ISSN 2776-5229

The Problematic of Authentic Assessment in PAI Learning under the Merdeka Curriculum: A Case Study of Affective and Psychomotor Competencies of Students in PAI Subjects

Achmad Junaidi[1], Cholilurrahman[2], and Moh Hafid Effendy[3], Maimun[4]
[1,2,3,4]Universitas Islam Negeri Madura, Pamekasan, Indonesia
*achjun56@gmail.com

ARTICLE INFO

Article history
Received September 29, 2026
Revised October 27, 2026
Accepted November 5, 2026

Keywords: Authentic Assessment; Affective Domain; Psychomotor Domain; Merdeka Curriculum; Islamic Religious Education

ABSTRACT
The Merdeka Curriculum positions assessment as an integral part of learning, including the affective and psychomotor domains of Islamic Religious Education (PAI) subjects. This study aims to describe authentic assessment practices as reported by teachers, the problems reported, the factors perceived to contribute to them, and the use of assessment results. A qualitative single case study was conducted at MTs Nahdlatul Ulum. Data were collected through semi-structured interviews with four teachers, each teaching one PAI subject, and analyzed comparatively across subjects through coding, categorization, and theme development. Teachers reported five stages of assessment practice: planning, indicator development, selection of instruments and rubrics, observation or performance assessment, and collection of learning evidence; the use of results formed a sixth stage. Six problems were reported, including limited depth of instruments, difficulty operationalizing affective indicators, unequal opportunities for individual performance, inadequate documentation, and potential inconsistency in assessment decisions. Perceived contributing factors were limited time and class size, technical difficulties in developing instruments, and limited documentation facilities. Assessment results were used for feedback, remedial activities, enrichment, and instructional improvement, although informally. The challenges relate more to technical and structural consistency than to conceptual understanding. As the study relied on interviews in a single madrasah, the findings represent teacher-reported practices and were not verified through classroom observation or document analysis.

1. INTRODUCTION

The Merdeka Curriculum positions assessment not merely as a means to determine final grades, but as an integral part of the planned, implemented, and utilized learning cycle to determine follow-up actions (Badan Standar, Kurikulum, dan Asesmen Pendidikan, 2024). In Islamic Religious Education (PAI), this demand is broader because learning outcomes encompass the internalization of religious values manifested in attitudes and skills (Baroroh et al., 2022), so assessments need to capture students' competencies comprehensively through learning evidence that is not limited to written tests (Rozikin et al., 2026). In madrasah tsanawiyah, PAI is divided into four standalone subjects, namely Islamic Cultural History (SKI), Creed and Morals, Jurisprudence, and Quran and Hadith; the term "PAI learning" in this article refers to all four collectively. Authentic assessment is relevant because it evaluates abilities through tasks and performances that reflect the application of competencies in meaningful situations (Ismail et al., 2022), such as reading the Quran, practicing worship, or being honest and responsible. In this study, a practice is categorized as authentic if it meets four operational criteria: (1) related to meaningful tasks or situations, not merely recalling information; (2) based on evidence of actual performance, in the form of observed behavior, work

performance, or products; (3) using criteria or rubrics, no matter how simple; and (4) its results are connected to learning follow-up. These criteria serve as an analytical lens so that practices reported by informants, such as presentations or observations, are not deemed authentic without a clear operational basis (Ismail et al., 2023., (Imanudin, 2023). Its application demands a complex series of tasks, namely determining indicators (Sirinding et al., 2023), selecting instruments, developing rubrics, observing performance, collecting evidence, and utilizing results. This complexity is even more pronounced in the affective and psychomotor domains (Forisma et al., 2023., Wijaya et al., 2023). The affective domain (attitudes, interests, motivation, and value appreciation) cannot be determined through a single measurement, thus requiring clear indicators, continuous observation, and behavioral evidence (Hazyimara et al., 2024., Nawali et al., 2024), while the psychomotor domain demands competence to be demonstrated through real actions, such as religious practices and reading the Al-Qur'an (Syamsiah & Prastowo, 2022). Rozikin et al. (2026) propose an integrated authentic assessment model for PAI that includes portfolios, religious projects, self and peer assessments based on introspection, and continuous behavioral observation. However, the review by Rustam et al. (2025) concludes that in practice, the affective and psychomotor assessments of PAI tend to be limited, unsustainable, and lack systematic instruments. Previous research shows a similar pattern of obstacles (Table 1). Prihantoro (2021) concluded from 27 studies that the implementation of authentic PAI assessments is not yet optimal, among other reasons, due to limited teacher understanding and minimal socialization and training, while Rosidah et al. (2021) found that teachers have a sufficient understanding of the concept but struggle with rubric development and grading. In the implementation of the Merdeka Curriculum, obstacles such as time constraints and suboptimal feedback were reported by Rahayu and Istikomah (2024), the complexity of instruments in SKI by Sahlan et al. (2023), and the lack of teacher understanding and limited resources by Fadlillah and Kusaeri (2024).

