



Evaluating Arabic Language Competence Achievement in Indonesian *Madrasah Tsanawiyah*: A Criterion-Referenced Approach

Fakhrudin Aman¹, Abdul Basith², Miftahul Ula³, Ahmad Taufiq⁴
^{1,2,3,4} Universitas Islam Negeri K.H. Abdurrahman Wahid Pekalongan
<https://doi.org/10.21580/rjtpd.v3i02.33624>

ABSTRACT

Criterion-Referenced Assessment (PAP) is a strategic evaluation instrument in Arabic language learning, yet its implementation in madrasah still faces challenges in standard construction, achievement interpretation, and pedagogical follow-up. This study aims to explore the dynamics of PAP in evaluating Arabic language learning among seventh-grade students at MTs At-Taqwa Jatingarang. This study employs a qualitative approach with a case study design, involving teachers, students, parents, and vice principal as informants. Data were collected through semi-structured interviews, observation, and document analysis, then analyzed thematically with source and technique triangulation. The findings reveal that competence standard construction is contextual through negotiation between national curriculum and madrasah characteristics. Achievement interpretation is conducted through a qualitative matrix based on four *maharah* skills, enabling precise identification of competence gaps. Furthermore, PAP transforms into the heart of a continuous learning improvement cycle through remedial programs, enrichment, instructional strategy adaptation, and collaboration with parents. This study offers a new conceptual model that transforms PAP from an administrative instrument into a participatory and improvement-oriented evaluation ecosystem.

Keywords: Criterion-Referenced Assessment, Arabic Language Learning, Learning Evaluation, Madrasah, *Maharah* Skills.

Contact: fakhrudin.aman25004@mhs.uingusdur.ac.id *

© 2025 the author: All rights reserved. The authors agree that this article remains permanently open access under the terms of the Research Journal on Teacher Professional Development

1. Introduction

Evaluation of Arabic language learning in madrasahs plays a strategic role in ensuring the quality of Islamic education and students' communicative literacy (Wijaya et al., 2025). Normatively, Benchmark-Referenced Assessment (PAP) is designed as a diagnostic instrument that measures the level of absolute competency mastery based on predetermined standards, not as a competitive tool among students (Pangastuti & Munfaati, 2018). However, social facts on the ground indicate a significant gap between the idealism of PAP and its evaluation practices. Most madrasahs are still trapped in the dominance of Norm-Referenced Assessment (PAN), which reduces competency achievement to mere class ranking, thereby losing its formative function (Azijah et al., 2026). Academic problems that arise are when teachers experience

difficulty translating national competency standards into contextual and meaningful PAP practices (Susilowati et al., 2024). As a result, PAP is often reduced to a summative administrative burden that is unable to provide pedagogical feedback for learning improvement (Kiswanto et al., 2024). This gap between the evaluation philosophy centered on competency mastery and the reality of ranking-oriented practice creates an urgency to explore how the PAP is actually operationalized within the unique madrasah ecosystem.

Previous research on Arabic language learning evaluation has generally focused on three approaches that have not addressed the dynamics of the PAP holistically. Amal's (2024) study focused more on the development and statistical testing of PAP instruments in public schools, without exploring the process of standards negotiation at the teacher level. Furthermore, Laati & Pangerang's (2026) study described the implementation of PAP in a normative-descriptive manner but failed to analyze how evaluation results were interpreted into pedagogical decisions. Meanwhile, Nurhijatina et al.'s (2023) study measured the effectiveness of PAP-based learning outcomes but positioned evaluation as a separate entity from the continuous learning improvement cycle. These three approaches leave a significant gap in the literature, particularly in the context of madrasahs with their unique characteristics. The novelty of this research lies in its in-depth qualitative exploration of the entire PAP dynamic, from the contextual construction of standards, through the qualitative matrix-based interpretation of the four maharah skills, to its transformation into the heart of the continuous improvement cycle. This approach offers a new perspective that goes beyond quantitative measurement to understanding the vibrant evaluation ecosystem.

Based on this literature gap, this study aims to explore in depth three fundamental aspects: how competency standards are constructed and translated into PAP practices, how teachers interpret student competency achievement, and how PAP results are used as a basis for decision-making and learning follow-up. MTs At-Taqwa Jatingarang was chosen as the research location because this madrasah exhibits distinctive and progressive PAP practices, where teachers actively negotiate the demands of the national curriculum (KMA 183 of 2019) with the local Islamic context, and has a structured and collaborative follow-up mechanism. The urgency of this research lies in its multidimensional contribution: theoretically, this study enriches the literature on Islamic education evaluation by offering a participatory and reflective conceptual model of PAP; practically, this research provides operational guidance for madrasah teachers in transforming PAP from an administrative instrument into an empowering diagnostic tool; and policy-wise, the findings of this study can serve as a strategic reference for the Ministry of Religious Affairs in formulating guidelines for quality assurance of Arabic language learning evaluation at the madrasah level nationally.

