

Measurement, Assessment, and Evaluation in Arabic Language Learning: Distinctions, Principles, and Purposes

Muhammad Maulana Arif¹, Herdah²

¹²Institut Agama Islam Negeri Parepare

E-mail: mthmaulana@gmail.com | herdah@iainpare.ac.id

Tel: 085242325148 | 085299227995

Naskah Diterima Tanggal 15/06/2026; Direvisi Akhir Tanggal 19/06/2026; Disetujui Tanggal 21/06/2026

doi: 10.37216/lughatuad-dhat.v7i1.3397

Abstract

Measurement, assessment, and evaluation are three foundational concepts in education that are often used interchangeably in Arabic language learning (ALL). Integrative studies specifically analyzing the differences among these three concepts along with their principles and objectives in the context of ALL in Indonesia remain limited, which constitutes a research gap addressed in this study. This article aims to analyze the fundamental differences among them, identify the principles underpinning their application, and examine their specific objectives in ALL. Employing a library research method with a conceptual analysis approach, this study reveals that measurement is quantitative-descriptive, assessment is interpretive-pedagogical, and evaluation is holistic-programmatic. Their application in ALL should be guided by the principles of validity, authenticity, language skills integration, and continuity. Their objectives encompass diagnosing Arabic language proficiency, monitoring learning progress, and comprehensively enhancing ALL program quality.

Keywords: Arabic language learning, Measurement, Assessment, Evaluation, Learning outcomes

Abstrak

Pengukuran, penilaian, dan evaluasi adalah tiga konsep fundamental dalam pendidikan yang kerap dipahami secara tumpang tindih dalam pembelajaran bahasa Arab (PBA). Kajian yang secara integratif menganalisis perbedaan ketiga konsep beserta prinsip dan tujuannya dalam konteks PBA di Indonesia masih sangat terbatas, inilah celah yang diisi penelitian ini. Artikel ini bertujuan menganalisis perbedaan fundamental antara ketiganya, mengidentifikasi prinsip-prinsip penerapannya, serta mengkaji tujuannya secara spesifik dalam PBA. Dengan metode kajian pustaka dan pendekatan analisis konseptual, kajian ini menemukan bahwa pengukuran bersifat kuantitatif-deskriptif, penilaian bersifat interpretatif-pedagogis, dan evaluasi bersifat holistik-programatik. Penerapannya dalam PBA harus berlandaskan prinsip validitas, autentisitas, keterpaduan keterampilan, dan keberlanjutan. Tujuannya meliputi diagnosis kemampuan berbahasa Arab, pemantauan perkembangan belajar, dan peningkatan kualitas program PBA secara komprehensif.

Kata-kata kunci: Pembelajaran Bahasa Arab, Pengukuran, Penilaian, Evaluasi, Hasil Belajar

I. INTRODUCTION

Assessment constitutes an integral component of the learning process that cannot be separated from teaching activities themselves. In Indonesian educational practice, particularly

in the context of Arabic language learning (ALL), three terms are frequently used interchangeably without clear distinction: measurement (pengukuran), assessment (penilaian), and evaluation (evaluasi). Yet, these three concepts theoretically possess fundamentally different meanings, functions, procedures, and scopes. This conceptual confusion is not merely a terminological issue; it directly impacts the quality of assessment and evaluation practice in the field.¹

Mardapi affirms that measurement, assessment, and evaluation constitute three distinct yet mutually complementary activities that form a hierarchy within the educational process.² Measurement produces numerical data, assessment interprets that data within a pedagogical context, and evaluation employs assessment results as one consideration in judging overall program effectiveness. When Arabic language teachers fail to understand these distinctions, the common consequences are the use of inappropriate instruments, erroneous result interpretation, and pedagogical decisions unsupported by adequate data.

Arabic language learning as a foreign language in Indonesia possesses distinctive characteristics that differ from the learning of other subjects. The complexity of the Arabic linguistic system, encompassing aspects of phonology, morphology, syntax, and semantics, demands a comprehensive, authentic, and communicatively appropriate approach to measurement and assessment. Arifin maintains that language learning evaluation requires a specialized approach tailored to the objectives and characteristics of the target language, including the integrated consideration of the four language skills.³

Nurgiyantoro has addressed language learning assessment comprehensively in his seminal work in the field,⁴ while Fanani has specifically examined authentic assessment in ALL.⁵ Mustofa and Chollisni have investigated ALL program evaluation in Islamic schools (madrasah).⁶ Hidayat and Wahyudin have identified challenges in developing authentic assessment instruments for ALL at the senior secondary level (madrasah aliyah).⁷ However, studies that analyze in an integrative and comparative manner the fundamental differences

¹Ubaid Ridho, "Evaluasi dalam Pembelajaran Bahasa Arab," *An-Nabighoh: Jurnal Pendidikan dan Pembelajaran Bahasa Arab*, Vol. 20, No. 1 (2018), p. 22.

