



Development of an Edpuzzle-Based *Maharah Istima'* Test for Junior High School Students

تطوير اختبار مهارة الاستماع القائم على منصة إدبازل لطلاب المرحلة المتوسطة

Iin Stevani Katili^{1*}, Ibnu Rawandhy N. Hula², Ratni Bt. Hj. Bahri³, Hairuddin⁴, Zohra Yasin⁵

^{1,2,3,4,5}Universitas Islam Negeri Sultan Amai Gorontalo, Indonesia

*Corresponding E-mail: iinstevani.k@gmail.com

ABSTRACT

The assessment of *maharah istima'* remains challenging because teachers often rely on conventional evaluation techniques and have limited access to structured digital assessment instruments. This study aimed to develop an Edpuzzle-based *Maharah Istima'* Test for eighth-grade students at SMP-IT Al-Faruq, Gorontalo City. The study employed a Research and Development approach using the ADDIE model, which comprises analysis, design, development, implementation, and evaluation. The developed product integrated animated conversation videos as listening materials, interactive multiple-choice items, test instructions, answer keys, and online assessment through Edpuzzle. The product was evaluated through expert validation, practicality assessment, and psychometric item analysis. Expert validation yielded an average feasibility score of 96.7%, indicating that the product was highly feasible. Item analysis showed that 34 of the 36 items were valid, while the reliability coefficients of the four test packages ranged from 0.744 to 0.829. The difficulty analysis identified 26 easy items and 8 moderately difficult items. The discrimination analysis indicated that most items had adequate discriminating power, while 40 of the 56 distractors functioned effectively. Teacher and student responses produced an overall practicality score of 91.4%, which was categorized as highly practical. These findings indicate that the developed *Maharah Istima'* Test is highly feasible, practical, and supported by satisfactory empirical item quality for use as a formative assessment instrument within the present trial context.

Keywords: Arabic Language Assessment; Edpuzzle; *Maharah Istima'* Test; Test Development.

مستخلص البحث

يهدف هذا البحث إلى تطوير اختبار مهارة الاستماع القائم على منصة إدبازل (Edpuzzle) لطلبة الصف الثامن بمدرسة SMP-IT Al-Faruq بمدينة غورontalo. واعتمد البحث منهج البحث والتطوير (Research and Development) باستخدام نموذج ADDIE الذي يشتمل على خمس مراحل، وهي: التحليل (Analysis)، والتصميم (Design)، والتطوير (Development)، والتنفيذ (Implementation)، والتقييم (Evaluation). وتمثلت المنتجات المطوّرة في مقاطع فيديو حوارية متحركة بوصفها موادّ للاستماع، وأسئلة اختيار من متعدد مدمجة في منصة Edpuzzle، وتعليمات الاختبار، ومفتاح الإجابة، وآلية تقديم الاختبار إلكترونياً. وجمعت البيانات من خلال الملاحظة، والمقابلة، والتحكيم من قبل الخبراء، ونتائج أداء الطلبة، وتحليل جودة فقرات الاختبار، واستبانة استجابات المعلم والطلبة. وأظهرت النتائج أن اختبار مهارة الاستماع القائم على منصة Edpuzzle حصل على متوسط نسبة صلاحية بلغ ٩٦,٧% وفق تقويم خبراء المادة، واللغة، والوسائط، مما يدل على أنه عالي الصلاحية. كما أظهرت نتائج تحليل الفقرات أن ٣٤ فقرة من أصل ٣٦ فقرة كانت صالحة، في حين تراوحت معاملات الثبات (Cronbach's Alpha) بين ٠,٧٤٤ و ٠,٨٢٩. وأشارت نتائج تحليل مستوى الصعوبة إلى وجود ٢٦ فقرة سهلة و ٨ فقرات متوسطة الصعوبة، بينما يبين تحليل قوة التمييز وجود فقرة واحدة ممتازة، و ١٨ فقرة جيدة، و ١١ فقرة مقبولة، و ٤ فقرات ضعيفة. كما أثبت تحليل المشتتات أن ٤٠ مشتتاً من أصل ٥٦ مشتتاً كانت فعالة. وبلغ متوسط نسبة القابلية للتطبيق ٩١,٤% بناءً على استجابات المعلم والطلبة، وهو ما يندرج ضمن فئة عالية القابلية للتطبيق. وبناءً على ذلك، يمكن الاستنتاج أن اختبار مهارة الاستماع القائم على منصة Edpuzzle يتمتع بدرجة عالية من الصلاحية، وجودة مناسبة لفقراته، ودرجة عالية من القابلية للتطبيق، مما يجعله أداة مناسبة لتقويم مهارة الاستماع في تعليم اللغة العربية.

الكلمات المفتاحية: تطوير الاختبار؛ اختبار مهارة الاستماع؛ إدبازل؛ تقويم اللغة العربية.

Citation:

Katili, I. S., Hula, I. R. N., Bahri, R. B. H., Hairuddin, & Yasin, Z. (2026). Development of an Edpuzzle-Based *Maharah Istima'* Test for Junior High School Students. *Al-Muyassar: Journal of Arabic Education*, 5 (3), 1112-1137.

INTRODUCTION

Arabic language learning aims to develop four language skills, namely *maharah istima'* (listening), *maharah kalam* (speaking), *maharah qira'ah* (reading), and *maharah kitabah* (writing) (Asyrofi & Pransiska, 2019). Among these skills, *maharah istima'* serves as the foundation for language acquisition because learners receive linguistic input through listening before producing spoken or written language (Afyuddin & Maarif, 2023; Chalik, 2021). Therefore, listening assessment is necessary to determine the extent to which students can comprehend spoken Arabic and to provide information that can be used to improve subsequent learning (Hilmi, 2018; Kosbandhono, 2013).

Despite its importance, the implementation of the *Maharah Istima'* Test in schools still encounters several challenges. Listening assessment is generally conducted using conventional methods, such as teachers reading texts aloud or playing audio recordings followed by paper-based questions (Andayani & Madani, 2023; Arifin, 2012). These practices tend to be less interactive, require considerable time for scoring, and provide limited information about students' listening performance. Consequently, teachers often experience difficulties in conducting continuous formative assessment and obtaining objective evidence of students' listening achievement (Hilmi, 2018).

A preliminary study conducted at SMP-IT Al-Faruq, Gorontalo City, revealed similar conditions. Classroom observations and interviews with the Arabic language teacher showed that the school had not yet implemented a structured *Maharah Istima'* Test aligned with specific listening competency indicators. In addition, *Al-'Arabiyyah Bayna Yadaik* Volume 1A, which is used as the primary instructional material, provides listening exercises but does not include a systematic formative test specifically designed to measure students' listening achievement. Consequently, the teacher required an assessment instrument that could measure students' listening comprehension more systematically, objectively, and efficiently. These findings indicate a practical gap between the need for structured formative listening assessment and the assessment instruments currently available in the classroom.

Recent developments in educational technology provide opportunities to address this problem through technology-enhanced assessment (Asiri, 2021; Rachman et al., 2023). Edpuzzle is one digital platform that allows teachers to integrate questions directly into videos, record students' responses, provide immediate scoring or feedback, and organize assessment results electronically (Zuhriyah, 2025). These functions are particularly relevant to listening assessment because audiovisual input and assessment

questions can be presented within the same digital environment, while students' responses can be recorded automatically. Previous studies have also reported that Edpuzzle can support more engaging and interactive learning activities (Achmad et al., 2021; Agustin et al., 2024; Hamid, 2024; Nabila, 2024). Therefore, Edpuzzle has the potential not only to function as a learning medium but also to serve as a platform for developing structured digital listening assessment.

