الكلمات المفتاحية: تطوير المنهج؛ مهارة الكلام؛ المدخل التواصلية؛ التعلم القائم على المهام؛ نموذج ثياغاراجان

INTRODUCTION

Speaking is the most direct productive dimension of language competence. In speaking, learners turn their knowledge of sounds, vocabulary, and structures into oral performance that serves a communicative purpose. Knowing a rule or recognizing a word in a written context is not enough. Learners must retrieve language under time pressure, organize it into comprehensible discourse, consider the listener and the situation, and manage interaction at their proficiency level. For this reason, speaking development is a complex process that combines accuracy, fluency, pronunciation, linguistic resources, and interaction management (Bygate, 1987; Goh & Burns, 2012; Luoma, 2004).

In teaching Arabic to non-native speakers, these components become more complex. Learners face phonological, morphological, and syntactic differences between their first language and Arabic. They also have limited natural exposure outside the classroom, fear making mistakes, and receive little speaking time per learner. The literature identifies the language environment, repeated practice, and structured feedback as decisive factors in turning linguistic knowledge into actual oral use. This holds especially in boarding school settings, which can make Arabic part of daily academic life (Amrullah et al., 2023; Atabik et al., 2023; Nasir et al., 2024; Teimouri et al., 2019).

A comprehensive textbook or an official curriculum does not guarantee that the curriculum takes the same form in the classroom. The intended curriculum may cover all four skills, yet lesson time, teacher priorities, class size, assessment tools, and student levels reshape it during implementation. Researchers should therefore analyze a curriculum as objectives, content, activities, assessment, and an implementation environment, not as a list of topics. Language curriculum development scholars stress the need to start with needs and context analysis and then align outcomes, content, procedures, and assessment (Asy'ari et al., 2026; Graves, 2000; Nation & Macalister, 2010; Richards, 2001).

Al-Arabiyyah Bayna Yadayk is a widely used textbook in Arabic language institutions. It offers gradual progression, varied topics, and integrated skills. However, preliminary observation at Al-Wadi Al-Mubarak Islamic Boarding School and interviews with program staff showed that classroom practice leaned toward reading, writing, and theoretical content, while structured speaking opportunities remained limited. Teachers and students reported several difficulties: little practice time, mixed proficiency levels, large classes, weak fluency and confidence, and fear of making mistakes. These observations do not point to a flaw in the textbook itself. They point to a need for implementation-level adaptation that makes speaking a central goal with dedicated time, activities, and assessment.

Communicative language teaching (CLT) provides a basis for this adaptation. It shifts the focus of learning from presenting structures to using them to convey meaning,

and it builds interaction, information gaps, linguistic choice, and feedback into the lesson (Brown & Lee, 2015; Richards, 2006). Task-based language teaching (TBLT) complements this orientation. It organizes learning around tasks with a clear goal and an observable outcome, such as getting acquainted, requesting information, describing a person, or playing a role. Vocabulary and structures then serve as resources for completing the task rather than as isolated goals (Bryfonski & McKay, 2019; Ellis, 2003; Nunan, 2004; Willis & Willis, 2007).

Pair and group activities increase learner speaking time and distribute roles. Role play offers a safe context for linguistic experimentation and repeated performance. Formative assessment, rating rubrics, self-assessment, and portfolios help teachers monitor performance over time instead of relying on a single final knowledge test. Feedback works best when it clarifies the goal, the current level of performance, and the next step, and when learners take part in reviewing their own performance (Panadero & Lipnevich, 2022; Wisniewski et al., 2020; Yan et al., 2022).

The literature underpinning this study falls into three strands. The first strand addresses the design of Arabic performance curricula based on standard frameworks such as ACTFL and Brown's model. It offers a basis for defining competencies and performance indicators, but it does not necessarily combine the development of a multi-unit product with testing in a pretest-controlled design (Borham et al., 2024; Hasanah et al., 2021). The second strand examines the language environment, intensive programs, and skill development in Islamic boarding schools. It shows the importance of practice and institutional support, yet it mostly describes programs or implementation mechanisms. It rarely documents an integrated sequence that starts with needs analysis and ends with a student book, a teacher's guide, and performance assessment (Atabik et al., 2023; Aziz et al., 2025; Roziqi et al., 2025; Siregar et al., 2023). The third strand, the CLT and TBLT literature, provides well-established design principles. However, it does not specify how to adapt the units of *Al-Arabiyyah Bayna Yadayk* to the boarding school under study (Ellis, 2017; Richards, 2006; Willis & Willis, 2007). This study does not claim new theoretical components. It operationalizes needs analysis, the 4-D model, CLT, TBLT, and performance assessment within a multi-unit product and then evaluates that product through expert judgment and a quasi-experimental design with pretest control.