²Djemari Mardapi, *Pengukuran, Penilaian, dan Evaluasi Pendidikan*, 3rd ed. (Yogyakarta: Parama Publishing, 2020), p. 5.

³Zainal Arifin, *Evaluasi Pembelajaran: Prinsip, Teknik, dan Prosedur*, Rev. ed. (Bandung: PT. Remaja Rosdakarya, 2022), p. 4.

⁴Burhan Nurgiyantoro, *Penilaian Pembelajaran Bahasa Berbasis Kompetensi*, 3rd ed. (Yogyakarta: BPFE, 2021), p. 6.

⁵Muhammad Zainul Fanani, "Authentic Assessment dalam Pembelajaran Bahasa Arab: Konsep, Prinsip, dan Implementasi," *Arabiyatuna: Jurnal Bahasa Arab*, Vol. 5, No. 2 (2021), p. 303.

⁶Bisri Mustofa and Annisa Chollisni, "Evaluasi Program Pembelajaran Bahasa Arab di Madrasah: Model dan Implementasi," *Lisanul Arab: Journal of Arabic Learning and Teaching*, Vol. 11, No. 2 (2022), p. 117.

⁷Nurul Hidayat and Dinn Wahyudin, "Pengembangan Instrumen Penilaian Autentik dalam Pembelajaran Bahasa Arab di Madrasah Aliyah," *Al-Ta'rib: Jurnal Ilmiah Program Studi Pendidikan Bahasa Arab*, Vol. 11, No. 1 (2023), p. 47.

among the three concepts, measurement, assessment, and evaluation, along with their specific principles and objectives in the context of ALL in Indonesia remain very scarce. This is the academic gap that the present article seeks to address.

Furthermore, national educational policy through the Merdeka Curriculum, introduced in 2022, mandates a more diverse, authentic, and learner-centered approach to assessment. Rahmawati notes that the effective implementation of formative assessment in ALL continues to face considerable challenges, particularly regarding teachers' understanding of diverse assessment concepts and techniques.⁸ This further underscores the urgency of a study that systematically addresses the conceptual foundations of measurement, assessment, and evaluation in ALL.

Effendy highlights that one of the fundamental weaknesses in ALL implementation in Indonesia is the lack of clarity in establishing Arabic language competency measurement standards, which results in inconsistent assessment outcomes across educational institutions.⁹ Hermawan adds that Arabic language teachers need adequate competence not only in teaching, but also in designing valid, reliable, and authentic assessment systems appropriate to the characteristics of the Arabic language.¹⁰

Based on the foregoing background, this article specifically aims to: (1) analyze the fundamental differences among measurement, assessment, and evaluation within the perspective of educational theory; (2) identify and examine the principles underlying the application of each concept in the context of ALL; and (3) investigate the specific objectives of applying these three concepts in ALL. Through this study, it is hoped that Arabic language educators, curriculum developers, and educational policymakers will acquire a stronger and more comprehensive conceptual foundation for designing and implementing effective, equitable, and sustainable systems of measurement, assessment, and evaluation in Arabic language learning.

II. METHOD

This study employs a library research method (*kajian pustaka*) with a conceptual analysis approach (*analisis konseptual*). The library research method was selected because this study is descriptive-analytical in nature, aimed at exploring, describing, and analyzing theoretical concepts based on relevant and credible scholarly literature sources. Zed defines

⁸Fina Rahmawati, "Penerapan Penilaian Formatif dalam Meningkatkan Kemampuan Bahasa Arab Siswa: Studi Kasus di MAN 1 Makassar," *Jurnal Pendidikan Bahasa Arab dan Kebahasaaraban*, Vol. 10, No. 1 (2023), p. 91.

⁹Ahmad Fuad Effendy, *Metodologi Pengajaran Bahasa Arab*, 6th ed. (Malang: Misykat, 2021), p. 195.

¹⁰Acep Hermawan, *Metodologi Pembelajaran Bahasa Arab*, Rev. ed. (Bandung: PT. Remaja Rosdakarya, 2022), p. 185.

library research as a series of activities concerned with methods of data collection from written sources, reading and note-taking, and processing research materials.¹¹

The data sources in this study consist of two categories. First, primary sources comprising textbooks in the fields of educational evaluation, measurement and assessment, and Arabic language learning methodology published between 2017 and 2025. Second, secondary sources comprising peer-reviewed journal articles accredited at Sinta 1 through Sinta 3 (Indonesia's national journal accreditation tiers) that specifically address assessment, evaluation, and measurement in the context of Arabic language education, as well as relevant national educational policy documents.