Several previous studies have examined digital platforms in language learning and assessment. Previous research developed a Moodle-based assessment for *maharah istima'* (Ismail & Musthafa, 2018), while another study applied Kahoot to Arabic language assessment (Hanafi et al., 2019). Studies involving Edpuzzle have generally emphasized its use as a learning medium, its influence on student engagement, or students' perceptions of its use in listening activities (Achmad et al., 2021). However, these studies have not primarily focused on developing an Edpuzzle-based Arabic listening test for junior high school students and simultaneously examining its feasibility, psychometric item quality, and practicality. In particular, empirical evidence remains limited regarding an Edpuzzle-based *Maharah Istima' Test* that is systematically developed and evaluated through expert validation, item validity, reliability, difficulty index, discrimination index, distractor effectiveness, and user practicality.

Table 1 summarizes several previous studies related to the use of digital platforms in Arabic language learning and assessment.

Table 1. State of the Art of Previous Studies

Researchers	Research Focus	Main Findings	Research Limitations
Ismail & Musthafa (2018)	Development of Moodle-based <i>Maharah Istima'</i> assessment	Produced a digital listening assessment using Moodle.	Did not employ Edpuzzle and did not conduct comprehensive psychometric item analysis.
Hanafi et al. (2019)	Arabic assessment using Kahoot	Digital assessment improved classroom evaluation.	Focused on Kahoot implementation rather than complete test development.
Achmad et al. (2021)	Edpuzzle in learning	Edpuzzle increased students' learning interest and classroom engagement.	Focused on learning media instead of assessment instrument development.
Hamid (2024)	Edpuzzle for listening assessment	Students perceived Edpuzzle positively in listening assessment.	Did not develop a listening test instrument nor analyze item quality.
Agustin et al. (2024)	Edpuzzle-based learning media	Produced interactive learning media using Edpuzzle.	Focused on media development rather than Arabic listening assessment.
This Study	Development of Edpuzzle-Based <i>Maharah Istima'</i> Test	Developed a complete assessment instrument including animated listening materials, interactive items, expert validation, practicality evaluation, and comprehensive psychometric item analysis.	Limited to one school and 30 eighth-grade students.

Based on the comparison presented in Table 1, the specific research gap addressed in this study concerns not merely whether Edpuzzle can support learning, but whether an Edpuzzle-based *Maharah Istima'* Test can be systematically developed for eighth-grade students and demonstrate adequate feasibility, psychometric item quality, and practicality. Previous studies have investigated digital learning platforms and listening assessment separately, whereas evidence concerning a complete Edpuzzle-based Arabic listening assessment instrument supported by both expert judgment and empirical item analysis remains limited.

The novelty of this study lies in integrating animated Arabic conversation videos,

interactive multiple-choice items embedded in Edpuzzle, expert validation, practicality evaluation, and comprehensive psychometric item analysis within a single assessment-development framework for eighth-grade students. Thus, Edpuzzle is treated not merely as a medium for delivering instructional content, but as a platform for constructing and administering an empirically evaluated *Maharah Istima'* Test. This combination distinguishes the present study from previous research that primarily examined Edpuzzle as a learning medium or focused on users' perceptions of its implementation.

Accordingly, this study aims to develop an Edpuzzle-based *Maharah Istima'* Test for eighth-grade students at SMP-IT Al-Faruq, Gorontalo City, and to examine the resulting instrument in terms of its feasibility, psychometric item quality, and practicality. Specifically, the study addresses how the test is developed through the ADDIE process, whether it meets expert-defined feasibility criteria, how its items perform empirically, and how practical the instrument is for teachers and students.

RESEARCH METHOD

This study employed a Research and Development (R&D) approach using the ADDIE development model, which consists of five sequential stages: Analysis, Design, Development, Implementation, and Evaluation (Hidayat & Nizar, 2021; Waruwu, 2024). The model was used to systematically develop, validate, implement, and refine an Edpuzzle-based *Maharah Istima'* Test. The study was developmental and descriptive in nature, focusing on the development process, product feasibility, empirical item quality, and practicality rather than on measuring instructional effectiveness.

Development Procedures

During the **Analysis stage**, a needs analysis was conducted through classroom observation, an interview with the Arabic language teacher, and a review of the instructional materials used at the research site. The analysis focused on existing *maharah istima'* learning and assessment practices, students' characteristics, available facilities, the use of technology in assessment, and the need for a structured listening assessment instrument. The results of this stage provided the basis for determining the characteristics and components of the product to be developed.

During the **Design stage**, the objectives of the test and the indicators of *maharah istima'* were formulated. The indicators included the ability to recognize vocabulary or expressions heard, understand the meaning of simple sentences, identify main information, identify detailed information, and select an appropriate response to an utterance or conversation. The listening materials were selected from *Al-'Arabiyyah Bayna*

Yadaik Volume 1A based on the materials previously studied by the students. Four formative test packages containing a total of 36 items were designed. Packages 1 and 2 each consisted of eight items, while Packages 3 and 4 each consisted of ten items. The design stage also included the preparation of the test blueprint, conversation scripts, animated-video design, test items, test instructions, answer key, and the placement of questions at predetermined points in the Edpuzzle videos.

During the **Development stage**, the designed materials were transformed into animated conversation videos and integrated into Edpuzzle. The videos were uploaded and linked to Edpuzzle, and the test items were embedded at selected points so that students answered questions after listening to the relevant parts of the conversations. The initial product was subsequently evaluated by three experts specializing in material, language, and instructional media. Both quantitative ratings and qualitative comments from the experts were used as the basis for revising the materials, language, test items, instructions, audiovisual components, and technical presentation before classroom implementation.

During the **Implementation stage**, the revised product was administered to 30 eighth-grade students at SMP-IT Al-Faruq, Gorontalo City. Before completing the test, the students received instructions on accessing Edpuzzle, listening to the animated conversations, answering the embedded questions, and completing the assessment. Students' responses were recorded through Edpuzzle and subsequently used for empirical item analysis. After the test, the Arabic language teacher and students completed practicality questionnaires regarding their experience with the developed assessment instrument.

During the **Evaluation stage**, the product was reviewed based on expert-validation results, students' test responses, psychometric item analysis, and teacher and student responses. Items or product components that did not meet the predetermined criteria were revised as necessary. The final output of this stage was the refined Edpuzzle-based *Maharah Istima'* Test. The overall development procedure is illustrated in Figure 1.

Figure 1. Development Process of the Edpuzzle-Based Maharah Istima' Test Using the ADDIE Model



Research Site and Participants

This study was conducted at SMP-IT Al-Faruq, Gorontalo City, Indonesia, during the even semester of the 2025/2026 academic year. The implementation stage involved 30 eighth-grade students as product-trial participants and one Arabic language teacher who participated in the practicality assessment.

Three expert validators were selected based on their academic qualifications and areas of expertise. The material expert held a doctoral degree, served as a lecturer in Arabic Language Education at IAIN Sultan Amai Gorontalo, and had expertise in Arabic grammar, particularly *nahwu* and *sharaf*. The language expert held a doctoral degree, was an Arabic Language Education lecturer at Universitas Muhammadiyah Gorontalo, and had expertise in linguistics and Arabic language use. The media expert held a doctoral degree, was affiliated with IAIN Sultan Amai Gorontalo, and had academic and professional experience in technology and digital media. Each expert evaluated only the aspects corresponding to his field of expertise.