2. Literatur Review

Arabic language learning in madrasahs plays a strategic role in holistically shaping students' communicative competence and Islamic literacy. Learning evaluation, particularly in the four essential skills (listening, speaking, reading, and writing), requires an approach that not only measures final results but also facilitates a process of continuous improvement (Jannah et al., 2026). However, various studies show that evaluation practices in madrasahs are still dominated by competitive Norm-Referenced Assessment (NPA), which reduces competency achievement to mere rankings and positions evaluation as a summative judgment tool rather than a formative instrument (Syafuruddin & Herdah, 2025). This condition causes students to experience evaluative anxiety and teachers lose the momentum to make targeted learning improvements. According to the Criterion-Referenced Assessment theory, initiated by Glaser (1963) and developed by Hambleton (1985), evaluation will be more meaningful when the interpretation of results is based on predetermined absolute competency benchmarks, rather than through relative comparisons between students. Teachers need to adopt Benchmark-Referenced Assessment (BRA), which can map specific competency mastery and provide diagnostic feedback.

Benchmark-Referenced Assessment (BRA) is an evaluation strategy that measures students' mastery of a set of explicitly defined competencies or learning indicators. This model is designed to create an objective, transparent, and mastery-learning-oriented evaluation system (Bloom, 1968). In its implementation, teachers establish clear success criteria and use instruments aligned with learning objectives, so students know exactly what is expected of them. In Arabic language learning, BRA is considered highly effective because the four maharah skills have measurable and hierarchical competency standards (Syafuruddin & Herdah, 2025). According to Richards and Rodgers (2001), authentic language assessment must reflect language use in real contexts, not simply mastery of grammatical structures. Through the PAP, teachers can construct assessment rubrics that comprehensively map student achievement across each dimension of competence, so that evaluation is no longer reduced to a single number but rather to a comprehensive and meaningful competency profile.

Competency achievement is a fundamental indicator for determining the success of the learning process (Amanda & Nadzjia, 2026). According to Black and Wiliam (2011), competency achievement in the Assessment for Learning paradigm encompasses not only the cognitive domain but also the affective and psychomotor domains, which are integrated with each other. In Arabic language learning, these three domains are closely interconnected; students are not only required to understand the rules of grammar and grammar theoretically but also to demonstrate them communicatively in their listening, speaking, reading, and writing skills (Wisnu et al., 2025). The poor quality of evaluation result interpretation is often caused by inaccurate assessment instruments and approaches, resulting in the resulting feedback being unable to guide students toward specific improvements (Awwawina et al., 2026). The PAP presents a solution to this problem because it provides an interpretive framework based on the gap between actual achievement and the targeted competency benchmarks (Nurmanto & Ashari, 2023). Through qualitative matrix analysis based on the Maharah, teachers can precisely identify which skills students are experiencing difficulty in, allowing for differentiated and targeted learning follow-up.

Various previous studies have examined the implementation of evaluation in Arabic language learning, but most have focused on quantitative and descriptive-normative approaches. Alfath's (2019) study, for example, primarily measured the effectiveness of the PAP instrument through statistical tests in public schools, without exploring the dynamics of standard construction at the teacher level. On the other hand, Kiswanto's (2024) study described the implementation of PAP normatively without analyzing how evaluation results are transformed into pedagogical decisions. Meanwhile, research by Nurfauzan (2026) and Nugrahawan (2025) discussed Arabic language evaluation as a separate entity from the continuous learning improvement cycle. This literature gap indicates that there has been no in-depth study exploring how PAP is qualitatively operationalized in the madrasah context—from negotiating national curriculum standards with local contexts, through qualitative matrix-based interpretation, to its transformation into a continuous improvement cycle. Based on these studies, this study aims to fill this gap by offering a new perspective on the holistic and contextual dynamics of PAP.

In addition to an appropriate evaluation approach, the success of PAP implementation is also heavily influenced by contextual factors, teacher pedagogical competence, and institutional policies. Vygotsky and Cole's (2018) social constructivism theory emphasizes that effective evaluation must consider students' zones of proximal development, with teachers acting as facilitators who provide scaffolding based on assessment results. In practice in madrasahs, teachers often face challenges in the form of the rigidity of national curriculum documents that do not always align with the characteristics and readiness of students in the field (Widyasari, 2025). This condition can cause competency standards to be too rigid and unresponsive to diverse learning needs. Therefore, pedagogical flexibility and madrasah policy support are needed that allow teachers to adapt national standards to local contexts without reducing academic rigor. PAP implemented with a participatory and reflective approach will be able to create a humanistic evaluation atmosphere, reduce students' evaluative anxiety, and ultimately improve the quality of Arabic language competency achievement overall.

3. Methods

This study uses a qualitative approach with a case study design to explore in depth the dynamics of Benchmark Assessment (BAP) in evaluating Arabic language learning for seventh-grade students at MTs At-Taqwa Jatingarang. The case study design was chosen because the research focuses on understanding a contemporary phenomenon in a natural and specific context, by considering the relationship between competency standard setting, assessment practices, interpretation of student achievement, and utilization of evaluation results in the learning process (Yin, 2018). Research data were collected through semi-structured interviews, observations, and document analysis. Interviews were used to explore informants' experiences and meanings regarding the implementation of BAP, interpretation of competency achievement, and follow-up of evaluation results; observations were conducted to understand evaluation and learning practices in their natural context; while document analysis was used to examine learning tools, assessment instruments, achievement criteria or standards, and relevant evaluation documents. The use of various data collection techniques enabled researchers to gain a more comprehensive understanding of the phenomenon being studied while strengthening the credibility of the findings through triangulation of techniques and data sources (Creswell & Poth, 2018). Informants were selected purposively based on their direct involvement and the relevance of their knowledge to the research phenomenon, while considering the principle of information power in qualitative research (Malterud et al., 2016).