Data were collected through systematic documentary study, involving the gathering, reading, recording, and classifying of relevant literature sources based on the topic and focus of inquiry. Data analysis employed two primary techniques: content analysis (*analisis isi*) and comparative conceptual analysis (*analisis komparatif konseptual*). The analytical stages comprised: (1) identification and inventory of key concepts from various literature sources; (2) categorization and classification based on the definitional, principled, and purposive aspects of each concept; (3) comparison between concepts to identify their fundamental similarities and differences; and (4) synthesis and conclusion-drawing based on the overall findings of the analysis.¹²

The validity of this study is ensured through source triangulation, confirming each conceptual finding from a minimum of three different and mutually corroborating reference sources. Additionally, this study employs a conceptual member checking technique by verifying each interpretation of a concept against the definitions or explanations provided by field experts.

III. FINDINGS AND DISCUSSION

1. Definitions and Fundamental Differences: Measurement, Assessment, and Evaluation

a. Measurement in Arabic Language Learning

Measurement constitutes a process of quantifying or assigning numerical values to a characteristic or attribute of an object based on established rules or procedures. Mardapi defines measurement as the process of determining the magnitude of an attribute using a standardized measuring instrument, yielding numerical data.¹³ Measurement is essentially the activity of

¹¹Mestika Zed, *Metode Penelitian Kepustakaan*, Rev. ed. (Jakarta: Yayasan Obor Indonesia, 2022), p. 3.

¹²Matthew B. Miles, A. Michael Huberman, and Johnny Saldana, *Qualitative Data Analysis: A Methods Sourcebook*, 4th ed. (Thousand Oaks: SAGE Publications, 2020), p. 10.

¹³Mardapi, *Pengukuran, Penilaian, dan Evaluasi Pendidikan*, p. 14.

comparing something against a particular unit of measure, such that its results are quantitative and value-free, meaning that measurement results only describe facts without rendering judgments about whether those results are good, poor, adequate, or inadequate.

In the context of ALL, measurement means the process of assigning scores or numerical values to students' Arabic language ability using valid and reliable instruments. Measurement instruments commonly used in ALL include various types of Arabic language tests: component-language tests (nahwu, sharaf, mufradat, ashwat) and language-proficiency tests encompassing the four maharah: maharah al-istima' (listening), maharah al-kalam (speaking), maharah al-qira'ah (reading), and maharah al-kitabah (writing), as well as rating scales for measuring productive skills such as speaking and writing in Arabic.¹⁴

Arikunto distinguishes measurement from assessment by stating that to measure is to compare something against a quantitative standard, while to assess is to make a decision about something in terms of qualitative good–bad judgments.¹⁵ Kusaeri and Suprananto affirm that measurement in education is a systematic process for obtaining numerical descriptions of the degree to which individuals possess certain characteristics.¹⁶ The quality of measurement in ALL is principally determined by two criteria: validity and reliability. Validity ensures that the instrument measures what it is intended to measure, while reliability ensures the consistency of measurement results when administered repeatedly under equivalent conditions.¹⁷

b. Assessment in Arabic Language Learning

Assessment is the process of collecting, integrating, and interpreting information about student learning and development for the purposes of pedagogical decision-making. Nurgiyantoro defines assessment as the process of obtaining information in any form that can serve as the basis for decisions about students, whether pertaining to their curriculum, their program, or school policies.¹⁸ Unlike measurement, which merely produces numbers, assessment encompasses a broader, pedagogically dimensional process: not only gathering data but also interpreting that data in light of established learning objectives.

In the tradition of language education, language assessment is categorized into several types based on function and timing. Formative assessment is administered during the learning process with the aim of providing constructive feedback to improve learning. Summative

¹⁴M. Soenardi Djiwandono, *Tes Bahasa: Pegangan Bagi Pengajar Bahasa* (Jakarta: PT Indeks, 2011), p. 15.

¹⁵Suharsimi Arikunto, *Dasar-dasar Evaluasi Pendidikan*, 3rd ed. (Jakarta: Bumi Aksara, 2021), p. 3.

¹⁶Kusaeri and Suprananto, *Pengukuran dan Penilaian Pendidikan*, 2nd ed. (Yogyakarta: Graha Ilmu, 2022), p. 7.

¹⁷M. Abdul Hamid, *Mengukur Kemampuan Bahasa Arab untuk Studi Islam* (Malang: UIN Maliki Press, 2013), p. 18.

¹⁸Nurgiyantoro, *Penilaian Pembelajaran Bahasa Berbasis Kompetensi*, p. 8.

assessment is administered at the end of a learning period to determine overall achievement of learning outcomes. Diagnostic assessment is administered before instruction begins to identify students' prior knowledge and learning needs. Placement assessment aims to group students according to their levels of Arabic language proficiency.¹⁹