Product Description

The product developed in this study was an Edpuzzle-based *Maharah Istima' Test* integrating animated Arabic conversation videos as listening materials, objective test items embedded in the videos, test instructions, an answer key, and an online assessment system through Edpuzzle. The test content was derived from *Al-'Arabiyyah Bayna Yadaik* Volume 1A and was aligned with the listening indicators and learning materials used at the research site.

Research Instruments

The instruments used in this study are summarized in Table 2.

Table 2. Research Instruments

Instrument	Source/Respondent	Main Purpose and Characteristics
Observation sheet	Classroom	To identify existing <i>maharah istima'</i> learning and assessment practices and the need for product development
Interview guide	Arabic Language Teacher	To identify assessment problems, students' characteristics, available facilities, and teacher needs
Material expert validation sheet	1 material expert	20 indicators assessing content, listening materials, items, answer choices, answer key, and appropriateness for junior high school students; five-point Likert scale
Language expert validation sheet	1 language expert	15 indicators assessing Arabic accuracy, vocabulary, sentence structure, meaning, pronunciation, intonation, speech rate, instructions, and readability; five-point Likert scale
Media expert validation sheet	1 media expert	22 indicators assessing display, audiovisual quality, accessibility, test flow, question-video integration, Edpuzzle functions, result monitoring, and usability; five-point Likert scale
Teacher response questionnaire	1 Arabic teacher	12 items covering interest, content, usability, and benefit; four-point Likert scale
Student response questionnaire	30 students	11 items covering interest, content, usability, and benefit; four-point Likert scale
Maharah Istima' Test	30 students	Four test packages comprising 36 items used to obtain test-response data for empirical item analysis

The research instruments consisted of observation sheets, interview guides, expert validation sheets, teacher and student response questionnaires, and the developed Maharah Istima' Test. As shown in Table 2, each instrument served a different function in collecting the qualitative and quantitative data required at various stages of the development process.

Expert Validation and Feasibility Analysis

Product feasibility was evaluated by the three expert validators using separate validation sheets corresponding to their areas of expertise. Each indicator was rated using a five-point Likert scale, with scores interpreted as follows: 5 = highly feasible/highly appropriate, 4 = feasible/appropriate, 3 = moderately

feasible/moderately appropriate, 2 = less feasible/less appropriate, and 1 = not feasible/not appropriate. The validators were also provided with space to give qualitative comments and recommendations for revision.

The feasibility percentage was calculated using:

$$P = \frac{S_o}{S_m} \times 100\%$$

where P represents the feasibility percentage, S_o is the obtained score, and S_m is the maximum possible score.

The percentages were interpreted using the following criteria: 81–100% = Highly Feasible, 61–80% = Feasible, 41–60% = Less Feasible, 21–40% = Not Feasible, and 0–20% = Very Infeasible. A product was considered suitable for further implementation when it reached the Feasible or Highly Feasible category. Regardless of the percentage obtained, qualitative recommendations from the validators were considered in revising the product before implementation.

Practicality Assessment

The practicality of the developed assessment instrument was evaluated using teacher and student response questionnaires administered after the implementation stage. Both questionnaires covered four aspects: interest, content, usability, and benefit. The teacher questionnaire contained 12 items, with three items assigned to each aspect. The student questionnaire contained 11 items: three items each for interest, content, and usability, and two items for benefit. The questionnaire blueprint specified the indicators and item allocation for each aspect.

Representative teacher items included *“The Edpuzzle-based Maharah Istima’ Test is easy to use in learning activities”* and *“Edpuzzle features make it easier for teachers to present listening materials and test questions.”* Representative student items included *“I can easily open and complete the test through Edpuzzle”* and *“The test instructions in Edpuzzle are easy to understand.”*

The questionnaires used a four-point Likert response format, scored as 4 = Strongly Agree, 3 = Agree, 2 = Disagree, and 1 = Strongly Disagree. No separate pilot test or expert-validation procedure was conducted for the teacher and student response questionnaires. Therefore, their results were treated as descriptive user-response data for evaluating product practicality rather than as evidence of the psychometric validity of a standardized questionnaire. Practicality scores were calculated using:

$$P = \frac{S_o}{S_m} \times 100\%$$

The percentages were interpreted as 81–100% = Highly Practical, 61–80% = Practical, 41–60% = Less Practical, 21–40% = Not Practical, and 0–20% = Very Impractical. The overall practicality score was obtained from the teacher and student response results.

Data Analysis

The collected data were analyzed using qualitative and quantitative techniques. Qualitative data obtained from classroom observation, teacher interviews, and experts' comments were analyzed descriptively through data organization, selection, presentation, and interpretation. These data were used to identify development needs and determine the revisions required during the development process.

Quantitative data from expert-validation sheets and practicality questionnaires were converted into percentages according to the formulas and interpretation criteria described above. Students' responses to the developed test were scored dichotomously, with 1 for a correct answer and 0 for an incorrect answer, and were analyzed to examine item validity, reliability, difficulty index, discrimination index, and distractor effectiveness (Kadir, 2015; Qadir et al., 2024).

Item validity was examined using the Pearson Product-Moment correlation between each item score and the total test score. With 30 participants, the critical correlation coefficient at the 0.05 significance level was $r_{table} = 0.361$. An item was considered valid when $r_{calculated} > 0.361$ and the significance value was less than 0.05; otherwise, the item was considered for revision or replacement. The same decision rule was used in the empirical analysis of the four test packages.

Test reliability was examined separately for each test package using Cronbach's Alpha in IBM SPSS Statistics version 32. The coefficients were interpreted according to the predetermined criteria used in this study: 0.80–1.00 = Very High, 0.60–0.79 = High, 0.40–0.59 = Moderate, 0.20–0.39 = Low, and 0.00–0.19 = Very Low. The reliability analysis therefore focused on the internal consistency of the items within each test package. The final empirical results showed that all four packages fell within the High or Very High categories.

Item difficulty was calculated using:

$$P = \frac{B}{N}$$

where P is the difficulty index, B is the number of students who answered the item correctly, and N is the total number of students who answered the item. The criteria were 0.00–0.30 = Difficult, 0.31–0.70 = Moderate, and 0.71–1.00 = Easy.

Discrimination power was analyzed by dividing the 30 students into upper and lower groups. The upper group consisted of the nine students with the highest scores, representing approximately 27% of the sample, while the lower group consisted of the nine students with the lowest scores. The discrimination index was calculated as:

$$D = \frac{B_U}{J_U} - \frac{B_L}{J_L}$$

where B_U is the number of students in the upper group who answered the item correctly, J_U is the number of students in the upper group, B_L is the number of students in the lower group who answered correctly, and J_L is the number of students in the lower group.

The discrimination indices were interpreted as 0.70–1.00 = Very Good, 0.40–0.69 = Good, 0.20–0.39 = Fair, 0.00–0.19 = Poor, and <0.00 = Very Poor. Items classified as Poor or Very Poor were considered for revision or replacement.

Distractor effectiveness was evaluated by examining the distribution of students' selections for each incorrect option. A distractor was classified as functioning when selected by at least 5% of the test participants, less functioning when selected by fewer than 5%, and non-functioning when it was not selected by any participant. A distractor selected more frequently by the upper group than the lower group was marked for revision.