The study adopted Thiagarajan's 4-D model because it offers a procedural path that moves from define to design, develop, and disseminate (Thiagarajan et al., 1974). The study carried out the first three stages in the field and treated wide dissemination as a later stage that requires validation in other institutions and samples. Evaluation did not form a separate final stage. It accompanied analysis, design, expert review, implementation, and revision. This approach makes the model a tool for building a product that others can examine and improve, not a formal template.

Accordingly, the study addressed four questions. What are the characteristics of the current curriculum and the speaking development needs at the boarding school? How can the developed speaking units be built on needs analysis, CLT, and TBLT? How do teachers implement these units in the classroom? How much does speaking performance change in the experimental group compared with the control group? The study aims to offer a usable instructional product and a methodological and contextual contribution. It shows how to operationalize existing curricula within a specific boarding school setting. It also separates the significance of results within the sample from any claim of theoretical innovation or broader causal generalization.

Theoretical Basis for Speaking Curriculum Development

Speaking curriculum development starts from the concept of communicative competence. This concept refers to the ability to choose language that fits the purpose, the listener, and the context, in addition to structural accuracy. Success, therefore, does not mean an error-free sentence. A sentence may be correct yet fail to produce the required response or to sustain the interaction. A speaking curriculum must combine accuracy, fluency, appropriateness, and turn management, with a progression that suits beginners. At the A1 level, for example, realistic goals include giving simple personal information, understanding familiar questions, and managing short exchanges, not delivering complex speeches (Council of Europe, 2020).

The definition of speaking directly shapes the choice of assessment instrument. If speaking means memorizing vocabulary or repeating a text, discrete questions may suffice. If speaking means communicative performance, learners must face a task that requires production and response. For this reason, the study's rubric divided performance into pronunciation, linguistic resources, interaction, and expression. This structure shows where learners improve and where they remain weak, and it reduces the influence of the rater's general impression. Speaking assessment scholars stress clear tasks, explicit rating criteria, and rater consistency, because an oral score emerges from the interaction of learner, task, and rater (Fulcher, 2015; Luoma, 2004).

Fluency requires a balance between planning, time pressure, and repetition. A beginner may know a phrase but retrieve it slowly. The more often learners use a phrase for similar purposes, the more readily available it becomes. Mechanical repetition alone, however, is not enough. It may produce memorized performance that does not transfer to new situations. The units therefore moved from repetition to information exchange and then to tasks, and they let learners repeat activities with a different partner or in a different context. This sequence combines the security of guided practice with the flexibility that real speaking demands (Abdi Tabari et al., 2025; Bygate, 2010; Suzuki, 2021).

Peer interaction is a learning resource, not merely a way to arrange seating. In pair dialogue, learners must listen, respond, ask for clarification, and repair misunderstandings. These processes do not occur in monologue. A peer can offer a model close to the learner's own level and can provide more speaking time than a pattern in which the teacher questions one student at a time. Effective interaction, however, requires assigned roles, a defined outcome, and monitoring of language use. Without these, group work becomes unbalanced or shifts to the first language (Sato & Ballinger, 2016).

Content selection depends on needs analysis, not on abstract linguistic progression alone. Greetings and introductions, for example, matter because students need them in the boarding school. Teachers can embed pronouns, question words, nationality, and country names in this topic. Competence develops when vocabulary and structures serve this communicative event and when learners use them repeatedly to introduce themselves, introduce a classmate, and ask for new information. Content organization thus combines situation, function, and structure instead of presenting each in a separate track.

The teacher's guide plays a key role in implementation fidelity. A student unit may be well designed, yet a teacher may turn the task into explanation or cut interaction time if the goal, timing, outcome, and correction method remain unclear. The guide should therefore go beyond answer keys. It should explain the pedagogical rationale, the sequence of support, how to form pairs and groups, and when to delay correction or

intervene. These details make the product transferable across teachers and reduce its dependence on individual effort.

Formative assessment links the written curriculum with the implemented curriculum. When a teacher notices that most students understand the vocabulary but do not initiate questions, the teacher can adjust the next activity to create more chances to ask questions. When self-recordings reveal recurring pronunciation problems, the teacher can add a short drill before students repeat the task. Assessment thus does more than produce a score at the end of a unit. It generates information that redirects teaching, which is the essence of assessment for learning (Black & Wiliam, 2009; Lipnevich & Panadero, 2021).

RESEARCH METHOD

The study used a research and development (R&D) approach to build instructional units for speaking development. It combined this approach with a two-group quasi-experimental design with a pretest and a posttest. The development path followed Thiagarajan's 4-D model through the define, design, and develop stages. The study did not carry out the wide dissemination stage, because its aim was to produce a context-appropriate model and to test its effectiveness at a preliminary level. The researchers used evaluation continuously at every stage to review decisions, materials, and instruments.