Fanani affirms that assessment in ALL should ideally be authentic (authentic assessment), meaning assessment that measures students' Arabic language ability through real-world tasks reflecting language use in everyday contexts and meaningful communicative situations.²⁰ Assessment in ALL also encompasses a more personal and reflective dimension: self-assessment and peer assessment, both of which encourage students to participate actively in evaluating their own learning, develop metacognitive skills, and build awareness of their strengths and weaknesses in Arabic language ability.²¹

c. Evaluation in Arabic Language Learning

Evaluation is the concept with the broadest scope among the three under discussion. Stufflebeam and Zhang define evaluation as the process of delineating, obtaining, reporting, and applying descriptive and judgmental information about the merit, worth, and significance of some object in order to guide decision-making, support accountability, and disseminate understanding.²² Arikunto elaborates that evaluation is the activity of collecting information about how something works, which information is then used to determine the appropriate alternative course of action.²³

In the context of ALL, evaluation extends well beyond the evaluation of student learning outcomes. Mustofa and Chollisni categorize ALL evaluation into three hierarchical levels: (1) classroom-level evaluation, focused on the effectiveness of learning activities in the classroom; (2) program-level evaluation, assessing the overall effectiveness of the ALL program, including curriculum, methods, materials, and resources; and (3) policy-level evaluation, assessing the relevance and effectiveness of Arabic language education policy at the institutional or national level.²⁴

ALL evaluation may employ various models, including: Stufflebeam's CIPP Model (Context, Input, Process, Product) which evaluates context, input, process, and product; the

¹⁹Imam Asrori, Muhammad Thohir, and M. Ainin, *Evaluasi Pembelajaran Bahasa Arab* (Malang: Misykat, 2017), p. 22.

²⁰Fanani, "Authentic Assessment dalam Pembelajaran Bahasa Arab," p. 305.

²¹Dina Indriana, "Evaluasi Pembelajaran dan Penilaian Autentik dalam Pembelajaran Bahasa Arab," *Al-Ittijah: Jurnal Keilmuan dan Kependidikan Bahasa Arab*, Vol. 10, No. 2 (2018), p. 38.

²²Daniel L. Stufflebeam and Guili Zhang, *The CIPP Evaluation Model: How to Evaluate for Improvement and Accountability* (New York: The Guilford Press, 2017), p. 12.

²³Arikunto, *Dasar-dasar Evaluasi Pendidikan*, p. 3–4.

²⁴Mustofa and Chollisni, "Evaluasi Program Pembelajaran Bahasa Arab di Madrasah," p. 119–120.

Kirkpatrick Model evaluating four levels, response, learning, behavior, and results; and Stake's Countenance Model comparing planned conditions against actual conditions. The selection of an appropriate evaluation model depends greatly on the evaluation's objectives and the characteristics of the ALL program being evaluated.²⁵²⁶

Table 1.
Fundamental Comparison of Measurement, Assessment, and Evaluation in ALL

Aspect	Measurement	Assessment	Evaluation
Definition	Process of numerically quantifying attributes of Arabic language ability	Process of collecting and interpreting learning information for pedagogical decision-making	Process of holistic appraisal of ALL programs, systems, and policies
Nature of Results	Quantitative (score/number), value-free	Qualitative & quantitative, interpretive	Qualitative, holistic, programmatic, normative
Focus	Student ability attributes/characteristics	Individual learning progress & needs	Programs, policies, curriculum, ALL system
Instruments	Standardized tests, scoring rubrics, rating scales	Tests, portfolios, observation, authentic tasks, learning journals	Questionnaires, in-depth interviews, FGDs, document analysis
Timing	At any time as needed	Continuous: before, during, and after learning	Periodic: end of semester, year, or program
Decision-Maker	Teacher/instructor	Teacher & students	Evaluation team, principal, policymakers
Primary Objective	Generate accurate data on Arabic language ability	Support sound pedagogical decisions	Improvement and development of ALL programs

2. Principles of Measurement, Assessment, and Evaluation in ALL

a. Principles of Measurement in ALL

For measurement in ALL to yield meaningful and dependable data, a number of principles must be observed. Mardapi identifies five principal principles of educational measurement.²⁷ First, validity, meaning that the measurement instrument must genuinely measure the ability construct intended. In ALL, for example, a test of maharah al-kalam must directly measure oral Arabic ability, not grammatical recall in written form. This validity encompasses content validity, construct validity, and criterion validity.

Second, reliability, that is, the consistency of measurement results when administered repeatedly to the same group under equivalent conditions. Third, objectivity, meaning that

²⁵Ii Rahmanudin et al., "Evaluasi Kurikulum Pembelajaran Bahasa Arab di SMA Islam Al Azhar 24 Boarding School," Taqdir, Vol. 8, No. 2 (2022), p. 231.

²⁶Aziz Fitriani, "Evaluasi Program E-Learning pada Prodi Pendidikan Bahasa Arab Universitas Muhammadiyah Yogyakarta dengan Model CIPP," Maharaat: Jurnal Pendidikan Bahasa Arab, Vol. 3, No. 2 (2021), p. 114.