Table 1. Summary of previous research on authentic assessment in PAI learning

Researcher (year)	Context and methods	Focus	Main findings
Rosidah et al. (2021)	Descriptive; FGD with 50 teachers in Mojokerto	The readiness of teachers to implement authentic assessment (from preparation to reporting)	The teacher understands the concept well enough, but has difficulty creating rubrics and processing grades, and has not yet received special training.
Prihantoro (2021)	Systematic review of 27 studies; PAI elementary and middle school	Failure in the implementation of authentic PAI assessment	Related to the limited understanding of teachers as well as the lack of socialization and training
Syamsiah & Prastowo (2022)	Field qualitative; SD	Authentic assessment in the psychomotor domain of PAI	Planning and practice (e.g., prayer practice) are carried out; the results are used for follow-up.
Sahlan et al. (2023)	Case study; SKI in madrasah	Authentic assessment of SKI in the Merdeka Curriculum	The complexity of instrument preparation, the numerous rubrics and items, and the time constraints
Hazyimara et al. (2024)	Descriptive qualitative; elementary school in Malang	Assessment of the affective domain in PAI	Planning, implementation, and evaluation; follow-up for students who have not reached the assessment threshold
Fadilah et al. (2025)	Instrumental case study; one public high school in Bandung	Psychomotor assessment of PAI and teacher readiness	Analytic rubrics, direct observation, video portfolios, and performance tasks; results for feedback, remediation, and enrichment

Researcher (year)	Context and methods	Focus	Main findings
Indriani et al. (2025)	Case study; Elementary school in Bandung	Affective assessment in PAI	Limited documentation and more oral; the use of formal instruments has not been consistent.
Andrianto et al. (2026)	Qualitative case study; one MTs in Central Sulawesi; interviews, observations, and documentation	The readiness of PAI teachers to develop and use psychomotor assessment instruments	Teachers understand the concept, but their technical readiness to develop measurable indicators, analytical rubrics, and systematic documentation is limited; constraints include assessment literacy, technical training, time, and the number of students.

The synthesis of Table 1 shows three unanswered issues. First, most studies examine teacher readiness or one of the domains (affective or psychomotor), whereas both are interconnected: the internalization of values in the affective domain ideally manifests in skills in the psychomotor domain. Second, studies are primarily conducted at the elementary and high school levels; at the madrasah tsanawiyah level, the only research traced by the author is by Andrianto et al. (2026), which is limited to psychomotor instruments. Third, studies tend to stop at listing obstacles or a single stage of assessment, thus not indicating at which stage in the assessment cycle issues arise and what distinguishes the problems (symptoms) from the factors perceived to contribute. The results are also not uniform: some studies place teachers' understanding as a source of obstacles (Prihantoro, 2021), while others emphasize technical difficulties (Rosidah et al., 2021; Andrianto et al., 2026). This research fills that gap by mapping the issues of affective and psychomotor assessments simultaneously throughout the entire assessment cycle, which includes five stages of practice (planning, determining indicators, selecting instruments and rubrics, observation or practice, and collecting learning evidence) and the utilization of results as the sixth stage, by distinguishing issues from factors perceived to contribute and comparing subjects within the PAI cluster. The research was conducted at MTs Nahdlatul Ulum to answer four questions: (1) how the practice of authentic assessments in the affective and psychomotor domains is reported by teachers; (2) what issues are reported at the assessment stages; (3) what factors are perceived by teachers to contribute to them; and (4) how assessment results are utilized for feedback, remediation, enrichment, and learning improvement. Because it is interview-based, the findings are interpreted as practices reported by teachers, not as implementations that have been directly verified.