The research informants consisted of one Arabic language teacher, six seventh-grade students, three parents/guardians, and one vice principal for curriculum, as presented in Table 1. The Arabic language teacher was positioned as a key informant because of her direct involvement in the planning, implementation, interpretation, and follow-up of the PAP. Six students were selected to represent a variety of experiences and learning outcomes, while three guardians were involved to obtain family perspectives on understanding and responding to student evaluation results. The vice principal for curriculum provided an institutional perspective on the curriculum policies and evaluation systems implemented at the madrasah. Data were analyzed thematically through data organization, familiarization, initial coding, category development, theme identification and review, and interpretations that addressed the research focus (Braun & Clarke, 2021). Analysis was conducted iteratively by comparing data from interviews, observations, and documents to identify patterns, similarities, differences, and relationships between themes. The analysis process also involved stages of data condensation, data presentation, and continuous drawing and verification of conclusions (Miles et al., 2014). Data validity was strengthened through triangulation of sources and techniques, member checking of important information from informants, and analytical discussions to ensure that the researcher's interpretations remained rooted in the empirical data and accurately represented the participants' perspectives (Lincoln & Guba, 1985).

Tabel 1. Informan Penelitian

No	Informant Code	Number	Selection Criteria and Relevance of Information
1	GBA	1	Arabic language teacher who teaches Grade VII and is directly involved in the planning, implementation, interpretation, and follow-up of Criterion-Referenced Assessment (CRA/PAP).
2	SA 1–6	6	Grade VII students who participate in Arabic language learning and CRA-based assessment, selected to represent variations in learning achievement and diverse learning experiences.
3	WMD	3	Parents or guardians of participating students, included to provide perspectives on the acceptance, understanding, communication, and follow-up of students' assessment results within the family

			environment.
4	WKM	1	Vice Principal for Curriculum, who has knowledge of curriculum policies, assessment systems, competency achievement standards, and assessment practices at the madrasah level.

4. Results and Discussion

The Construction of Competency Standards in Arabic Language Learning Benchmark Assessment Practices

The construction of Arabic language competency standards at MTs At-Taqwa Jatingarang is rooted in the integration of the national curriculum and madrasah characteristics. A teacher formulated the standards based on official documents, as she explained: "We refer to KMA 183 of 2019 as the main basis for determining competencies" (GBA, W1). This process did not stop at the document; it was further developed into measurable indicators. This was confirmed through an analysis of the syllabus document, which demonstrated a specific and detailed mapping of maharah skills. The teacher further emphasized, "Each indicator must reflect the minimum achievement expected of seventh-grade students" (GBA, W2). Learning tool documentation revealed a clearly structured assessment rubric. In practice, these standards were adapted to the local context, with a teacher stating, "We added vocabulary for the madrasah environment to make it more contextual" (GBA, W3). The competency standards serve not merely as administrative requirements, but as a measurable and relevant operational foundation for comprehensive student evaluation at this madrasah each academic semester.

Operationalizing these standards into Benchmark-Referenced Assessment (BRA) instruments requires adjustments to students' cognitive and affective well-being. Teachers design questions and assignments based on clear benchmarks, explaining, "The BRA instrument is designed to measure mastery of specific material, not to compare students across students" (GBA, W4). This approach ensures the objectivity of the evaluation. Observations in a seventh-grade classroom indicate that teachers consistently use the speaking and writing skills checklists during the learning process, which are then summarized directly on the BRA sheet. This consistency is crucial for maintaining construct validity. Furthermore, teachers acknowledge challenges in developing assessment weights, stating, "Determining fair weighting between listening and speaking skills sometimes requires lengthy discussions with colleagues" (GBA, W5). Nevertheless, this collaboration produces a comprehensive, tested BRA instrument that fully represents competency standards and provides a highly accurate picture of individual student achievement in a comprehensive, structured, and sustainable manner. The dynamics of translating standards into PAP practices were also influenced by the responses and characteristics of seventh-grade students. A teacher recognized that rigid standards needed to be softened without diminishing the essence of competencies, stating, "Seventh-grade students are still in transition from elementary school, so the assessment approach must be more persuasive" (GBA, W6). This flexibility was key to the success of the evaluation. Students also shared their perspectives on the established standards. One student stated, "The PAP questions felt challenging, but we knew the minimum standards that had to be achieved" (SA 1, W7). This understanding demonstrated good transparency in the standards. Another student added, "The teacher always explained the assessment criteria before the assignments were given, so we weren't surprised" (SA 2, W8). A teacher concluded this dynamic emphatically, "The key to PAP in seventh-grade is clear communication of expectations between teachers and students from the beginning of the lesson" (GBA, W9). This demonstrates that participatory standard construction can create a conducive and meaningful evaluation ecosystem.