The combined results of item validity, reliability, difficulty, discrimination, and distractor analyses were used to determine whether individual items should be retained, revised, or replaced in the final version of the *Maharah Istima'* Test.

For the psychometric analysis, response data were obtained from the same trial group of 30 eighth-grade students across the four test packages. The packages consisted of 8, 8, 10, and 10 items, respectively. All 30 students completed each of the four test packages during the product trial. Thus, the same cohort contributed item-response data to all four packages. Students' responses to each item were coded dichotomously as 1 for a correct response and 0 for an incorrect response. The item-level response data were

analyzed separately for each test package; therefore, item validity and Cronbach's Alpha were calculated within each package rather than by combining all 36 items into a single reliability coefficient. The descriptive distribution of students' overall scores was not used as the input for the reliability calculation. Before test administration, all participants received technical instructions on accessing Edpuzzle, listening to the videos, answering the embedded questions, and completing the test. However, no separate familiarization trial or experimental control was conducted to estimate possible learning or familiarity effects associated with repeated use of the Edpuzzle format. Therefore, such effects cannot be completely ruled out and are considered a limitation of the present study.

Scope of Product Evaluation

This study did not examine the instructional effectiveness of Edpuzzle in improving students' *maharah istima'*. Therefore, no pretest-posttest design, experimental treatment, or control group was employed. Edpuzzle functioned as the digital platform for developing and administering the assessment instrument rather than as an independent instructional treatment. The evaluation was limited to the product-development process, expert-based feasibility, empirical item quality, and practicality based on teacher and student responses. Future research may examine the effectiveness of the developed instrument or its instructional use through an experimental or quasi-experimental design.

RESULT AND DISCUSSION

The Development stage focused on producing the initial version of the Edpuzzle-based *Maharah Istima'* Test. The developed product consisted of animated conversation videos as listening materials, interactive multiple-choice questions embedded within the videos, test instructions, and answer keys. Following the product development, expert validation was conducted involving a material expert, a language expert, and a media expert. The material expert recommended adjusting the depth of the listening materials to suit junior high school students and reviewing several distractors. The language expert recommended improving the clarity of meaning, pronunciation, speech rate, and the wording of several questions and answer choices. The media expert recommended providing clearer and more detailed instructions for using the Edpuzzle platform. These recommendations were incorporated into the product revision before classroom implementation.

After the revision process had been completed, the final product was integrated into the Edpuzzle platform, enabling all assessment components to function as a unified digital assessment instrument. The developed product comprised animated conversation videos, interactive multiple-choice questions embedded at specific points in the videos, and an automated reporting system that records students' responses and scores. The main components of the developed product are presented in Figure 2.

Figure 2. Main Components of the Developed Edpuzzle-Based Maharah Istima' Test

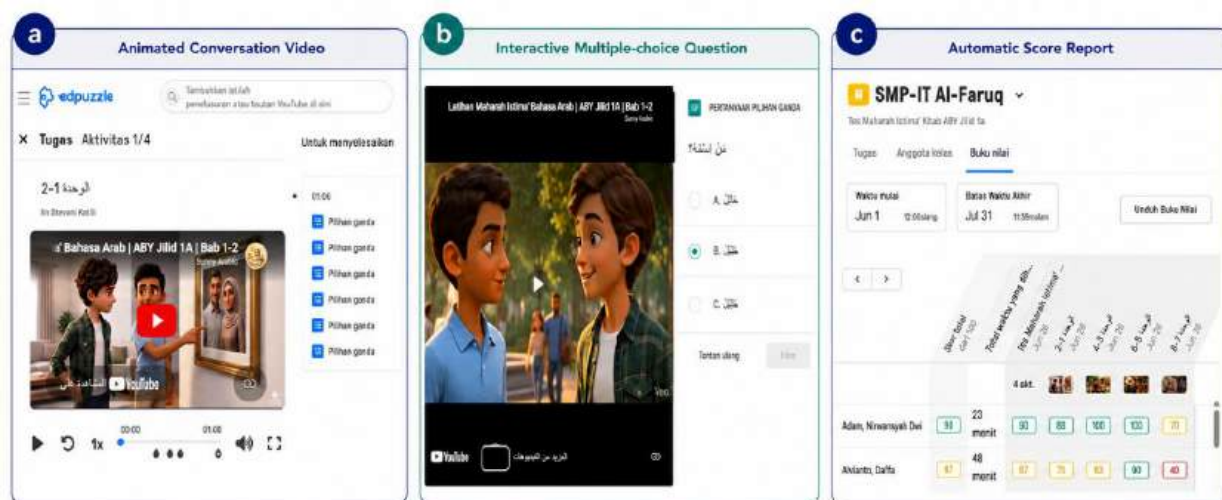
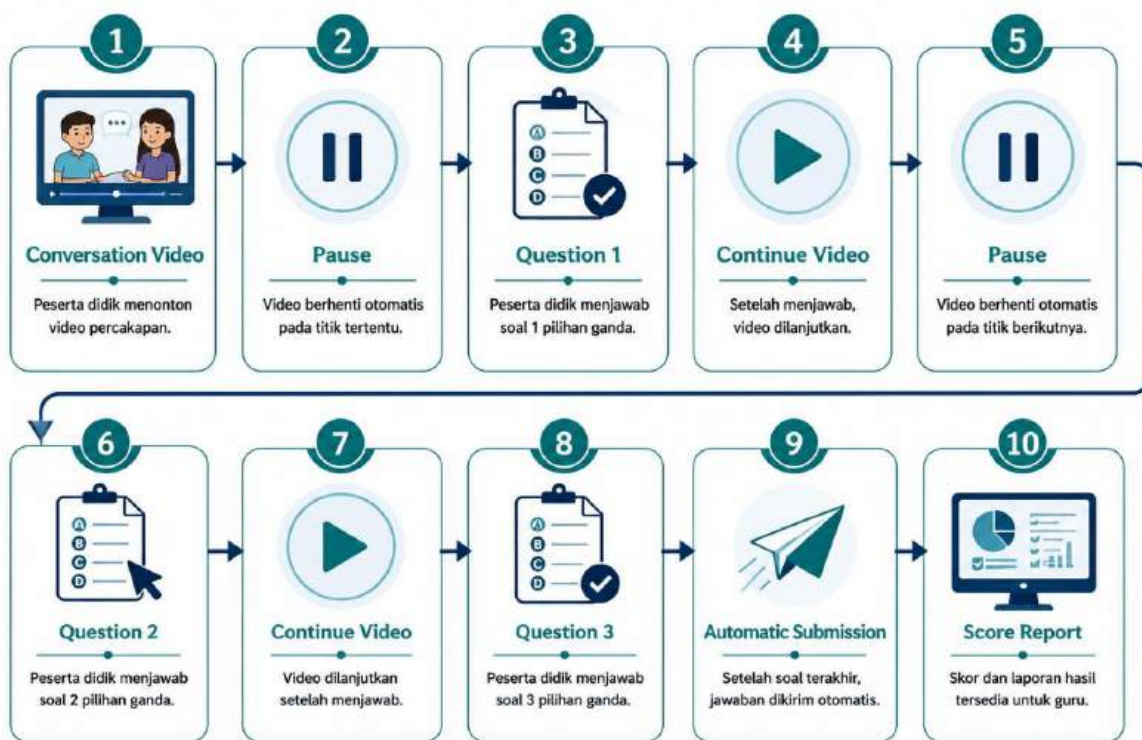


Figure 2 illustrates the principal components of the developed assessment instrument. The first display presents the animated conversation video used as the listening material. The second display shows an interactive multiple-choice question embedded directly within the video, allowing students to answer questions during the listening activity. The third display presents the automatic score report generated by Edpuzzle after students complete the assessment. The integration of these components enables teachers to administer listening assessments digitally while providing immediate access to students' assessment results.