2. METHODS

This research uses a qualitative approach with a single case study design to gain an in-depth understanding of the authentic assessment practices in the affective and psychomotor domains as reported by teachers. The case unit is MTs Nahdlatul Ulum, with the case boundary being the assessment practices in four subjects of the PAI cluster during the implementation of the Merdeka Curriculum at the madrasah (since 2022). The four subjects function as units of internal comparison, not four separate cases, so the cross-subject analysis is a comparison within a single case, not a multi-case study. The case context includes three learning groups (grades VII, VIII, and IX) each consisting of 8, 14, and 10 students, with one meeting per week for each subject in the PAI cluster. Data on madrasah assessment facilities were not specifically collected; descriptions of the facilities in the results section should be read as the informants' perceptions, not as an audit of the facilities. Informants were purposively selected based on three criteria: (1) permanent teachers actively teaching one of the PAI subject group subjects; (2) directly involved in the planning and

implementation of assessments; and (3) have implemented the Merdeka Curriculum in the subjects they teach. Four teachers who met these criteria each represented one subject (Table 2), ensuring that all subjects in the PAI subject group at this madrasah were represented. This number is limited, so the sufficiency of the data is considered not from the number of informants, but from three things: each subject is represented by a teacher who conducts its assessment, the interviews trace the entire assessment cycle with the same guidelines, and the response patterns can be compared among informants. Consequently, the typical practices of a subject cannot be compared with those of other teachers in the same subject, and this is acknowledged as a limitation. The identities of the informants are anonymized with initials; certification status is included as it is relevant to the variation in practice (NH, KH, and IM are certified, MM is not).

Table 2. Research informant profiles

Initial	Subject	Started teaching (year)	Certification status
NH	History of Islamic Culture (SKI)	2009	Certified
MM	Akidah Akhlak	2016	Not yet certified
KH	Fiqh	2008	Certified
IM	Al-Qur'an Hadith	2010	Certified

Data were collected through semi-structured face-to-face interviews over two weeks in September 2026, each lasting approximately 30–40 minutes, recorded with the informants' permission, and transcribed verbatim. The guidelines contain 24 main questions according to four problem formulations, namely assessment practices (seven questions), issues (eight), perceived contributing factors (six), and utilization of results (three), with the same main questions for all four informants and contextual probing questions. The analysis was conducted comparably across subjects through five stages: (1) verbatim transcription; (2) open coding to identify units of meaning relevant to the problem formulation; (3) grouping codes into categories; (4) theme formation through data reduction, data presentation, and conclusion drawing; and (5) compiling a comparison matrix across subjects to trace consistent and differing patterns. Differences in responses among informants are maintained and reported as variations, not standardized. In the results section, each finding is accompanied by the initials of the supporting informants, and findings that are sourced from only one informant are marked as indications, not patterns across informants. The validity of the data is examined through a comparison of the responses of the four informants on the same theme to assess the consistency of the information. This procedure should be understood with limitations: the four informants come from the same madrasah with relatively similar work contexts, so the similarity in their answers may reflect shared conditions or similar reporting habits, rather than evidence that the practice occurs in the classroom. This comparison therefore tests the internal consistency of the teachers' reports, not the truth of the practice. The research did not use classroom observations, document reviews, or member checking, so the findings are positioned as self-reported practices and perceptions by the teachers, and the informants' statements are distinguished from the researchers' interpretations. The research obtained permission from the madrasah and verbal consent (informed consent) from each informant regarding the purpose of the research, the use of interview results for publication, as well as the right not to answer certain questions or withdraw. Confidentiality is maintained with initials. The research was not submitted to the institutional ethics committee because the risks are considered minimal, as it involves interviews about teachers' professional practices, not personally sensitive topics**.**

3. RESULTS AND DISCUSSION

RESULTS

The results are presented according to four problem formulations. Each finding is accompanied by the initials of the supporting informants; statements sourced from only one

informant are referred to as such. Comparison with previous research is focused on the Discussion section.