²⁷Mardapi, Pengukuran, Penilaian, dan Evaluasi Pendidikan, p. 18.

measurement results must not be influenced by the rater's subjectivity. This principle is especially important in measuring productive Arabic skills such as kalam and kitabah, where inter-rater reliability must be maintained through the use of clear, detailed rubrics and rater training. Arifin affirms that objectivity in language measurement depends greatly on the quality of the scoring rubrics employed.²⁸

Fourth, practicality, meaning that measurement instruments must be feasibly administered given constraints of time, cost, and available resources without sacrificing data quality. Fifth, normativity, meaning that measurement results must be interpreted against a particular norm or standard, either a group norm (norm-referenced) comparing individual performance with group performance, or a criterion norm (criterion-referenced) comparing individual performance with the competency standard set in the Arabic language curriculum. Kusaeri and Suprananto further add the principle of clarity, meaning that instruments must employ language that is unambiguous and readily understood by students.²⁹

b. Principles of Assessment in ALL

Effective and meaningful assessment in ALL must rest on a set of integrated principles. Nurgiyantoro formulates comprehensive principles of language assessment as follows.³⁰ First, integration, meaning that assessment must be designed as an organic part of the learning process, not as a separate, unrelated activity. Second, authenticity, which requires that assessment tasks reflect genuine, meaningful, and student-relevant Arabic language use. Fanani emphasizes that authenticity in ALL assessment encompasses both task authenticity, the degree to which tasks resemble real-world Arabic language use, and response authenticity, the freedom afforded to students in expressing their Arabic language ability without being overly constrained by rigid response formats.³¹

Third, learner involvement, which encourages students to play an active role in the assessment process through self-assessment, learning reflection, and independent target-setting. Fourth, transparency, meaning that assessment criteria and standards must be explicitly communicated to students prior to assessment. Fifth, continuity, meaning that assessment must be conducted on an ongoing basis throughout the learning process. Hermawan adds the principle of comprehensiveness, stating that ALL assessment must encompass all aspects of Arabic language competence, not only cognitive aspects, but also psychomotor aspects (active

²⁸Arifin, *Evaluasi Pembelajaran*, p. 22.

²⁹Kusaeri and Suprananto, *Pengukuran dan Penilaian Pendidikan*, p. 9.

³⁰Nurgiyantoro, *Penilaian Pembelajaran Bahasa Berbasis Kompetensi*, p. 12.

³¹Fanani, "Authentic Assessment dalam Pembelajaran Bahasa Arab," p. 307.

language ability) and affective aspects (attitudes and motivation toward Arabic language learning).³²

Rahmawati, drawing on her research in Islamic schools, identifies feedback as the most crucial principle in ALL formative assessment. Specific, timely, and constructive feedback has been shown to significantly enhance students' motivation and Arabic language ability. According to her, effective feedback in ALL must include: identification of specific Arabic language errors, explanation of the causes of those errors, and concrete, actionable correction suggestions for students.³³

c. Principles of Evaluation in ALL

Comprehensive and impactful evaluation in ALL must be grounded in scientifically validated principles. Stufflebeam and Zhang identify four standards of sound evaluation: utility, feasibility, propriety, and accuracy.³⁴ When elaborated in the context of ALL, these four standards yield the following operational principles.

First, comprehensiveness: evaluation must encompass all components of the ALL program. Second, relevance: evaluation must be relevant to the specific objectives, needs, and context of the ALL program being evaluated. Third, objectivity: the evaluation process and its conclusions must be grounded in strong, verifiable empirical data and evidence. Fourth, participatory engagement: sound ALL evaluation involves all stakeholders, Arabic language teachers, students, parents, school principals, supervisors, and graduate-user institutions. Fifth, utility: evaluation findings must be practically and concretely applicable to improving and developing the ALL program. Sixth, continuity: ALL program evaluation must be conducted periodically and in a planned manner. Seventh, systematicness: evaluation must follow a structured, step-by-step procedure grounded in a sound methodological framework. Eighth, meaningfulness: evaluation must produce information that is genuinely beneficial for ALL program development, not merely fulfilling administrative obligations.

3. Objectives of Applying Measurement, Assessment, and Evaluation in ALL

a. Objectives of Measurement in ALL

Measurement in ALL serves several specific, interrelated objectives. Kusaeri and Suprananto formulate the objectives of educational measurement that can be contextualized within ALL as follows.³⁵ First, to determine and document students' level of Arabic language

³²Hermawan, *Metodologi Pembelajaran Bahasa Arab*, p. 190.

³³Rahmawati, "Penerapan Penilaian Formatif," p. 93.

³⁴Stufflebeam and Zhang, *The CIPP Evaluation Model*, p. 18.