In the define stage, the researchers analyzed students' characteristics, levels, and needs. They examined the textbook, the schedule, and teaching and assessment practices, and they gathered the views of administrators and teachers through interviews and observation. In the design stage, they set speaking outcomes, organized vocabulary and structures, built the activity sequence, and drafted the student book, the teacher's guide, the test, and the rating rubric. In the develop stage, experts in Arabic language teaching, curriculum, and measurement reviewed the product and instruments, and the researchers made the necessary revisions before the field trial.

The study took place at Al-Wadi Al-Mubarak Islamic Boarding School in Bogor with first-year junior secondary students who are non-native speakers of Arabic. The available sample consisted of 32 students in two equal groups: 16 in the experimental group and 16 in the control group. The available study records do not document individual random assignment. The study therefore treats the two groups as non-randomized groups within a quasi-experimental design and does not rule out selection bias, despite the similar pretest means. No documented a priori power analysis determined the sample size, so the small sample limits the precision of estimates and generalization. The control group followed the existing method, while the experimental group used all the developed speaking units. Implementation followed the boarding school's approved schedule and allocated three lessons to each unit, which allowed students to move from guided input to independent performance.

The product consisted of a student book and a teacher's guide covering the speaking units in Volume 1 of *Al-Arabiyyah Bayna Yadayk*. Each unit included the level and time allocation, learning outcomes, vocabulary and structures, media, teaching steps, a communicative task, and assessment and homework. The sequence began with a warm-up and activation of prior knowledge. It then moved to vocabulary presentation, listening, and repetition, followed by dialogue, pair or group work, and role play. It ended with the "I Speak Fluently" activity and an assessable task.

The researchers collected qualitative data through interviews, observation,

curriculum content analysis, and expert feedback. They used these data to diagnose the problem, guide the design, and interpret implementation. They collected quantitative data through an oral performance pretest and posttest and an analytic rating rubric. The performance criteria covered four main domains: phonological and pronunciation skills, linguistic and structural skills, communicative and interactive skills, and oral expression skills. The study thus measured speaking as a composite performance, not as knowledge of vocabulary or grammar alone. However, the available research records do not specify whether one or more independent raters scored the responses, and they report no inter-rater reliability coefficient such as the ICC. The study therefore claims no inter-rater reliability. The sensitivity of scores to rater judgment is a measurement limitation that future studies should address.

The researchers analyzed the data with means, standard deviations, paired-samples t-tests within groups, independent-samples t-tests between groups, and analysis of covariance (ANCOVA) to control for the pretest. Tests of the homogeneity of regression slopes and homogeneity of variance preceded the ANCOVA. The researchers also calculated partial eta squared to estimate effect size in the analysis model and computed the 95% confidence interval for the adjusted difference from the error variance and degrees of freedom of the ANCOVA model. The significance level was 0.05. The researchers interpreted the results within the limits of the non-randomized quasi-experimental design, the sample, and the context.

Table 1. Components of the Developed Speaking Units and Their Instructional Sequence

Component	Procedural Content	Function in Speaking Development
Learning outcomes	Formulating observable and measurable oral behavior	Directing content, activities, and assessment toward a specific performance
Warm-up and input	Picture or situation, vocabulary, listening and repetition	Activating knowledge and building phonological and lexical resources
Guided practice	Dialogue, questions and answers, information exchange	Moving learners from imitation to controlled use
Communicative task	Pair or group work, role play, problem situation	Creating a communicative purpose and a real need to speak
Independent production	Short talk, presentation, or description without reading	Developing fluency, organization, and confidence
Assessment	Rating rubric, formative and self-assessment, portfolio	Monitoring performance and giving actionable feedback

RESULTS AND DISCUSSION

Curriculum Diagnosis and the Need for Development

The curriculum analysis showed that the textbook follows a graded and integrated overall structure and covers all four skills. This makes it a suitable basis for teaching Arabic to non-native speakers. However, actual implementation did not distribute time and activities evenly. Reading, writing, and theoretical explanation dominated, while speaking opportunities remained limited and irregular. Observation, interviews, and materials analysis all pointed to the same diagnosis. This finding directed development toward reorganizing implementation rather than abandoning or replacing the textbook.

The main needs were more speaking opportunities for each student, attention to mixed proficiency levels, a progression from pronunciation and short sentences to dialogue and independent expression, less fear of making mistakes, and assessment linked to oral behavior. Studies in Islamic boarding schools confirm the importance of the language environment, repeated practice, and institutional support. They also show that the presence of Arabic in a program does not ensure that all students produce it (Aziz et al., 2025; Fakhroh & Hikmah, 2023; Ilyas et al., 2024).

This diagnosis leads to an important distinction between the curriculum as material and the curriculum as an implemented experience. Material may contain rich dialogues and vocabulary, yet its effect on speaking stays limited if explanation takes up most of the lesson, if performance stays limited to choral repetition, or if written questions measure success. The product therefore redistributed control over time, talk, and assessment in the classroom. It gave students more time to perform, assigned clearer interactive roles, and used assessment that describes actual performance.