Following the completion of the development and revision processes, the Edpuzzle-based Maharah Istima' Test was implemented with 30 eighth-grade students at SMP-IT Al-Faruq, Gorontalo City. The implementation aimed to evaluate the practicality of the developed assessment instrument and to collect students' test results for subsequent item quality analysis. During this stage, the Arabic language teacher and students also provided responses regarding the usability of the product, the clarity of the listening materials, the presentation of the test items, and the usefulness of Edpuzzle as a platform for conducting listening assessment.

The implementation was conducted through the Edpuzzle platform according to the assessment workflow designed during the Development stage. Students first watched the animated conversation videos as listening materials. At predetermined points, the videos paused automatically to display interactive multiple-choice questions. After selecting an answer, students continued watching the videos until the next question appeared. This sequence continued until all test items had been completed. Students' responses were recorded automatically, and the assessment results became immediately accessible to the teacher through the reporting system provided by Edpuzzle. The implementation workflow is presented in **Figure 3**.

Figure 3. Workflow of the Edpuzzle-Based Maharah Istima' Test



As illustrated in Figure 3, the assessment process integrates listening activities, interactive questioning, automatic submission, and score reporting into a single digital environment. This workflow allows students to complete the listening assessment independently while enabling teachers to obtain students' performance data efficiently. The automatic recording of students' responses also minimizes manual scoring and facilitates the subsequent analysis of item validity, reliability, difficulty index, discrimination index, and distractor effectiveness.

The data collected during the implementation stage served as the basis for evaluating the practicality of the developed assessment instrument and for conducting

psychometric item analysis. Product feasibility had previously been established through expert validation during the Development stage. The findings of these analyses are presented in the following sections.

Product Feasibility Based on Expert Validation

The feasibility of the developed Maharah Istimah Test was evaluated by three experts specializing in Arabic language content, language use, and instructional media. The validation focused on the appropriateness of the listening materials, alignment with learning objectives and listening indicators, linguistic accuracy, vocabulary selection, sentence structure, audio quality, video presentation, usability, and the suitability of the assessment format implemented through Edpuzzle. The experts also assessed the clarity of the test instructions and the appropriateness of the answer key.

Overall, the validation results indicated that the developed assessment instrument fulfilled the feasibility criteria in terms of content, language, and media. Several suggestions for improvement were provided before classroom implementation. The material expert recommended adjusting the depth of the listening materials to suit junior high school students and reviewing several distractors. The language expert recommended improving the clarity of meaning, pronunciation, speech rate, and the wording of several questions and answer choices. The media expert recommended providing clearer and more detailed instructions for using the Edpuzzle platform. These recommendations were used as the basis for revising the product before classroom implementation. The recapitulation of expert validation results is presented in Table 3.

Table 3. Recapitulation of Expert Validation Results

Validator	Max Score	Score Obtained	Percentage	Category
Material Expert	100	97	97%	Highly Feasible
Language Expert	75	69.5	92.7%	Highly Feasible
Media Expert	110	109	99%	Highly Feasible
Average	285	275.5	96.7%	Highly Feasible

Source: Expert validation results, 2026.

As shown in Table 3, the overall feasibility score reached 96.7%, indicating that the developed Maharah Istimah Test was highly feasible for classroom implementation. These findings demonstrate that the product satisfied the essential requirements of a digital Arabic listening assessment instrument in terms of content appropriateness, linguistic quality, and media presentation. Consequently, the product was considered appropriate for implementation and subsequent evaluation through students' trials and

psychometric item analysis.

Item Quality Analysis

The psychometric quality of the developed *Maharah Istima'* Test was examined using item validity, reliability, difficulty index, discrimination index, and distractor effectiveness. The analysis used item-level responses obtained from the same trial cohort of 30 eighth-grade students across the four test packages. Students' responses were scored dichotomously, with 1 assigned to a correct response and 0 to an incorrect response. Each test package was analyzed separately; therefore, item validity and Cronbach's Alpha were calculated within each package rather than by combining all 36 items into a single reliability coefficient. The descriptive distribution of students' overall scores was not used as the basis for calculating reliability.

Item validity analysis showed that 34 of the 36 developed items (94.44%) met the validity criterion. All eight items in Packages 1 and 2 were valid. In Package 3, nine of the ten items were valid, whereas item 1 did not meet the criterion ($r = 0.335$, $p = 0.$). Similarly, nine of the ten items in Package 4 were valid, while item 8 was invalid ($r = 0.333$, $p = 0.072$). These two items were therefore identified for revision or replacement in the final product.

Reliability was calculated separately for each test package using Cronbach's Alpha. The coefficients were 0.744 for Package 1, 0.818 for Package 2, 0.796 for Package 3, and 0.829 for Package 4, indicating high to very high internal consistency according to the criteria adopted in this study. These coefficients represent the consistency of item responses within each package and should not be interpreted as a single reliability estimate for the entire set of 36 items.

The difficulty analysis was conducted on the 34 valid items. The results identified 26 easy items and 8 moderately difficult items, with no items classified as difficult. The predominance of easy items is understandable because the instrument was designed for formative assessment of materials that the students had previously studied. Nevertheless, the absence of difficult items indicates a relatively restricted range of item difficulty and suggests the need for further refinement to provide a more balanced distribution of item difficulty.

The discrimination analysis showed that 1 item was classified as very good, 18 as good, 11 as fair, and 4 as poor. The four items with poor discrimination were Package 2 item 3, Package 3 items 8 and 9, and Package 4 item 1. Package 3 item 8 yielded a

discrimination index of 0.000, indicating that the item did not adequately differentiate between students in the upper and lower performance groups. These items were therefore identified as requiring revision or replacement.

Distractor analysis showed that 40 of the 56 distractors (71.4%) functioned effectively. Fourteen distractors were selected by fewer than 5% of the participants and were therefore classified as less functioning, while two distractors required revision because they were selected more frequently by students in the upper group than those in the lower group. These results indicate that most distractors functioned as intended, although several answer options still required refinement.

The overall psychometric results and the decision criteria used in the analysis are summarized in Table 4.

Table 4. Summary of Psychometric Item Analysis

Analysis Aspect	Decision Criteria	Result
Item Validity	$r_{calculated} > 0.361$ and $p < 0.05$	34 of 36 items valid (94.44%)
Reliability	0.80–1.00 = Very High; 0.60–0.79 = High; 0.40–0.59 = Moderate; 0.20–0.39 = Low; 0.00–0.19 = Very Low	P1 = 0.744; P2 = 0.818; P3 = 0.796; P4 = 0.829
Difficulty Index	0.00–0.30 = Difficult; 0.31–0.70 = Moderate; 0.71–1.00 = Easy	26 easy; 8 moderate; 0 difficult
Discrimination Index	0.70–1.00 = Very Good; 0.40–0.69 = Good; 0.20–0.39 = Fair; 0.00–0.19 = Poor; <0.00 = Very Poor	1 very good; 18 good; 11 fair; 4 poor
Distractor Effectiveness	Selected by $\geq 5\%$ = functioning; <5% = Less Functioning; not selected = Non-functioning	40 of 56 functioning (71.4%)

Note: Discrimination analysis used the upper–lower group method, with nine students, approximately 27% of the sample, assigned to each group.