³⁵Kusaeri and Suprananto, *Pengukuran dan Penilaian Pendidikan*, p. 14.

competency mastery in a measurable and objective manner, enabling meaningful comparisons against established competency standards. Second, to provide accurate and dependable baseline data for teachers and educational institutions in designing more targeted and efficient Arabic language learning programs. Third, to identify students requiring special attention and intervention, whether students experiencing difficulties below the minimum standard or those gifted and potentially outstanding in Arabic who require enrichment. Fourth, to fulfill the needs of educational accountability, with valid and reliable measurement data serving as empirical evidence of Arabic language learning achievement that can be accounted for to various stakeholders. Fifth, to support research and development in the field of ALL.³⁶

b. Objectives of Assessment in ALL

Effendy states that the objectives of assessment in Arabic language teaching encompass dimensions far broader than merely the assignment of final grades.³⁷ Nurgiyantoro classifies the objectives of language assessment into five comprehensive functional categories.³⁸ First, the diagnostic objective: to identify in detail the strengths and weaknesses of individual students' Arabic language ability so that teachers can design and implement appropriately differentiated instructional interventions. Diagnostic assessment in ALL can reveal, for instance, whether a student's difficulties lie in aspects of nahwu, vocabulary mastery, receptive skills, or productive skills in Arabic.

Second, the formative objective: to monitor student learning progress on a continuous basis throughout the learning process and to provide constructive feedback that can be immediately used to improve learning strategies. Rahmawati demonstrates in her research that the consistent and high-quality implementation of formative assessment can significantly improve students' Arabic language ability, particularly in speaking and writing.³⁹ Third, the summative objective: to determine and document students' Arabic language learning achievement for a given period as the basis for assigning report-card grades, making promotion decisions, or awarding Arabic language proficiency certificates.

Fourth, the placement objective: to group students according to their levels of Arabic language ability so they may learn within more homogeneous groups, receive appropriately leveled instruction, or determine suitable further study programs. Fifth, the motivational objective: well-designed and well-implemented assessment, particularly authentic assessment

³⁶Muhammad Matsna and Erta Mahyudin, *Pengembangan Evaluasi dan Tes Bahasa Arab* (Tangerang: Al-Kitabah, 2012), p. 37.

³⁷Effendy, *Metodologi Pengajaran Bahasa Arab*, p. 198.

³⁸Nurgiyantoro, *Penilaian Pembelajaran Bahasa Berbasis Kompetensi*, p. 45.

³⁹Rahmawati, "Penerapan Penilaian Formatif," p. 95.

that affords students genuine experiences of success in using Arabic, has been shown to enhance students' intrinsic motivation to continue developing their Arabic language ability. Sixth, the selection objective: to identify and determine students who meet certain Arabic language ability qualifications, for example for admission to Arabic language education departments or scholarship selection to Arab-speaking countries.

c. Objectives of Evaluation in ALL

Mustofa and Chollisni formulate the objectives of ALL program evaluation hierarchically by level of evaluation.⁴⁰ The first objective is to assess the overall effectiveness of the ALL program, whether it has successfully achieved the learning objectives stipulated in the curriculum, the extent to which students have attained the expected Arabic language competencies, and how efficiently resources have been deployed. The second objective is to systematically identify the strengths and weaknesses of every component of the ALL program, including curriculum relevance, teaching material quality, instructional method effectiveness, Arabic teacher competence, adequacy of learning facilities, and learning environment conditions.

The third objective is to provide evidence-based information to Arabic language educational policymakers at the school, regional, and national levels to support more rational, accountable, and quality-oriented decision-making. Arifin affirms that sound evaluation must produce recommendations that are concrete, implementable, and grounded in robust data.⁴¹ The fourth objective is to measure the relevance of the ALL program to students' genuine needs, the demands of a workforce increasingly requiring Arabic language proficiency, and advances in Arabic linguistics and language education scholarship. The fifth objective is to promote public accountability in the delivery of ALL programs. The sixth objective is to develop and refine evaluation models best suited to the specific characteristics and context of ALL in Indonesia, given that existing evaluation models have largely been developed in English-language learning or Western educational contexts not directly applicable to ALL in Indonesian Islamic educational institutions.

4. Hierarchical Relationship and Integrated Implementation in the ALL System

Measurement, assessment, and evaluation in ALL cannot be understood or implemented in isolation; they must be regarded as a mutually supportive system forming a logical hierarchy. Hidayat and Wahyudin affirm that understanding the hierarchical relationship among these

⁴⁰Mustofa and Chollisni, "Evaluasi Program Pembelajaran Bahasa Arab di Madrasah," p. 123.

⁴¹Arifin, *Evaluasi Pembelajaran*, p. 60.

three concepts is a prerequisite competency that every professional Arabic language teacher must possess.⁴² Measurement occupies the foundational position in this hierarchy: it provides the quantitative data that constitute the raw material for the assessment process. Assessment occupies the intermediate level: it processes measurement data together with other qualitative information, such as observation records, portfolios, and anecdotal notes, to produce sound pedagogical decisions about Arabic language learning and teaching. Evaluation occupies the highest position: it employs assessment results together with various other information sources to appraise the overall effectiveness of the ALL program.