Integrating Needs Analysis, Design, and Assessment

The development stages show that needs analysis was not an opening step that ended with a report. It served as a reference for later decisions. The problem of limited practice became a decision to increase pair and group work. Mixed proficiency levels led to graded levels of support. Fear of making mistakes led to activities that begin with safe imitation before independent performance. Limited oral assessment led to a rating rubric and a portfolio. As a result, each diagnosed problem can be traced to a component of the product.

This traceability is an important criterion of development quality, because it prevents the addition of attractive activities that serve no specific goal. The units examine every activity through three questions. What performance does it target? What linguistic resource does it require? What evidence will show that learning has occurred? If the answers are unclear, the activity is redesigned or removed. This principle reflects the idea of alignment among outcomes, learning, and assessment in curriculum design (Graves, 2000; Tomlinson, 2013).

Expert review also showed that language accuracy alone does not determine product validity. The review covered the clarity of objectives, the fit of content to level, the sequence of steps, the authenticity of tasks, the feasibility of assessment, and the clarity of teacher instructions. Expert feedback led to revisions before the trial, which reduced the chance that implementation problems would stem from unclear materials. Expert review does not replace field testing, but it provides a logical and professional check before learners invest time in the trial.

The development cycle also helped distinguish between modification and development. Modification may involve only changing an example or adding an

exercise. Development reconsiders the relationships among the components of a unit. The study kept the textbook's topics and overall sequence, but it rewrote the outcomes, added tasks, changed the distribution of speaking time, and introduced performance criteria. The result was therefore a deep adaptation within the existing framework, not a separate book unrelated to the institutional curriculum.

Product Design and Pedagogical Features

Development produced a student book and a teacher's guide covering all target speaking units in Volume 1, not a single pilot unit. Each unit specified performance-based objectives, situation-linked vocabulary and structures, a sequence from warm-up to task, and feedback criteria. Experts reviewed the alignment of objectives, content, activities, skills, and assessment, and the researchers revised the materials before implementation.

The design linked each linguistic resource to a communicative function. Students did not merely memorize greeting and introduction phrases. They used them in a task in which they got to know a classmate and then introduced that classmate to the class. Students did not practice question words in isolated sentences. They used them to obtain missing information. This structure lets students notice a form, practice it, and then use it to reach a communicative outcome, in line with the principles of CLT and TBLT (Ellis, 2017; Long, 2015).

Pair and group work extended speaking time compared with teacher-centered question-and-answer patterns. Role play created a safe psychological distance that allowed experimentation and repeated performance, while the fluency activity let students move from a read text to organized speech. These features match the view that fluency development requires meaningful repetition, familiar content, appropriate communicative pressure, and chances to improve performance in later attempts (Nation & Newton, 2020; Suzuki & Hanzawa, 2022).

Assessment did not serve as a separate conclusion. Teachers used observation and the rating rubric during performance and added self-assessment and portfolios to track recordings, dialogues, and presentations. Clear criteria turned feedback from a general judgment into information about pronunciation, structures, interaction, and the organization of expression. This type of feedback supports learning when learners know where they stand and what they need to adjust (Carless & Boud, 2018; Hattie & Timperley, 2007; Hoo et al., 2022; Molloy et al., 2020; Nicol & Macfarlane-Dick, 2006).

Pedagogical Interpretation of the Quantitative Differences

The intensity of retrieval and production opportunities explains the large gain in the experimental group's mean. In each unit, students encountered vocabulary and structures, retrieved them in dialogue, reorganized them in a situation or role, and then produced an independent outcome. These successive cycles narrow the gap between knowledge stored in memory and knowledge available for rapid use. Repeating performance with different partners also prevents a phrase from becoming tied to a single text and makes it more flexible (Abdi Tabari et al., 2025; Suzuki, 2021).

The distribution of participation may have an effect parallel to that of content. In a teacher-centered class, total speaking time may be large while each student gets only a few seconds. Simultaneous pairs let more students speak at the same time. When roles come with a clear outcome, each student must ask, answer, or give information (Nation & Newton, 2020; Sato & Ballinger, 2016). The measure of curriculum renewal, therefore, is not the number of new pages but the amount of purposeful language activity it offers

each learner.

Brief feedback may also have contributed to the improvement. It focused on errors that impede meaning or recur, and teachers delayed some correction until after the task to keep speech flowing. This differs from correcting every error immediately, which may interrupt interaction and increase hesitation (Li, 2010). The rating rubric lets teachers group observations under clear domains and then set a small goal for the next attempt, such as improving a question word or making an answer more coherent (Luoma, 2004; Wisniewski et al., 2020).

Ongoing regular instruction, exposure to textbook content, and familiarity with the test format may explain the modest improvement in the control group. This point matters, because gains in both groups prevent automatic attribution of every change to the intervention. However, the large difference in the size of the gains and the persistence of the difference after controlling for the pretest support a link between the outcome and the combined intervention. Even so, possible teacher effects, participant motivation, and within-group interaction remain plausible in a quasi-experimental design.