Alignment between learning objectives, materials, instruments, and assessment standards is a key parameter for the success of the PAP construction. The Deputy Principal emphasized the importance of internal audits of these tools, stating, "We review the PAP syllabus and instruments at the beginning of each

semester to ensure alignment" (WKM, W10). This preventative measure ensures content validity. Teachers also make periodic adjustments to the materials, explaining, "If any material is too difficult, we revise the indicators without lowering the basic competency standards" (GBA, W11). This adaptation demonstrates the teachers' professionalism in maintaining quality. Students feel the impact of this alignment, with one stating, "The material taught always aligns with what is tested in the PAP" (SA 3, W12). This consistency builds students' confidence in facing evaluations. The harmonious integration of all evaluation components directly ensures that the Benchmark-Referenced Assessment (BEA) serves as an objective, fair measuring tool, accurately reflecting student competency achievement and is academically accountable.

Interpreting Student Competency Achievement Based on Benchmark-Referenced Assessment

Interpreting Benchmark-Referenced Assessment (PAP) results at MTs At-Taqwa Jatingarang goes beyond understanding the meaning behind student achievement. A teacher unequivocally explains this approach: "I don't just look at the final score, but analyze error patterns to understand where students' difficulties lie" (GBA, W13). This analytical approach allows teachers to identify specific root causes. The Vice Principal also emphasized the importance of comprehensive interpretation, stating, "PAP results should be read as a reflection of the learning process, not simply as a label of success or failure" (WKM, W14). This institutional perspective reinforces that evaluation aims for improvement, not judgment. Furthermore, the teacher added, "Each score tells a story about the efforts and obstacles experienced by students during the learning process" (GBA, W15). This interpretation of the research findings transformed the PAP from a measurement tool into a diagnostic instrument rich in pedagogical information.

The interpretation process continued with a systematic and transparent categorization of competency achievement. The teacher grouped students into four categories based on established benchmarks, explaining, "Students achieving 85-100% are categorized as excellent, 70-84% as good, 60-69% as adequate, and below 60% need remedial work" (GBA, W16). This categorization facilitated the identification of learning needs. Observations showed that teachers consistently used the standardized rubric to ensure objectivity in the grouping. Students also understood this categorization system, with one stating, "We know where we stand and what needs to be improved to achieve a better category" (SA 1, W17). This transparency motivated students to improve their achievement. Parents also responded positively to the system, with one stating, "The achievement categories help us understand our child's condition and how to support them at home" (WMD 1, W18). A teacher corroborated this finding, stating, "Clear categorization eliminates ambiguity and provides specific direction for students, parents, and teachers themselves" (GBA, W19). The system creates a participatory and improvement-oriented evaluation ecosystem.

Identifying competency gaps was a primary focus in the PAP interpretation. The teacher analyzed the gap pattern across the skills of reading comprehension, revealing, "The majority of students mastered the skills of reading comprehension (maharah istima') and reading comprehension (qira'ah), but experienced significant difficulties in the skills of reading comprehension (kalam) and reading comprehension (kitabah)" (GBA, W20). This finding is consistent with observations showing low frequency of speaking and writing practice in class. Documentation of the PAP results confirmed this pattern, with 68% of students scoring below the benchmark for productive skills. Students also acknowledged this challenge, with one stating, "I find it easier to understand vocabulary than to pronounce it in complete sentences" (SA 2, W21). This gap indicates a need for strengthening productive skills. The teacher responded reflectively to this finding, saying, "This gap signals that learning methods need to involve more direct practice" (GBA, W22). Thus, systematic gap identification provides a clear roadmap for targeted learning interventions.

Factors influencing competency achievement are multidimensional, encompassing both internal and external aspects that interact. Teachers identified internal factors as the primary determinants, explaining, "Learning motivation and home practice habits significantly influence the speed of competency mastery" (GBA, W23). External factors also play a significant role, particularly the supportive learning environment.

The Deputy Principal added an institutional perspective, stating, "The availability of a language laboratory and the intensity of Arabic language exposure in the madrasah environment also determine student achievement" (WKM, W24). The dynamic interaction between these factors results in variations in achievement that cannot be reduced to mere numbers but must be read as a complete competency profile. To capture this complexity, teachers developed a qualitative interpretation matrix that maps student mastery of each Maharah skill, as presented systematically in Table 2 below.

Table 2. Interpretation of Arabic Competency Achievement Based on Maharah Skills

Achievement Category	Mahārah Istimā' (Listening)	Mahārah Kalām (Speaking)	Mahārah Qirā'ah (Reading)	Mahārah Kitābah (Writing)	Pedagogical Implications
Excellent	Able to understand complex oral instructions without repetition	Able to respond spontaneously using accurate <i>nahwu-sharaf</i> structures	Able to read vowelised texts with fluent <i>tartīl</i> and high accuracy	Able to produce coherent, well-structured paragraphs	Students have exceeded the competency standards and are ready for cross- <i>mahārah</i> enrichment activities
Good	Able to understand contextual conversations with only minor errors	Able to speak fluently with minimal vocabulary support	Able to read vowelised texts with accurate pronunciation and good fluency	Able to write compound sentences with relatively accurate structures	Students have achieved the competency standards but require further reinforcement in specific areas
Fair	Able to understand individual vocabulary items but experiences difficulty comprehending complete sentences	Able to respond using fragmented sentences with frequent pauses	Able to read vowelised texts with teacher assistance or guidance	Able to write simple sentences but demonstrates varying errors in spelling and grammatical structure	Students demonstrate stronger receptive than productive language skills and require instructional scaffolding
Requiring Remedial Support	Experiences difficulty distinguishing basic Arabic phonemes	Unable to independently formulate and deliver simple oral responses	Reads texts by gradually sounding out individual letters	Able to copy texts but unable to independently construct sentences	Students require intensive intervention, differentiated instruction, and targeted remedial programmes