Authentic Assessment Practices as Reported by Teachers

The four informants understood authentic assessment as an evaluation that does not stop at written tests but encompasses students' attitudes and skills during learning. MM articulated it, "Authentic assessment is an evaluation that looks at students' abilities in real situations during learning, especially how the knowledge learned is reflected in their attitudes and behaviors" (Interview with MM, September 7, 2026). In the planning phase, all four informants reported a uniform pattern: learning objectives were established first, followed by the aspects to be assessed. NH and MM mentioned that the instruments were prepared before or simultaneously with the learning materials (according to MM, they were often still simple), while KH and IM described the same work sequence in the practice of worship and the Qur'an and Hadith materials. In the affective domain, all four informants derived indicators from the value content of the taught material, not from a list of generic attitudes. MM exemplified the indicator of honesty from the honesty in completing tasks when the material discussed honesty, whereas KH and IM linked it to the value of verses, hadiths, or the seriousness in following practice rules. The main instrument for the affective domain in the four subjects is direct observation, with notes or simple journals (NH, MM); self-assessment is not yet routine (NH, MM) and peer assessment is not always conducted because it requires time (NH). Affective observations are accumulated from several activities because not all students can be observed in depth in one meeting. The collected evidence consists of assignment results, observation notes, and practical or presentation scores, while visual documentation has not yet been routine. The four informants did not conclude their assessments from a single incident; MM stated, "Students' attitudes are not always the same every day. So I can't just look at one incident" (Interview with MM, September 12, 2026). In the psychomotor domain, the forms of assessment, implementation, and typical challenges vary by subject (Table 3). Since the data comes from interviews without observation and document review, the five stages are understood as practices reported by the teachers.

Table 3. Comparison of psychomotor assessment practices and follow-up across subjects as reported by informants.

Subject (informant)	Focus and forms of psychomotor assessment	Implementation and specific challenges reported	Reported follow-up
SKI (NH)	Skills in delivering historical narratives: presentations, concept maps, role-playing historical figures	Presentation and simulation; striving to see individual contributions, not just group performance; group contributions are uneven.	Modeling and small group practice before the presentation; explaining again in your own words.
Akidah Akhlak (MM)	Ability to practice moral values: simulation and role-playing	Presentation and simulation; the dilemma of detailed versus practical rubrics; affective indicators are difficult to concretize.	Example of contextual cases and discussion; reinforcement of discussion rules
Fikih (KH)	The accuracy of the sequence of worship movements: ablution, prayer, funeral rites	Individual practice in turns; time can run out if evaluated one by one; limited facilities force adjustments in practice; the process is also evaluated.	Re-explanation followed by group practice; peer tutor
Al-Qur'an Hadis (IM)	Accuracy of recitation, pronunciation, fluency, and memorization	Practice individual reading in turns; adhering to the established criteria.	Gradual reading practice with individual guidance; memorization of related verses

Reported Issues and Perceived Contributing Factors

Six reported issues are spread throughout the assessment cycle, not centered on grading. Table 4 shows interview evidence for each issue along with the perceived contributing factors. Five issues are supported by more than one informant; a concrete example of the difficulty in concretizing affective indicators comes only from MM, thus it is read as an indication, as well as the second side of the instrument dilemma (an instrument that is too simple does not adequately represent the condition of the students). MM's formulation of the dilemma is, "If it's too detailed, teachers have difficulty recording all students. If it's too simple, the assessment results do not accurately reflect the students' conditions" (Interview with MM, September 12, 2026).

Table 4. Issues, interview evidence, perceived contributing factors, and their implications for assessment quality

Problematic	Interview evidence (informant)	Factors perceived to contribute	Implications for assessment quality
1. The dilemma of depth and practicality of instruments	NH, KH, IM: too many indicators or criteria make observation difficult, so the most important indicators and a simple rubric were chosen. Only MM: the instrument is too simple and does not adequately represent the students' conditions.	Structural: time and number of students	Some dimensions of competence are not captured in the rubric.
2. Difficult to concretize attitude indicators	MM: Silent students are not necessarily behaving poorly, and active students are not necessarily better, so the indicators must be clear and not based on the teacher's impression. The four informants: the teachers' constraint lies in the technical skills of formulating indicators (general statements).	Technical: the skill of formulating indicators has not yet addressed conceptual socialization	Risk of bias toward surface behavior appearance
3. The scope of the observation is not comprehensive.	The four informants: not all students can be observed in depth in one meeting (NH: observations are conducted gradually); MM: the more students there are, the more behaviors go unobserved.	Structural: interaction of time and number of students	Evidence of inconsistent attitudes among students is not documented evenly.
4. Praktik bergantian dan kontribusi kelompok tidak merata	KH: if the practice is done one by one, time can run out for assessment (Fiqh and Quran Hadith take turns); NH: striving to see the contribution of each student, not just the group's performance	Structural: duration of individual practice and number of students	Opportunities for performance are not equal; the representation of individual performance evidence is potentially diminished.
5. Weak documentation of performance evidence	All four informants: visual documentation is not yet routine; NH, MM, IM: limited photo, video, or digital media facilities; KH: limited facilities force adjustments in practice.	Facilities: limitations of documentation facilities	Traceability of performance evidence is reduced for verification and reporting.
6. Threats to the consistency of assessment decisions	MM: do not conclude from a single event; MM and IM: adhere to criteria; KH: consider the process. This strategy has not yet been supported by systematic recording instruments (researcher interpretation).	Technical: the absence of systematic recording instruments	Consistency depends on the diligence and consideration of the teacher.