The significance of the pretest in the ANCOVA shows that prior performance remained an important factor in posttest results. Pedagogically, this observation matters more than treating the pretest as a mere statistical variable. It means that students enter the intervention with different resources and that the design should offer additional support paths for those who start at a lower level. Future implementations can use diagnostic results to form flexible pairs or to give some students extra audio input and exercises.

Readers should not interpret the effect size apart from the characteristics of the measure. The total test score ranges up to twenty points, the sample is relatively homogeneous, and the intervention targets the performance that the test measures. These factors may inflate the value (Plonsky & Oswald, 2014). The more accurate scientific report therefore presents means, standard deviations, test statistics, and effect size together and then describes the limitations, instead of using the label "high effectiveness" without context. The study followed this approach by separating statistical significance from pedagogical generalization.

The results nonetheless offer the institution important practical evidence. They show that reorganizing existing units can produce a measurable difference without replacing the textbook entirely. This makes the development easier to adopt, because teachers and administrators keep working with familiar content while gaining clearer procedural guidance. Sustainability improves when the innovation connects to the schedule, available resources, and training rather than to an add-on project that the program struggles to absorb (Richards, 2001; Tomlinson, 2013).

Unit Implementation and the Shift in Classroom Roles

Implementation showed that the graded organization helped students move from constrained responses to more independent production. The teacher began with a phonological and lexical model and then gradually reduced support. Students moved from repetition to completing dialogues, exchanging information, performing roles, and speaking without reading. The teacher's role did not disappear. It shifted from sole source of language to time manager, facilitator of interaction, and provider of brief feedback (van de Pol et al., 2010).

Gradual progression proved important because pushing learners directly into free speech may raise anxiety and lead them to rely on their first language (Teimouri et al., 2019). The move from listening input to guided practice and then to tasks provided a

linguistic base before teachers asked for production. Repeating structures in different situations helped students consolidate forms without separating them from meaning. Short tasks also made participation possible for students at different levels, because teachers could adjust the amount of information and the length of each turn to learner ability.

Observation also revealed that unit success depended not on content alone but on classroom management, time allocation, and rules of interaction. Group work without clear roles may let one student dominate, and role play without a goal may become superficial acting (Sato & Ballinger, 2016). The teacher's guide therefore included step-by-step instructions, timing for each stage, the expected type of outcome, and the method for giving feedback. These procedural details turn the communicative principle into a repeatable practice.

Pretest and Posttest Results

Before the intervention, the control group had a mean of 9.38 (SD = 1.740), and the experimental group had a mean of 9.75 (SD = 1.571). The difference was not statistically significant ($t = 0.639$, $p = 0.528$), which indicates similar initial performance. After the intervention, the control group mean rose to 10.63, while the experimental group mean rose to 17.19. The experimental group gained 7.44 points, compared with 1.25 points in the control group.

Table 2. Pretest and Posttest Means for Both Groups

Group	n	Pretest Mean	Pretest SD	Posttest Mean	Posttest SD	Gain
Control	16	9.38	1.740	10.63	1.668	1.25
Experimental	16	9.75	1.571	17.19	1.328	7.44

A paired-samples t-test showed a significant difference within the experimental group between the two measurements ($t = 36.551$, $p < 0.001$). The posttest comparison between groups also showed a difference in favor of the experimental group ($t = 12.31$, $p < 0.001$). This result suggests that the improvement did not stem from regular exposure or test familiarity alone, because the control group improved only slightly, while the experimental group changed far more.

ANCOVA Results and Effect Size

The data met the assumptions of homogeneity of regression slopes and homogeneity of variance, with significance levels above 0.05 in both tests. After controlling for the pretest, the adjusted mean was 10.770 for the control group and 17.042 for the experimental group, an adjusted difference of 6.272 points in favor of the experimental group. The 95% confidence interval for this difference ranged from 5.688 to 6.856, which shows the uncertainty around the estimate within the sample. The ANCOVA showed a significant group effect on posttest performance ($F(1, 29) = 481.853$, $p < 0.001$), and the pretest was a significant predictor of posttest performance ($F(1, 29) = 76.808$, $p < 0.001$).

Table 3. Summary of ANCOVA for Posttest Performance

Source of Variance	Sum of Squares	df	Mean Square	F	p	Partial η^2
Pretest	49.498	1	49.498	76.808	< .001	0.726
Group	310.529	1	310.529	481.853	< .001	0.943

Source of Variance	Sum of Squares	df	Mean Square	F	p	Partial η^2
Error	18.689	29	0.644	-	-	-

Partial eta squared for the group effect was 0.943, a very large effect size in the analysis model (Lakens, 2013; Plonsky & Oswald, 2014). However, this value does not mean that the curriculum alone explains 94.3 percent of speaking learning in all conditions. It is an estimate tied to the small sample, the measurement instrument, the duration of the intervention, and the variance within this context. It therefore indicates the strength of the difference within the model, not an absolute causal proportion that transfers to every institution.