Based on Table 2, the interpretation of competency achievement is no longer reduced to a numerical score, but rather mapped holistically through the mastery profile of the four maharah skills. This matrix allows teachers to read student achievement from a perspective and contextual perspective. A teacher emphasized the advantages of this approach, saying, "This table helps me see specifically where students need intervention, rather than just giving a final grade" (GBA, W25). Students also felt the benefits of this transparency, with one stating, "I now know that my weakness lies in speaking, not in all Arabic lessons" (SA 3, W26). Parents also supported this qualitative approach, with one stating, "Skills-based explanations make it easier for us to support our children in learning at home" (WMD 2, W27). This qualitative matrix-based interpretation transforms the PAP from a measuring tool into a diagnostic map rich in pedagogical meaning, accountable, and oriented towards improving differentiated learning.

Utilizing Benchmark-Referenced Assessment Results for Decision-Making and Learning Follow-Up

The utilization of Benchmark-Referenced Assessment (PAP) results at MTs At-Taqwa Jatingarang has transformed into an essential foundation for differentiated pedagogical decision-making. A teacher emphasized, "PAP results are not just numbers on a report card, but rather a database for determining who needs remedial work and who deserves enrichment" (GBA, W28). This follow-up is carried out systematically, not incidentally. The Vice Principal reinforced institutional commitment, saying, "We require every teacher to develop a follow-up program based on the analysis of PAP results before starting new material" (WKM, W29). Remedial programs are specifically designed to close competency gaps, while enrichment programs are aimed at broadening students' horizons. Students felt the direct impact of this policy, with one stating, "After the PAP scores came out, I took extra classes to improve my speaking skills" (SA 4, W30). The evaluation results directly translated data into concrete, personalized and targeted actions for each student.

In addition to specific programs, PAP results also triggered dynamic adaptations of classroom learning strategies. A teacher conducted in-depth reflection on each assessment cycle, stating, "The PAP findings showed that the lecture method was less effective for nahwu (literacy), so I switched to game-based learning" (GBA, W31). This change in strategy had a significant impact on student engagement. Students acknowledged a shift in the learning environment, stating, "The teacher's teaching style has become more fun and we understand better" (SA 5, W32). Positive responses were also evident in changes in learning habits at home. One parent observed, "At home, my children practice vocabulary more often because the teacher's methods have changed" (WMD 3, W33). These adaptations demonstrate that PAP serves as reflective feedback, forcing teachers to continually update their methodologies for greater and more meaningful learning effectiveness.

The collaborative dimension in utilizing PAP results is strengthened through intensive communication and affective mentoring. Teachers do not work alone, she explained, "We invite parents to discuss strategies for supporting their children based on PAP results, not just to give grades" (GBA, W34). This humanistic approach builds trust. Parents feel meaningfully involved: "Communication with teachers is very intense, we know exactly what we need to support at home" (WMD 1, W35). Students also feel the teacher's presence beyond the classroom: "Teachers always ask how I am and what difficulties I have after the exam" (SA 6, W36). The Deputy Principal added, "We designed our reporting system to be transparent and open up dialogue between the school and families" (WKM, W37). This synergy creates a strong network of support, ensuring that follow-up learning does not stop at the madrasah gates, but continues in an integrated manner throughout the home environment.

Figure 1. Model of the Cycle of Utilizing PAP Results for Continuous Learning Improvement



At the macro level, PAP results are integrated into long-term learning planning and curriculum evaluation. A teacher explained, "PAP data is archived as a roadmap for developing learning designs for the next semester to ensure they are more targeted" (GBA, W38). This integration ensures sustainable quality. The Vice Principal emphasized, "Evaluation results serve as vital input in teacher council meetings to review the effectiveness of the current curriculum" (WKM, W39). To comprehensively visualize this transformational flow, this study proposes a conceptual model depicting the closed-loop utilization of PAP as a living system, as schematically presented in Figure 1 below.

Figure 1 illustrates that the Benchmark Assessment (PAP) is part of a continuous improvement cycle that holistically connects assessment, interpretation, decision-making, and implementation. This model emphasizes that evaluation is the starting point for improvement, not the end of the learning process. A teacher summarized this philosophy by saying, "PAP is a mirror for teachers to reflect on, not a verdict for students" (GBA, W40). Students also experienced a more holistic sense of learning: "I feel supported to develop, not just judged by grades" (SA 1, W41). Benchmark-Referenced Assessment (BEA) at MTs At-Taqlwa Jatingarang has successfully metamorphosed from an administrative instrument into the heart of learning quality improvement that is responsive, accountable, and centered on the needs of students holistically.