The psychometric analysis was also used to identify problematic items that required further refinement. Decisions concerning item retention were not based on validity alone but considered the combined results of validity, difficulty, discrimination, and distractor analyses. For example, Package 2 item 3 met the item-validity criterion but showed poor discrimination, whereas Package 4 item 8 did not meet the item-validity

criterion. Table 5 presents examples of changes incorporated into the revised version of these problematic items.

Table 5. Examples of Revisions to Problematic Items

Item	Empirical Finding	Before Revision	Revised Version	Rationale
Package 2, Item 3	Poor discrimination	Image-based multiple-choice item with three answer alternatives	The revised item contained four image alternatives, including an additional distractor representing a narrow apartment	To increase the range and plausibility of distractors and reduce the likelihood of guessing
Package 4, Item 8	Did not meet the item-validity criterion	Multiple-choice item asking students to select an appropriate occupation for someone who likes children, with three occupation alternatives	The revised item contained four occupation alternatives, including an additional option (<i>engineer</i>)	To improve the quality and balance of the answer alternatives and strengthen item functioning

Source: Product revision, 2026.

These examples illustrate how empirical item information was considered in refining the final product. More broadly, Package 2 item 3, Package 3 items 8 and 9, and Package 4 item 1 required attention because of poor discrimination, while Package 3 item 1 and Package 4 item 8 required revision or replacement because they did not meet the validity criterion. The refinement process focused on improving item clarity, answer alternatives, distractor plausibility, item difficulty, and alignment with the listening materials and *maharah istima'* indicators.

Practicality Evaluation

The practicality of the developed *Maharah Istima' Test* was evaluated based on the responses of the Arabic language teacher and the students following the implementation stage. The practicality assessment covered four aspects: interest, content, usability, and benefit. The purpose of this evaluation was to determine whether the developed assessment instrument was practical and usable for classroom implementation.

The results indicated that both the teacher and the students responded positively to the developed assessment instrument. The teacher-response questionnaire yielded a practicality score of 95.0%, while the student-response questionnaire yielded an average

score of 87.8%. The overall practicality score reached 91.4%, which falls within the Highly Practical category. These findings indicate that the developed assessment instrument was easy to administer, simple to operate, and well accepted by its intended users. The summary of the practicality evaluation is presented in Table 6.

Table 6. Practicality Evaluation Results

Respondent	Score (%)	Category
Teacher	95%	Highly Practical
Students	87.8%	Highly Practical
Average	91.4%	Highly Practical

Source: Response questionnaire results, 2026.

The high teacher-response score indicates a positive perception of the usability and benefit of the developed assessment instrument. Similarly, the students' responses indicate positive perceptions of its interest, content, usability, and benefit. Taken together, these results support the Highly Practical category obtained from the response questionnaires. The results should be interpreted as descriptive user-response data concerning the practicality of the instrument rather than as evidence of improved learning outcomes.

Discussion

The present study developed an Edpuzzle-based *Maharah Istima'* Test that integrates animated conversation videos, interactive assessment items, expert validation, practicality evaluation, and psychometric item analysis within a single digital assessment instrument. Unlike previous studies that primarily examined Edpuzzle as a learning medium or focused on users' perceptions of its implementation, the present study positioned Edpuzzle as a platform for developing and administering an Arabic listening assessment instrument and empirically evaluating its item quality (Achmad et al., 2021; Agustin et al., 2024; Jusnidar et al., 2022; Rachman et al., 2023). This approach extends the use of Edpuzzle from instructional media to technology-enhanced assessment in Arabic language learning.

The expert-validation results indicate that the developed assessment instrument achieved a high level of feasibility in terms of material, language, and media. The revision process based on expert recommendations was an important part of the development process because several aspects of the listening materials, item formulation, speech rate, answer choices, and test instructions required refinement before classroom implementation. Such iterative revision is consistent with the developmental nature of the ADDIE model, in which evaluation and revision are incorporated throughout the

development process (Hidayat & Nizar, 2021; Waruwu, 2024).

The psychometric analysis provides initial empirical support for the quality of the developed *Maharah Istima'* Test, although several findings require cautious interpretation. The high proportion of items meeting the validity criterion and the Cronbach's Alpha coefficients ranging from 0.744 to 0.829 indicate satisfactory item-total relationships and internal consistency within the present trial. However, these coefficients were obtained from a relatively small sample of 30 students from a single school and should therefore be regarded as preliminary and context-dependent rather than as definitive evidence of psychometric stability across broader populations.

The predominance of easy items is understandable in relation to the formative purpose of the instrument, which was designed to assess students' comprehension of previously learned materials. Formative assessment is primarily intended to monitor students' learning progress and provide information for subsequent learning rather than to classify students according to achievement levels (Andayani & Madani, 2023; Arifin, 2012). Nevertheless, the absence of difficult items indicates a restricted range of item difficulty. Consequently, the instrument may be less sensitive in distinguishing students at the upper end of the ability distribution. Future refinement should therefore increase the proportion of moderately difficult and selectively more challenging items while maintaining alignment with the learning objectives.

Item validity and discrimination should also be interpreted as complementary rather than interchangeable indicators of item quality. Item validity reflects the relationship between an individual item and the total test score, whereas the discrimination index indicates how effectively an item differentiates between higher- and lower-performing students (Nurhalimah et al., 2022; Saputri et al., 2023). This distinction helps explain why some statistically valid items still demonstrated poor discrimination. Package 2 item 3, for example, met the item-validity criterion but showed poor discrimination. Similarly, Package 3 item 8 yielded a discrimination index of 0.000, indicating insufficient differentiation between the upper and lower performance groups. These findings show that a valid item does not automatically possess adequate discriminating power and that decisions concerning item quality should be based on multiple psychometric indicators rather than on validity alone.

The combined analysis of validity, difficulty, discrimination, and distractor functioning therefore provided a more comprehensive basis for item revision. Items with poor discrimination required review of their difficulty, wording, and answer

alternatives, while items that did not meet the validity criterion required more substantial revision or replacement. These empirical findings can also be interpreted in relation to the expert feedback obtained during product development. The material expert recommended adjusting the depth of the listening content to the students' level, which may partly explain the predominance of easy items after revision. Although this adjustment was pedagogically appropriate for junior high school students, the absence of difficult items indicates that the resulting difficulty range remained relatively narrow. The language expert also recommended adjustments to sentence structure, pronunciation, and speech rate. In a listening assessment, these features are important because excessively complex language or inappropriate audio pacing may interfere with comprehension independently of the intended listening construct. However, because these aspects were revised before the product trial, the present data cannot isolate their specific effects on individual item statistics. In contrast, distractor quality was directly reflected in the empirical analysis: several distractors attracted fewer than 5% of students, while some problematic items showed poor discrimination. These findings highlight the relevance of content alignment, distractor plausibility, and the technical presentation of the listening materials when refining item quality.