The result is consistent with the pedagogical logic of the intervention. The units increased the number of listening and response cycles, distributed speaking opportunities, linked structures to tasks, and provided feedback on actual performance. The difference reflects an integrated system of decisions rather than the effect of a single element such as role play or group work. This explains why objectives, activities, assessment, and the implementation guide need to be developed together (Bryfonski & McKay, 2019).

Several factors limit this interpretation: the lack of documented individual random assignment, possible selection bias, the small sample, its link to a single institution, and the absence of an a priori power analysis or a delayed posttest to check whether the effect persists. The lack of documented information on the number of raters and their inter-rater agreement also weakens the measurement inference (Koo & Li, 2016). In addition, implementation in a boarding school may benefit from an environment that differs from a day school. Replication therefore requires larger samples, stronger assignment or control, independent raters with ICC estimates, multiple teachers, delayed measurements, and monitoring of implementation fidelity.

Theoretical and Practical Contributions

The study does not contribute a new theoretical model, nor does it claim that the 4-D model, CLT, or TBLT are new components. It makes a methodological and contextual contribution by integrating these components into a traceable procedural sequence. The sequence includes analyzing the needs of learners at an Islamic boarding school, adapting all target speaking units in Volume 1 of *Al-Arabiyyah Bayna Yadayk*, building a student book and a teacher's guide, linking tasks to performance criteria, and evaluating the product through expert review and a pretest-controlled quasi-experimental trial. This sequence shows how known principles can operate within a coherent product that fits a specific institutional context, without turning the adaptation into a claim of general theoretical novelty.

In practice, institutions that use a ready-made textbook can improve how they implement it without abandoning its sequence or identity. Development begins by defining the target oral behavior, allocating time to it, building a task that makes it necessary, and setting a criterion for observing it. Teachers also need training in managing pairs, groups, and feedback, and in creating an environment that allows repeated attempts without turning correction into a source of anxiety (Ellis, 2017; Teimouri et al., 2019).

Curriculum development does not simply mean adding more activities. An overloaded unit may reduce the depth of practice. What matters is functional alignment

among the objective, input, task, and assessment. If the objective is managing a dialogue, the unit must create a need to exchange information, assign roles, and require turn-taking, and the assessment must measure responses and follow-up, not just the accuracy of isolated sentences (Fulcher, 2015; Nation & Macalister, 2010).

CONCLUSION

The study answered its questions by diagnosing the imbalance between the presence of skills in the materials and the speaking opportunities in implementation. It then built graded speaking units for Volume 1 of *Al-Arabiyyah Bayna Yadayk*, implemented them in the classroom, and compared performance changes between two groups. The development stage produced a student book and a teacher's guide that link needs analysis to outcomes, vocabulary, structures, communicative tasks, and formative and performance assessment. Implementation showed that the units can increase student participation and move students from guided repetition to dialogue, tasks, and more independent production.

Quantitatively, the experimental group mean rose from 9.75 to 17.19, while the control group mean rose from 9.38 to 10.63. The adjusted difference after controlling for the pretest was 6.272 points, with a 95% confidence interval from 5.688 to 6.856. This finding offers strong contextual evidence within the sample, not a final causal estimate open to unconditional generalization. Theoretically, the study proposes no new theory. It shows how curriculum development can connect with speaking instruction and performance assessment. Methodologically, it offers a traceable path that integrates needs analysis, the 4-D stages, communicative tasks, expert review, and quasi-experimental testing. Practically, it provides units and a guide that similar boarding schools can adapt to their learners, schedules, and resources. The small sample, the lack of documented random assignment and inter-rater agreement, the single institution, and the short measurement period limit the results. Future studies should therefore use larger samples, documented assignment procedures, independent raters, and delayed, multi-institution measurements.