Discussion

This study reveals three fundamental findings regarding the dynamics of Benchmark-Referenced Assessment (BEA) in Arabic language learning at MTs At-Taqlwa Jatingarang. First, the construction of competency standards is not mechanistic, but rather the result of creative negotiations between the national curriculum document (KMA 183 of 2019) and the local madrasah context. Teachers actively adapt assessment indicators by adding vocabulary to the madrasah environment, ensuring that competency standards are contextually relevant without sacrificing academic rigor (Azizi et al., 2026). Second, the interpretation of competency achievement is conducted through a qualitative matrix based on the four maharah skills (*istimā'*, *kalām*, *qirā'ah*, *kitābah*), rather than through numerical reduction. This approach allows teachers to holistically map students' competency profiles and identify specific gaps in productive skills. Third, the PAP transforms from an administrative instrument into the heart of a continuous improvement cycle, where evaluation results directly translate data into differentiated pedagogical actions—from remedial and enrichment programs to adapted learning strategies and collaboration with parents. These three findings collectively demonstrate that an effective PAP is not simply a measuring tool, but a vibrant, participatory, and improvement-oriented evaluation ecosystem.

The findings of this study align with and enrich the theoretical framework of Criterion-Referenced Assessment, initiated by Glaser (1963) and further developed by Hambleton (1985), which emphasizes that the interpretation of evaluation results must refer to established competency benchmarks, not to comparisons between students. In the context of Arabic, these findings also overlap with the Communicative Language Teaching theory (Richards & Rodgers, 2014), which emphasizes the importance of authentic assessment of communicative skills, not merely mastery of grammatical structures. Furthermore, this study confirms the principles of Assessment for Learning (Black & Wiliam, 2011), which positions evaluation as formative feedback for learning improvement, not merely a summative tool. From an Islamic education perspective, these findings also align with the evaluation philosophy in the *tarbiyah* tradition, which emphasizes a humanistic and holistic approach, where assessment is not intended to judge, but rather to guide students towards competency perfection (Al-Attas, 1979; Al-Ghazali, 2010). This integration between modern evaluation theory and Islamic educational values results in PAP practices that are not only academically valid, but also spiritually and pedagogically meaningful in the madrasah context.

The novelty of this research lies in three aspects that significantly differentiate it from previous studies. First, while Alfath's (2019) study focused on the PAP in public schools using a quantitative

approach, this study explores the PAP in the madrasah context through an in-depth qualitative lens, revealing the dynamics of negotiating standards between the national curriculum and the local Islamic context. Second, unlike Kiswanto's (2024) study, which only described the normative implementation of the PAP, this study reveals how teachers actively construct a qualitative interpretation matrix based on the four maharah—an innovation undocumented in previous Arabic language evaluation literature. Third, while Nurfauzan's (2026) and Nugrahawan's (2025) studies discussed the PAP as an evaluation tool separate from the learning cycle, this study proposes a closed-loop conceptual model that integrates the PAP into a continuous improvement cycle. These three novel aspects collectively fill a gap in the literature regarding how the PAP can be transformed from an administrative instrument into the heart of learning quality improvement in the context of Islamic education, while also offering a new analytical framework that can be replicated in other madrasah settings. The implications of this research's findings are multidimensional, encompassing theoretical, practical, and methodological domains. Theoretically, this study enriches the literature on Islamic education evaluation by offering a new conceptual framework for the integration of Criterion-Referenced Assessment (CRA) and tarbiyah values, a topic rarely explored empirically. The findings also contribute to Assessment for Learning theory by demonstrating how formative evaluation principles can be operationalized in the context of Arabic language learning in madrasas. Practically, this study offers an operational model of Criterion-Referenced Assessment (PAP) that can be adopted by Arabic language teachers in other madrasas, particularly in terms of developing a qualitative interpretation matrix based on maharah (religious value) and designing a continuous learning improvement cycle. This model provides concrete guidance for teachers in transforming Criterion-Referenced Assessment (PAP) from an administrative burden into a meaningful diagnostic tool. Methodologically, this study demonstrates the power of a qualitative case study approach in exploring complex evaluation dynamics, often beyond the reach of quantitative approaches (Hikmah et al., 2026). The use of triangulation of data sources (teachers, students, parents, and madrasah leaders) also provides a best practice example of how the credibility of qualitative findings can be strengthened through a multi-stakeholder perspective.

Based on these findings and implications, several policy recommendations need to be formulated to strengthen the implementation of the PAP at the madrasah and national levels. At the madrasah level, leaders need to develop clear Standard Operating Procedures (SOPs) for the PAP, covering the stages of standard construction, interpretation of results, and follow-up to learning. Madrasahs also need to allocate time and resources for regular teacher training, particularly in developing authentic PAP instruments and analyzing evaluation results based on qualitative matrices. At the Ministry of Religious Affairs level, national guidelines for the implementation of the PAP in madrasahs should be developed that integrate national competency standards (KMA 183) with contextual flexibility. These guidelines should be accompanied by examples of best practices from madrasahs that have successfully implemented the PAP effectively. At the national level, the government needs to consider integrating the PAP into the madrasah education quality assurance system, where PAP results are used not only for student evaluation but also as input for curriculum reviews and teacher professional development. Furthermore, a digital platform should be established to share PAP best practices among madrasahs, allowing for wider dissemination of local innovations. These policies will ensure that the PAP is not merely an evaluation instrument but also a driving force for systemic and sustainable improvement in the quality of Arabic language learning in madrasahs.