Hermawan depicts the ideal cycle of integrated implementation of these three concepts in ALL: teachers plan valid and reliable measurement to assess students' Arabic language competencies; measurement results are interpreted through a comprehensive assessment process; assessment information is used for sound pedagogical decision-making; and periodically, the entire learning process is evaluated to identify areas requiring improvement and development.⁴³

The principal challenges of this integrated implementation in Indonesia, particularly in the context of ALL at Islamic schools (madrasah) and Islamic higher education institutions, include: (1) limitations in Arabic language teachers' conceptual understanding of the differences and relationships among measurement, assessment, and evaluation; (2) limitations in teachers' technical competence in developing valid and reliable measurement instruments for various aspects of Arabic language ability; (3) limitations in time and resources for implementing a comprehensive and authentic assessment system; and (4) the absence of a standardized ALL program evaluation system consistently implemented at the national level. Addressing these challenges requires comprehensive educational policy, ongoing professional development programs for Arabic language teachers, and active collaboration among the various stakeholders in Arabic language education in Indonesia.

IV. CONCLUSION

In this systematic review of scholarly literature on measurement, assessment, and evaluation in Arabic language learning (ALL), three principal findings emerge. First, measurement, assessment, and evaluation constitute a hierarchical continuum with distinct epistemological foundations: measurement serves as the quantitative, value-neutral operationalization of language proficiency; assessment functions as the interpretive-

⁴²Hidayat and Wahyudin, "Pengembangan Instrumen Penilaian Autentik," p. 50.

⁴³Hermawan, *Metodologi Pembelajaran Bahasa Arab*, p. 193.

pedagogical process integrating qualitative and quantitative data for learning-oriented purposes; and evaluation operates as the holistic, programmatic appraisal of ALL systems directed toward policy formulation. Second, each construct is governed by discipline-specific principles—measurement by validity, reliability, objectivity, practicality, normativity, and clarity; assessment by integration, authenticity, learner involvement, transparency, continuity, comprehensiveness, and constructive feedback; and evaluation by comprehensiveness, relevance, objectivity, participatory engagement, utility, systematicness, and meaningfulness. Third, to operationalize these conceptual distinctions and their underlying principles, multilevel interventions are requisite: educators require sustained professional development programs facilitating reconceptualization of these three constructs; educational institutions must establish integrated, contextually-sensitive, and evidence-based measurement and assessment systems; and policymakers must promulgate nationally-standardized competency frameworks for ALL that demonstrate validity, reliability, and authenticity, thereby ensuring quality and equitable implementation across Indonesia's educational landscape.

V. REFERENCES

Books

- Arifin, Z. (2022). *Evaluasi Pembelajaran: Prinsip, Teknik, dan Prosedur* (Rev. ed.). Bandung: PT. Remaja Rosdakarya.
- Arikunto, S. (2021). *Dasar-dasar Evaluasi Pendidikan* (3rd ed.). Jakarta: Bumi Aksara.
- Effendy, A. F. (2021). *Metodologi Pengajaran Bahasa Arab* (6th ed.). Malang: Misykat.
- Hermawan, A. (2022). *Metodologi Pembelajaran Bahasa Arab* (Rev. ed.). Bandung: PT. Remaja Rosdakarya.
- Kusaeri, K., & Suprananto, S. (2022). *Pengukuran dan Penilaian Pendidikan* (2nd ed.). Yogyakarta: Graha Ilmu.
- Mardapi, D. (2020). *Pengukuran, Penilaian, dan Evaluasi Pendidikan* (3rd ed.). Yogyakarta: Parama Publishing.
- Miles, M. B., Huberman, A. M., & Saldana, J. (2020). *Qualitative Data Analysis: A Methods Sourcebook* (4th ed.). Thousand Oaks: SAGE Publications.
- Nurgiyantoro, B. (2021). *Penilaian Pembelajaran Bahasa Berbasis Kompetensi* (3rd ed.). Yogyakarta: BPFE.
- Stufflebeam, D. L., & Zhang, G. (2017). *The CIPP Evaluation Model: How to Evaluate for Improvement and Accountability*. New York: The Guilford Press.
- Zed, M. (2022). *Metode Penelitian Kepustakaan* (Rev. ed.). Jakarta: Yayasan Obor Indonesia.
- Asrori, I., Thohir, M., & Ainin, M. (2017). *Evaluasi Pembelajaran Bahasa Arab*. Malang: Misykat.
- Djiwandono, M. S. (2011). *Tes Bahasa: Pegangan Bagi Pengajar Bahasa*. Jakarta: PT Indeks.
- Hamid, M. A. (2013). *Mengukur Kemampuan Bahasa Arab untuk Studi Islam*. Malang: UIN Maliki Press.

Matsna, M., & Mahyudin, E. (2012). Pengembangan Evaluasi dan Tes Bahasa Arab. Tangerang: Al-Kitabah.