REFERENCES

- Abdi Tabari, M., Zhuang, J., & Farahanynia, M. (2025). Task repetition and L2 oral performance: A meta-analysis. *System*, 135, 103868. <https://doi.org/10.1016/j.system.2025.103868>
- Amrullah, A. F., Bakheit, B. M., & Huda, M. (2023). Development strategy of speaking skill for the Tsanawiyah level. *Ijaz Arabi Journal of Arabic Learning*, 6(2), 585–595. <https://doi.org/10.18860/ijazarabi.v6i2.23220>
- Asy'ari, H., Akhbar, K., Rufaiqoh, E., Ibrahim, N. I. E., Sulfiatin, S., & Muid, F. A. (2026). Arabic curriculum design for non-native speakers in multicultural settings: A systematic literature review. *Arabiyat: Jurnal Pendidikan Bahasa Arab dan Kebahasaaraban*, 13(1), 70–85. <https://doi.org/10.15408/a.v13i1.50924>
- Atabik, A., Yahya, M. S., & Mustajab. (2023). Life skills approach in Arabic language learning at Islamic boarding school-based Madrasah Aliyah. *Al-Hayat: Journal of Islamic Education*, 7(2), 327–342. <https://doi.org/10.35723/ajie.v7i2.378>
- Aziz, M. F., Mulyati, A., Ramdhani, Z. R., & Syahputra, I. (2025). The implementation of an intensive Arabic language instruction program: A case study at Al-Aqsa Islamic Boarding School. *Tadris Al-'Arabiyyah*, 4(1), 97–113. <https://doi.org/10.15575/ta.v4i1.44921>
- Black, P., & Wiliam, D. (2009). Developing the theory of formative assessment. *Educational Assessment, Evaluation and Accountability*, 21(1), 5–31. <https://doi.org/10.1007/s11092-008-9068-5>
- Borham, S. R., Ramli, S., & Abdul Ghani, M. T. (2024). AI concepts integration in developing e-muhadathat kits for non-Arabic speakers. *Ijaz Arabi Journal of Arabic Learning*, 7(3). <https://doi.org/10.18860/ijazarabi.v7i3.26568>
- Brown, H. D., & Lee, H. (2015). *Teaching by principles: An interactive approach to language pedagogy* (4th ed.). Pearson.
- Bryfonski, L., & McKay, T. H. (2019). TBLT implementation and evaluation: A meta-analysis. *Language Teaching Research*, 23(5), 603–632. <https://doi.org/10.1177/1362168817744389>
- Bygate, M. (1987). *Speaking*. Oxford University Press.
- Bygate, M. (2010). Speaking. In R. B. Kaplan (Ed.), *The Oxford handbook of applied linguistics* (2nd ed., pp. 412–423). Oxford University Press.
- Carless, D., & Boud, D. (2018). The development of student feedback literacy: Enabling uptake of feedback. *Assessment & Evaluation in Higher Education*, 43(8), 1315–1325. <https://doi.org/10.1080/02602938.2018.1463354>
- Council of Europe. (2020). *Common European Framework of Reference for Languages: Learning, teaching, assessment – Companion volume*. Council of Europe Publishing.
- Ellis, R. (2003). *Task-based language learning and teaching*. Oxford University Press.

- Ellis, R. (2017). Position paper: Moving task-based language teaching forward. *Language Teaching*, 50(4), 507–526. <https://doi.org/10.1017/S0261444817000179>
- Fakhiroh, N., & Hikmah, K. (2023). Enhancing Arabic language proficiency in Islamic boarding schools: A comprehensive study. *Indonesian Journal of Islamic Studies*, 11(3). <https://doi.org/10.21070/ijis.v11i3.1673>
- Fulcher, G. (2015). *Re-examining language testing: A philosophical and social inquiry*. Routledge.
- Goh, C. C. M., & Burns, A. (2012). *Teaching speaking: A holistic approach*. Cambridge University Press.
- Graves, K. (2000). *Designing language courses: A guide for teachers*. Heinle & Heinle.
- Hasanah, M., Mubaligh, A., Sari, R. R., Syarofah, A., & Prasetyo, A. (2021). Arabic performance curriculum development: Reconstruction based on ACTFL and Douglas Brown perspective. *Ijaz Arabi Journal of Arabic Learning*, 4(3), 779–801. <https://doi.org/10.18860/ijazarabi.v4i3.11900>
- Hattie, J., & Timperley, H. (2007). The power of feedback. *Review of Educational Research*, 77(1), 81–112. <https://doi.org/10.3102/003465430298487>
- Hoo, H.-T., Deneen, C., & Boud, D. (2022). Developing student feedback literacy through self and peer assessment interventions. *Assessment & Evaluation in Higher Education*, 47(3), 444–457. <https://doi.org/10.1080/02602938.2021.1925871>
- Ilyas, M., Rasmi, M., & Rusydi, M. (2024). Improving modern pondok students' Arabic language skills in Indonesia: Language institutions as language improvement central. *AL-ISHLAH: Jurnal Pendidikan*, 16(2). <https://doi.org/10.35445/alishlah.v16i2.5095>
- Koo, T. K., & Li, M. Y. (2016). A guideline of selecting and reporting intraclass correlation coefficients for reliability research. *Journal of Chiropractic Medicine*, 15(2), 155–163. <https://doi.org/10.1016/j.jcm.2016.02.012>
- Lakens, D. (2013). Calculating and reporting effect sizes to facilitate cumulative science: A practical primer for t-tests and ANOVAs. *Frontiers in Psychology*, 4, 863. <https://doi.org/10.3389/fpsyg.2013.00863>
- Li, S. (2010). The effectiveness of corrective feedback in SLA: A meta-analysis. *Language Learning*, 60(2), 309–365. <https://doi.org/10.1111/j.1467-9922.2010.00561.x>
- Lipnevich, A. A., & Panadero, E. (2021). A review of feedback models and theories: Descriptions, definitions, and conclusions. *Frontiers in Education*, 6, 720195. <https://doi.org/10.3389/feduc.2021.720195>
- Long, M. H. (2015). *Second language acquisition and task-based language teaching*. Wiley Blackwell.
- Luoma, S. (2004). *Assessing speaking*. Cambridge University Press.
- Molloy, E., Boud, D., & Henderson, M. (2020). Developing a learning-centred framework