5. Conclusion

This study concludes that the implementation of Benchmark Assessment (PAP) in Arabic language learning at MTs At-Taqwa Jatingarang has transcended conventional administrative functions into a vibrant, participatory evaluation ecosystem oriented toward continuous improvement. Key findings indicate that the construction of competency standards is not mechanistic, but rather the result of a creative negotiation between the demands of the national curriculum and the local madrasah context. Interpretation of

competency achievement is conducted comprehensively through a qualitative matrix based on the four Maharah skills, allowing for precise identification of specific gaps without numerical reduction. PAP results are transformed into the heart of the learning improvement cycle, where evaluation data is directly translated into differentiated pedagogical actions, such as remedial programs, enrichment programs, adaptation of instructional strategies, and intensive collaboration with parents. The synergy between contextual standards, diagnostic interpretation, and strategic follow-up demonstrates that effective PAP functions as a reflective instrument that empowers teachers and students, while ensuring accountability, fairness, and relevance for comprehensive Arabic language learning at the madrasah level. This innovative evaluation model fundamentally shifts the assessment paradigm from merely a summative measurement tool to a formative and empowering diagnostic instrument.

The implications of this study emphasize the need to integrate locally context-based formative evaluation into madrasah curriculum policies. However, the study's limitations lie in its focus on a single madrasah with a limited sample, requiring caution in generalizing the findings. Further research is recommended to comparatively test this PAP improvement cycle model across various madrasah settings using a mixed-methods approach, as well as to explore the integration of digital technology into qualitative interpretation matrices to strengthen the accuracy and efficiency of Arabic language learning evaluations more broadly. This will enrich the literature on Islamic education evaluation in a more comprehensive and applicable manner. Future studies should also incorporate the perspectives of madrasah supervisors.

6. References

- Al-Attas, S. M. N. (1979). *Preliminary Thoughts on the Nature of Knowledge and the Definition and Aims of Education* (pp. 19-47).
- Albab, U., Ta'rifin, A., Novianti, D., & Safitri, H. H. (2025). Exploring the Impact of Augmented Reality on Meaningful Learning in Islamic Religious Education: A Quantitative Analysis. *EDUKASI: Jurnal Penelitian Pendidikan Agama dan Keagamaan*, 23(1), 1–25.
- Alfath, K. (2019). Teknik pengolahan hasil asesmen: teknik pengolahan dengan menggunakan pendekatan acuan norma (PAN) dan pendekatan acuan patokan (PAP). *AL-MANAR: Jurnal Komunikasi dan Pendidikan Islam*, 8(1), 1-28.
- Al-Ghazali, I. (2010). *The Duties of Brotherhood in Islam*. Kube Publishing Ltd.
- Amal, I. (2024). Integrasi Pendekatan Acuan Patokan (PAP) Dan Pendekatan Acuan Normatif (PAN) Dalam Konteks Penilaian Pembelajaran Sekolah. *Ideguru: Jurnal Karya Ilmiah Guru*, 9(3), 1373-1380.
- Amanda, D. A. S., & Ndizija, A. N. (2026). Implementation of Problem-Based Learning in Enhancing Student Engagement and Learning Achievement in Islamic Religious Education for Children. *Research Journal on Teacher Professional Development*, 4(01), 19–38. <https://doi.org/10.21580/rjtpd.v4i01.24904>
- Awwawina, R. U., Waluya, B., & Kharisudin, I. (2026). Implementing Differentiated Learning to Enhance Mathematical Reasoning Readiness for University Entrance Assessment: A Classroom Action Research in Madrasah Aliyah. *Research Journal on Teacher Professional Development*, 4(01), 183–197. <https://doi.org/10.21580/rjtpd.v4i01.33576>
- Azijah, I. N., Nurkholifah, N. I., Rahma, S. N., Putri, M. M., & Udin, T. (2026). Analisis Butir Soal dan Ketuntasan Belajar Ditinjau dari Penilaian Acuan Norma dan Penilaian Acuan Patokan. *AL IBTIDA': Jurnal Program Studi Pendidikan Guru Madrasah Ibtidaiyah*, 14(1), 1-10.
- Azizi, N. D., Albab, A. W., Musrikah, & Chotimah, C. (2026). Contextual Teaching and Learning (CTL) Model to Improve Reading and Writing Literacy Skills of Elementary School Students. *Social Journal of Studies in Education*, 2(01), 31–46.
- Bloom, B. S. (1968). Learning for mastery. Instruction and curriculum. Regional education laboratory for the carolinas and virginia, topical papers and reprints, number 1. *Evaluation comment*, 1(2), n2.