Journal Articles

- Fanani, M. Z. (2021). Authentic Assessment dalam Pembelajaran Bahasa Arab: Konsep, Prinsip, dan Implementasi. *Arabiyatuna: Jurnal Bahasa Arab*, 5(2). <https://doi.org/10.29240/jba.v5i2.3015>
- Hidayat, N., & Wahyudin, D. (2023). Pengembangan Instrumen Penilaian Autentik dalam Pembelajaran Bahasa Arab di Madrasah Aliyah. *Al-Ta'rib: Jurnal Ilmiah Program Studi Pendidikan Bahasa Arab*, 11(1). <https://doi.org/10.23971/altarib.v11i1.5812>
- Mustofa, B., & Chollisni, A. (2022). Evaluasi Program Pembelajaran Bahasa Arab di Madrasah: Model dan Implementasi. *Lisanul Arab: Journal of Arabic Learning and Teaching*, 11(2). <https://doi.org/10.15294/la.v11i2.52318>
- Rahmawati, F. (2023). Penerapan Penilaian Formatif dalam Meningkatkan Kemampuan Bahasa Arab Siswa: Studi Kasus di MAN 1 Makassar. *Jurnal Pendidikan Bahasa Arab dan Kebahasaaraban*, 10(1). <https://doi.org/10.24042/jpba.v10i1.14520>
- Fitriani, A. Z. (2021). Evaluasi Program E-Learning pada Prodi Pendidikan Bahasa Arab Universitas Muhammadiyah Yogyakarta dengan Model CIPP. *Maharaat: Jurnal Pendidikan Bahasa Arab*, 3(2), 109–127. <https://doi.org/10.18196/mht.v3i2.11786>
- Haniefa, R. (2022). Implementasi Model Penilaian HOTS (Higher Order Thinking Skills) pada Penilaian Empat Keterampilan Berbahasa Arab. *Ta'limi: Journal of Arabic Education and Arabic Studies*, 1(1), 49–71. <https://doi.org/10.15575/talimi.v1i1.16857>
- Indriana, D. (2018). Evaluasi Pembelajaran dan Penilaian Autentik dalam Pembelajaran Bahasa Arab. *Al-Ittijah: Jurnal Keilmuan dan Kependidikan Bahasa Arab*, 10(2), 34–52. <https://doi.org/10.32678/alittijah.v10i2.1873>
- Laili, M. (2020). Ketepatan Konstruksi Butir Pilihan Ganda Bahasa Arab. *ALSUNIYAT: Jurnal Penelitian Bahasa, Sastra, dan Budaya Arab*, 3(2), 111–124. <https://doi.org/10.17509/alsuniyat.v3i2.25272>
- Qomariyah, L. (2022). Analisis Tingkat Kesukaran dan Daya Pembeda Butir Soal TOAFL Universitas Hasyim Asy'ari Tebuireng Jombang. *Lisanan Arabiya: Jurnal Pendidikan Bahasa Arab*, 6(1), 1–18. <https://doi.org/10.32699/liar.v6i1.2549>
- Rahmanudin, I., Anwar, M. S., Ahmad, B. M. B., & Qomari, N. (2022). Evaluasi Kurikulum Pembelajaran Bahasa Arab di SMA Islam Al Azhar 24 Boarding School. *Taqdir*, 8(2), 227–239. <https://doi.org/10.19109/taqdir.v8i2.13781>
- Ridho, U. (2018). Evaluasi dalam Pembelajaran Bahasa Arab. *An-Nabighoh: Jurnal Pendidikan dan Pembelajaran Bahasa Arab*, 20(1), 19–44. <https://doi.org/10.32332/an-nabighoh.v20i01.1124>
- Setyawan, C. E. (2015). Desain Evaluasi Program Pembelajaran Bahasa Arab. *Al-Manar: Jurnal Komunikasi dan Pendidikan Islam*, 4(1). <https://doi.org/10.36668/jal.v4i1.64>
- Fitriah, F., Abu, A., Malli, R., & Getteng, A. R. (2022). Evaluasi Program Camp Al-Qur'an di SMK Nasional Makassar: Studi Evaluasi Model Context, Input, Process, dan Product. *Al-Afkar: Journal for Islamic Studies*, 5(1), 312–333. <https://doi.org/10.31943/afkarjournal.v5i1.252>
- Makhisoh, I., & Suroyya, E. N. (2022). Analisa Butir Soal Ujian Madrasah Bahasa Arab Menurut Taksonomi Bloom. *Muhadasah: Jurnal Pendidikan Bahasa Arab*, 4(1), 1–18. <https://doi.org/10.30984/muhadasah.v4i1.1684>

- Nisa, A. K., & Al Ghifary, M. (2023). Implementasi Kurikulum Merdeka pada Pembelajaran Bahasa Arab di Madrasah Aliyah Negeri 1 Kendari. *Al Mi'yar: Jurnal Ilmiah Pembelajaran Bahasa Arab dan Kebahasaaraban*, 6(2), 627–642. <https://doi.org/10.35931/am.v6i2.2685>