- for feedback literacy. *Assessment & Evaluation in Higher Education*, 45(4), 527–540. <https://doi.org/10.1080/02602938.2019.1667955>
- Nasir, A., Juhri, J., & Karim, A. (2024). Multilingual education in Islamic boarding schools: Planning, implementation and institutional policy. *Ijaz Arabi Journal of Arabic Learning*, 7(2). <https://doi.org/10.18860/ijazarabi.v7i2.26151>
- Nation, I. S. P., & Macalister, J. (2010). *Language curriculum design*. Routledge.
- Nation, I. S. P., & Newton, J. (2020). *Teaching ESL/EFL listening and speaking* (2nd ed.). Routledge.
- Nicol, D. J., & Macfarlane-Dick, D. (2006). Formative assessment and self-regulated learning: A model and seven principles of good feedback practice. *Studies in Higher Education*, 31(2), 199–218. <https://doi.org/10.1080/03075070600572090>
- Nunan, D. (2004). *Task-based language teaching*. Cambridge University Press.
- Panadero, E., & Lipnevich, A. A. (2022). A review of feedback models and typologies: Towards an integrative model of feedback elements. *Educational Research Review*, 35, 100416. <https://doi.org/10.1016/j.edurev.2021.100416>
- Plonsky, L., & Oswald, F. L. (2014). How big is "big"? Interpreting effect sizes in L2 research. *Language Learning*, 64(4), 878–912. <https://doi.org/10.1111/lang.12079>
- Richards, J. C. (2001). *Curriculum development in language teaching*. Cambridge University Press.
- Richards, J. C. (2006). *Communicative language teaching today*. Cambridge University Press.
- Roziqi, M. A., Murdiono, M., Rahman, N. F., & Arif, M. (2025). Mapping of strategies and methods for learning Arabic speaking skills. *Ijaz Arabi Journal of Arabic Learning*, 8(1). <https://doi.org/10.18860/ijazarabi.v8i1.26939>
- Sato, M., & Ballinger, S. (Eds.). (2016). *Peer interaction and second language learning: Pedagogical potential and research agenda*. John Benjamins.
- Siregar, Z. U., Fauzia, E. L., & Fathurrahman, Y. (2023). Development of Arabic language skills for students at Syamsul Ulum Muhammadiyah Bandung Islamic Boarding School. *Ta'lim al-'Arabiyyah*, 7(1). <https://doi.org/10.15575/jpba.v7i1.24453>
- Suzuki, Y. (2021). Optimizing fluency training for speaking skills transfer: Comparing the effects of blocked and interleaved task repetition. *Language Learning*, 71(2), 285–325. <https://doi.org/10.1111/lang.12433>
- Suzuki, Y., & Hanzawa, K. (2022). Massed task repetition is a double-edged sword for fluency development. *Studies in Second Language Acquisition*, 44(2), 536–561. <https://doi.org/10.1017/S0272263121000358>
- Teimouri, Y., Goetze, J., & Plonsky, L. (2019). Second language anxiety and achievement: A meta-analysis. *Studies in Second Language Acquisition*, 41(2), 363–387. <https://doi.org/10.1017/S0272263118000311>

- Thiagarajan, S., Semmel, D. S., & Semmel, M. I. (1974). *Instructional development for training teachers of exceptional children: A sourcebook*. Indiana University.
- Tomlinson, B. (Ed.). (2013). *Developing materials for language teaching* (2nd ed.). Bloomsbury.
- van de Pol, J., Volman, M., & Beishuizen, J. (2010). Scaffolding in teacher–student interaction: A decade of research. *Educational Psychology Review*, 22(3), 271–296. <https://doi.org/10.1007/s10648-010-9127-6>
- Willis, D., & Willis, J. (2007). *Doing task-based teaching*. Oxford University Press.
- Wisniewski, B., Zierer, K., & Hattie, J. (2020). The power of feedback revisited: A meta-analysis of educational feedback research. *Frontiers in Psychology*, 10, 3087. <https://doi.org/10.3389/fpsyg.2019.03087>
- Yan, Z., Lao, H., Panadero, E., Fernández-Castilla, B., Yang, L., & Yang, M. (2022). Effects of self-assessment and peer-assessment interventions on academic performance: A meta-analysis. *Educational Research Review*, 37, 100484. <https://doi.org/10.1016/j.edurev.2022.100484>