- Braun, V., & Clarke, V. (2021). One size fits all? What counts as quality practice in (reflexive) thematic analysis? *Qualitative Research in Psychology*, 18(3), 328–352. <https://doi.org/10.1080/14780887.2020.1769238>
- Creswell, J. W., & Poth, C. N. (2018). *Qualitative inquiry and research design: Choosing among five approaches* (4th ed.). SAGE Publications.
- Hambleton, R. K. (1985). Criterion-referenced assessment of individual differences. In *Methodological and statistical advances in the study of individual differences* (pp. 393-424). Boston, MA: Springer US.
- Jannah, V. M., Salamah, U., & Ridlo, U. (2026). Pengembangan instrumen evaluasi maharah kitabah. *Jejak Digital: Jurnal Ilmiah Multidisiplin*, 2(1), 501-508.
- Kiswanto, D., Arista, D., Fitrah, I. J., Annisa, M. N., & Qomari, N. (2024). Implementasi penilaian acuan norma (PAN) dan penilaian acuan patokan (PAP) dalam pengolahan hasil belajar siswa. *Risalah Jurnal Pendidikan dan Studi Islam*, 10(3), 1207-1219.
- Laata, R., & Pangerang, R. U. L. (2026). Studi Komparasi Penilaian Acuan Patokan (PAP) dan Penilaian Acuan Norma (PAN) dalam Evaluasi. *Jejak digital: Jurnal Ilmiah Multidisiplin*, 2(4), 6436-6445.
- Lincoln, Y. S., & Guba, E. G. (1985). *Naturalistic inquiry*. SAGE Publications.
- Malterud, K., Siersma, V. D., & Guassora, A. D. (2016). Sample size in qualitative interview studies: Guided by information power. *Qualitative Health Research*, 26(13), 1753–1760. <https://doi.org/10.1177/1049732315617444>
- Miles, M. B., Huberman, A. M., & Saldaña, J. (2014). *Qualitative data analysis: A methods sourcebook* (3rd ed.). SAGE Publications.
- Nugrahawan, A. R. (2025). *Mengenai Evaluasi dalam Pembelajaran Bahasa Arab*. PT Bukuloka Literasi Bangsa.
- Nurfauzan, M. F., Al Iqram, M. T., Maizar Tri Al Iqram, R., & Ridlo, U. (2026). Evaluasi Pembelajaran Bahasa Arab sebagai Bahasa Asing: Konsep, Tantangan, dan Strategi. *Jejak digital: Jurnal Ilmiah Multidisiplin*, 2(1), 564-573.
- Nurhijatina, H., Larasati, N. J., & Bella, S. (2023). Analisis Classrooms Assessment: Remedial, Pengayaan, Pendekatan Acuan Patokan (PAP) dan Pendekatan Acuan Normatif (PAN). *Didaktik: Jurnal Ilmiah PGSD STKIP Subang*, 9(5), 3216-3230.
- Nurmanto, Y., & Ashari, M. Y. (2023). Penilaian Acuan Patokan Dalam Evaluasi Pendidikan Islam. *Ats-Tsaqofi: Jurnal Pendidikan dan Manajemen Islam*, 5(1), 39-51.
- Pangastuti, R., & MunfaAti, K. (2018). Penilaian Acuan Norma, Penilaian Acuan Patokan, Kriteria Ketuntasan Minimal Di Madrasah Ibtidaiyah an-Nur Plus Junwangi Krian Sidorajo Jawa Timur. *Tarbiyah Al-Awlad: Jurnal Kependidikan Islam Tingkat Dasar*, 8(2), 202-217.
- Raharjo, W., Rokhani, R., & Toha, A. M. (2025). A Hermeneutic Study of the Ethical Use of Technology in Arabic Language Education: A Perspective of Qur'anic Verses. *Journal of Islamic Education Thought and Development*, 1(1), 31–47. <https://doi.org/10.67046/jietd.v1i1.2>
- Richards, J. C., & Rodgers, T. S. (2001). *Approaches and methods in language teaching*. Cambridge university press.
- Susilowati, E., Shaleh, S., & ZulfatulKhoiriyah, Z. (2024). Analisis Analisis Langkah-Langkah Pendekatan Acuan Patokan (PAP) Asesmen dalam Pendidikan Islam. *Moderasi: Journal of Islamic Studies*, 4(2), 194-206.
- Syafuruddin, L., & Herdah, H. (2025). Interpretasi Hasil Tes Penilaian Acuan Norma (PAN) dan Penilaian Acuan Kriteria (PAK) dalam Evaluasi Pembelajaran Bahasa Arab. *Ta'limi| Journal of Arabic Education and Arabic Studies*, 4(2), 245-266.
- Vygotsky, L., & Cole, M. (2018). Lev Vygotsky: Learning and social constructivism. *Learning Theories for Early Years Practice*. UK: SAGE Publications Inc, 68-73.
- Widyasari, M. P. (2025). *Dilema Guru: Antara Kompetensi dan Tuntunan Kurikulum*. Indonesia Emas Group.

- Wijaya, B. R., Suja'i, S., & Munawaroh, S. (2025). Traditional Ethics in the Use of AI for Arabic Language Education: An Aristotelian Philosophy Perspective. *Journal of Islamic Education Thought and Development*, 1(1), 13–30. <https://doi.org/10.67046/jietd.v1i1.4>
- Wiliam, D. (2011). What is assessment for learning?. *Studies in educational evaluation*, 37(1), 3-14.
- Yin, R. K. (2018). *Case study research and applications: Design and methods* (6th ed.). SAGE Publications.