The findings of this study should also be distinguished from evidence of instructional effectiveness. The present study did not employ a pretest-posttest design, an experimental treatment, or a control group; therefore, students' test performance cannot be interpreted as evidence that the use of Edpuzzle improved their *maharah istima'*. This distinction is important because studies that aim to establish improvement in Arabic language skills generally require an experimental comparison of learning outcomes. For example, an experimental study on Arabic writing skills examined learning effectiveness by comparing experimental and control groups (Hilira et al., 2024). In contrast, Edpuzzle in the present study functioned as the digital platform through which the developed assessment was administered. The findings are therefore limited to product feasibility, psychometric item functioning, internal consistency, and practicality. Experimental or quasi-experimental research would be required to determine whether Edpuzzle-based instruction or assessment contributes to measurable improvements in listening achievement.

In addition, although students received technical instructions before completing the assessment, no separate familiarization trial was conducted to isolate possible effects of increasing familiarity with the Edpuzzle format. Repeated exposure to the same digital

assessment environment may have influenced students' ease of navigation or response behavior across the test sequence. This possibility should therefore be considered when interpreting the present findings and addressed in future validation studies.

The practicality evaluation further indicates that the developed assessment instrument was practical and usable for classroom implementation. The teacher response reached 95.0%, while the students' responses reached 87.8%, resulting in an overall practicality score of 91.4%. These descriptive response data suggest positive perceptions of the instrument in terms of interest, content, usability, and benefit. The findings are consistent with previous studies reporting positive student responses and engagement associated with the use of Edpuzzle in learning activities (Achmad et al., 2021; Hamid, 2024). However, the practicality findings should be interpreted as users' responses to the implementation of the assessment instrument rather than as evidence of improved learning outcomes.

From an instructional perspective, the developed test is most appropriately used as a formative assessment after students have completed the relevant lesson or thematic unit rather than as a one-time measure of final achievement. Each test package may be administered after the corresponding listening materials have been taught so that the resulting response data can inform subsequent instruction. After students complete the assessment, Edpuzzle automatically records their responses and scores, allowing the teacher to identify items or listening indicators that produce frequent errors. Rather than treating the score merely as a numerical outcome, the teacher can use these data to determine which vocabulary, expressions, detailed information, or listening contexts require further explanation and practice.

Operationally, this creates a formative feedback cycle consisting of assessment, identification of learning difficulties, instructional follow-up, and reassessment. When a substantial proportion of students answers a particular item incorrectly, the teacher may revisit the relevant listening material, provide focused explanation or repeated listening practice, and use similar items in a subsequent formative activity to determine whether comprehension has improved. Students' positive responses to the Edpuzzle format suggest that this cycle can be incorporated into the regular learning process, although technical orientation remains important for students who are not yet familiar with the platform. Therefore, the formative value of Edpuzzle in this study lies not only in automated scoring but also in making students' response data available promptly so that teachers can plan remediation and subsequent learning activities.

From a theoretical perspective, this study contributes to Arabic language assessment by demonstrating the importance of combining conceptual evaluation and empirical item analysis in the development of technology-based assessment instruments. Expert validation provides evidence concerning the appropriateness of material, language, and media, whereas empirical analysis provides information concerning item validity, reliability, difficulty, discrimination, and distractor functioning. The integration of these forms of evaluation provides a more comprehensive basis for refining Arabic language assessment instruments. This distinction represents the primary contribution of the present study because previous research has rarely combined expert validation, practicality assessment, and item quality analysis within a single assessment development framework (Ba'ul Chusna & Niswatul Fauziah, 2021; Hanafi et al., 2019).

From a practical perspective, the developed *Maharah Istima'* Test provides Arabic language teachers with an alternative digital assessment instrument that integrates listening materials, interactive assessment, automatic scoring, and assessment reports within a single platform. This finding also supports the broader use of digital platforms in Arabic language assessment, as technology-based test design can be adapted to different language skills and assessment purposes (Katili & Hula, 2025). Consequently, the developed instrument may assist teachers in conducting formative assessment more efficiently while maintaining acceptable standards of assessment quality (Andayani & Madani, 2023).

Overall, the contribution of this study lies in developing an Edpuzzle-based *Maharah Istima'* Test that is conceptually feasible, empirically evaluated, and practical within the context of its initial implementation. At the same time, the findings indicate that further item refinement and validation with larger and more diverse samples are required before stronger claims can be made regarding the psychometric stability and broader applicability of the instrument.

CONCLUSION

This study developed an Edpuzzle-based *Maharah Istima'* Test for eighth-grade students at SMP-IT Al-Faruq, Gorontalo City, using the ADDIE development model. The developed instrument integrates animated Arabic conversation videos, interactive test items, automatic scoring, and digital reporting within the Edpuzzle platform. Expert validation yielded an average feasibility score of 96.7%, while teacher and student responses produced an overall practicality score of 91.4%. In the product trial, 34 of the 36 items met the item-validity criterion, and the four test packages produced Cronbach's

Alpha coefficients ranging from 0.744 to 0.829. These results provide preliminary support for the feasibility, practicality, and empirical item quality of the developed assessment instrument in the context in which it was tested.

The psychometric findings should nevertheless be interpreted cautiously. The study involved only 30 students from a single school, and the item analysis showed a predominance of easy items, no difficult items, and four items with poor discrimination. These findings indicate that although the instrument is appropriate for formative assessment within the present context, further refinement is needed to improve the balance of item difficulty and its ability to differentiate students across a wider range of listening proficiency. The psychometric coefficients reported in this study should therefore not be generalized directly to other schools or student populations without further validation.

This study did not investigate improvement in students' listening achievement because no pretest-posttest design, experimental treatment, or control group was employed. In addition, no separate familiarization trial was conducted to isolate possible effects of increasing familiarity with the Edpuzzle format. Future studies should validate the instrument with larger and more diverse samples across multiple schools, examine revised items with a broader range of difficulty, and conduct independent psychometric cross-validation. Experimental or quasi-experimental studies may subsequently examine whether the use of Edpuzzle-based assessment or instruction contributes to measurable improvements in *maharah istima'*.

ACKNOWLEDGEMENT

The authors would like to express their sincere appreciation to the Headmaster, the Arabic language teacher, and the eighth-grade students of SMP-IT Al-Faruq, Gorontalo City, for their cooperation and valuable support throughout the development and implementation of the Edpuzzle-based Maharah Istima' Test. The authors are also grateful to the material, language, and media experts for their valuable evaluations and constructive feedback, which contributed significantly to improving the quality of the developed assessment instrument. In addition, the authors acknowledge the Postgraduate Program of IAIN Sultan Amai Gorontalo for its academic support throughout this research. This research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors.

REFERENCES

- Achmad, N., Ganiati, M., & Kur'aeni, D. N. (2021). Implementasi Edpuzzle dalam Meningkatkan Minat Belajar Peserta Didik pada Era New Normal. *UJMES*, 06(02), 46–51. <https://doi.org/10.30999/ujmes.v6i2.1283>
- Afyuddin, M. S., & Maarif, M. S. (2023). Istima' (Listening) Learning Strategy Used by Arabic Education Students. *Al-Irfan : Journal of Arabic Literature and Islamic Studies*, 6(1), 138–166. <https://doi.org/10.58223/alirfan.v6i1.6617>
- Agustin, S. E., Yandari, I. A. V., & Yuhana, Y. (2024). Pengembangan Media Pembelajaran Berbasis Aplikasi Edpuzzle Pada Materi Bangun Datar Kelas IV Sekolah Dasar. *Jurnal Ilmiah Profesi Pendidikan*, 9(1), 119–128. <https://doi.org/10.29303/jipp.v9i1.1839>
- Andayani, T., & Madani, F. (2023). Peran Penilaian Pembelajaran Dalam Meningkatkan Prestasi Siswa di Pendidikan Dasar. *Jurnal Educatio FKIP UNMA*, 9(2), 924–930. <https://doi.org/10.31949/educatio.v9i2.4402>
- Arifin, Z. (2012). *Evaluasi Pembelajaran* (W. B. Shidiq (ed.); Revisi 2). Direktorat Jendral Pendidikan Islam Kementerian Agama.
- Asiri, A. (2021). Edpuzzle : A Formative Assessment Tool. *WAESOL Educator*, 46(2), 34–35. <https://waesol.org/wp-content/uploads/2022/01/12.-Asiri-A.-2021.-Edpuzzle-A-formative-assessment-tool.-WAESOL-Educator-462-34-35..pdf>
- Asyrofi, S., & Pransiska, T. (2019). *Desain Pembelajaran Bahasa Arab: Model, Strategi dan Permainan Edukatif dalam Pembelajaran Bahasa Arab* (Cetakan 1). Pustaka Ilmu.
- Ba'ul Chusna, A., & Niswatul Fauziah, H. (2021). Pengembangan Instrumen Tes Kemampuan Bahasa Arab Bagi Mahasiswa IAIN Ponorogo. *Ibriez : Jurnal Kependidikan Dasar Islam Berbasis Sains*, 6(2), 263–270. <https://doi.org/10.21154/ibriez.v6i2.178>
- Chalik, S. A. (2021). Metode dan Strategi Pembelajaran Istima'. *Shaut Al- ' Arabiyah*, 9(2), 269–281. <https://doi.org/10.24252/saa.v9i2.31777>
- Hamid, S. F. (2024). Students' Perception of Using Edpuzzle in Listening Assessment. *PRIMACY Journal of English Education and Literacy*, 1(1), 15–24. <https://doi.org/10.33592/primacy.v1i1.2608>
- Hanafi, Y., Huda, I. S., & Rasydiana. (2019). Pengembangan Tes Pembelajaran Bahasa Arab Menggunakan Aplikasi Kahoot di MTsN 2 Kota Malang. *Prosiding Konferensi Nasional Bahasa Arab*, 5, 437–447. <https://prosiding.arab-um.com/index.php/konasbara/article/view/513>
- Hidayat, F., & Nizar, M. (2021). Model ADDIE (Analysis, Design, Development, Implementation and Evaluation) dalam Pembelajaran Pendidikan Agama Islam. *Jurnal Inovasi Pendidikan Agama Islam*, 1(1), 28–38. <https://doi.org/10.15575/jipai.v1i1.11042>
- Hilira, A. P., Kasmianti, Halomoan, & Syaf, M. N. (2024). The Effectiveness of Using the Hijaiyah Game within the Natural Method to Enhance Writing Skills. *Al-Kawakib*, 5(1), 20–26. <https://doi.org/10.24036/kwkib.v5i1.181>
- Hilmi. (2018). Teknik Evaluasi Aktual terhadap Maharah Lughawiyah. *Jurnal Mimbar Akademika*, 3(2), 173–200. <https://www.mimbarakademika.com/index.php/jma/article/download/66/65?>
- Ismail, M., & Musthafa, A. A. (2018). Pengembangan Materi Tes Keterampilan Menyimak Bahasa

Arab Menggunakan Moodle Untuk Meningkatkan Kemampuan Menyimak Mahasiswa Universitas Darussalam Gontor. *At-Ta'dib*, 13(2), 28–49. <https://doi.org/10.21111/at-tadib.v13i2.2467>

- Jusnidar, J., Mannahali, M., & Achmad, A. K. (2022). Media Edpuzzle dalam Pembelajaran Menyimak Bahasa Jerman. *Phonologie: Journal of Language and Literature*, 2(2), 188–193. <https://doi.org/10.26858/phonologie.v2i2.35298>
- Kadir, A. (2015). Menyusun dan Menganalisis Tes Hasil Belajar. *Al-Ta'dib*, 8(2), 70–81. <https://doi.org/10.31332/atdb.v8i2.411>
- Katili, I. S., & Hula, I. R. N. (2025). Interactive Arabic Writing Skills Test Design Using a Mentimeter: Research and Development. *Alsuna: Journal of Arabic and English Language*, 8(1), 53–70. <https://doi.org/10.31538/alsuna.v8i1.6596>
- Kosbandhono, E. (2013). Asesmen dan Evaluasi untuk Maharah Istima'. *Arabia*, 5(1), 1–12. <https://doi.org/10.21043/arabia.v5i1.1390>
- Nabila, D. A. P. (2024). *The Effectiveness Of Edpuzzle For Listening Skills* [Maulana Malik Ibrahim State Islamic University of Malang]. <http://etheses.uin-malang.ac.id/66702/1/200107110016.pdf>
- Nurhalimah, S., Hidayati, Y., Rosidi, I., & Hadi, W. P. (2022). Hubungan antara Validitas Item dengan Daya Pembeda dan Tingkat Kesukaran Soal Pilihan Ganda Pas. *Natural Science Educational Research*, 4(3), 249–257. <https://doi.org/10.21107/nser.v4i3.8682>
- Qadir, A., Huda, N., & Hermina, D. (2024). Analisis Butir Tes Tingkat Kesukaran, Daya Pembeda dan Efektivitas Pengecoh. *Al Furqan: Jurnal Agama, Sosial, Dan Budaya*, 3(3), 1450–1467. <https://publisherqu.com/index.php/Al-Furqan/article/view/1069/962>
- Rachman, A., Oktoviandry, R., Putri, S. D., Ningsih, A. G., Ulya, R. H., Indriyani, V., Sari, H. Y., Henanggil, M. D. F., & Juita, N. (2023). Pelatihan Pembuatan Video Pembelajaran Interaktif Berbantuan Aplikasi Edpuzzle Bagi Guru SMPN 4 Harau Kabupaten Lima Puluh Kota. *Community Development Journal: Jurnal Pengabdian Masyarakat*, 4(5), 10928–10933. <https://doi.org/10.31004/cdj.v4i5.21854>
- Saputri, H. A., Zulhijrah, Larasati, N. J., & Shaleh. (2023). Analisis Instrumen Assesmen : Validitas, Reliabilitas, Tingkat Kesukaran, dan Daya Beda Butir Soal. *Didaktik : Jurnal Ilmiah PGSD FKIP Universitas Mandiri*, 09(05), 2986–2995. <https://doi.org/10.36989/didaktik.v9i5.2268>
- Waruwu, M. (2024). Metode Penelitian dan Pengembangan (R&D): Konsep, Jenis, Tahapan dan Kelebihan. *Jurnal Ilmiah Profesi Pendidikan*, 9(2), 1220–1230. <https://doi.org/10.29303/jipp.v9i2.2141>
- Zuhriyah, N. (2025). Development of Interactive Video Using Edpuzzle to Improve Evaluation of Arabic Learning for Students of University. *ISoLEC (International Seminar on Language, Education, and Culture)*, 9(1), 64–69. <https://conference.um.ac.id/index.php/isolec/article/viewFile/10672/4414>