The factors perceived to contribute are divided into three groups. First, structural factors: all four informants identified time constraints and the number of students as the most dominant factors. NH stated, "The main factors are time constraints and the numerous aspects that need to be assessed." Teachers have to teach the material, manage the class, and simultaneously observe the students"

(Interview with NH, September 19, 2026), and MM added that the more students there are, the greater the likelihood that some behaviors will go unobserved. Second, technical factors: all four informants mentioned that teacher competence could be a constraint, entirely on the technical skills of developing indicators, instruments, and rubrics, not on conceptual understanding. This response was given as a general statement about teachers, not a self-assessment, so its interpretation is interpretative. MM expressed the need for examples of direct application, such as how to create a rubric specifically for Akidah Akhlak, not just the socialization of concepts. Third, the factor of facilities and documentation: all four informants believe that simple assessments (observation, presentation, direct practice) can still be conducted with the existing facilities, but NH, MM, and IM explicitly mention the limitations of photo, video, or digital media documentation facilities, while KH mentions that the limited facilities force some practices to be adjusted. The term "factor" here refers to factors perceived or identified by the informants, not causal relationships empirically tested. 3. Utilization of Assessment Results Although reporting the six issues, the four informants utilized the assessment results for diagnosis, feedback, remedial or enrichment, and improvement of subsequent learning. The assessment results were interpreted as information about the students' difficulties, not as a final score (NH). Feedback was directed toward adjusting treatments, such as reinforcing discussion rules when observations showed a lack of respect for peers' opinions (MM). Follow-up actions were twofold: remedial actions included guidance and opportunities to retake, while enrichment actions included in-depth assignments, including peer tutoring (KH). The most detailed example was provided by NH, "When I saw from the presentation results that many students understood the material but were unable to convey it well, in the next lesson I did not immediately ask them to present." I provided an example first... then gave them the opportunity to practice in small groups. "After that, a presentation is conducted" (Interview NH, September 19, 2026). A similar pattern was reported by three other informants (Table 3): MM, KH, and IM changed subsequent activities after diagnosing gaps in understanding and application, the sequence of steps not yet understood, and reading fluency issues. The connection to reassessment appears explicit in NH and KH, but implicit in MM and IM. This cycle relies on simple notes and teacher observations without structured documentation of student development.

DISCUSSION

The practices reported by the teacher show a relatively clear assessment framework, but are implemented with compromises at the technical and structural levels. The finding that teachers understand the concept but are hindered by technical readiness aligns with Rosidah et al. (2021) and Andrianto et al. (2026), but differs from Prihantoro (2021) and Fadlillah and Kusaeri (2024) who place the lack of teacher understanding as the dominant obstacle. This difference is likely related to the context: both studies are cross-location and cross-level reviews, whereas this research focuses on a single madrasah that has implemented the Merdeka Curriculum since 2022 and, according to MM, has received conceptual socialization but not technical assistance per subject. Adequate conceptual understanding does not automatically guaranty the ability to translate it into indicators, rubrics, and evidence systems. The relationship between time, technical ability, instruments, and the quality of evidence can be interpreted as a sequence (researcher's interpretation). Limited meeting time and the burden on teachers who have to teach while also observing encourage teachers to choose concise instruments (Azzahra & Maksum, 2025). Without technical assistance, this conciseness seems to be achieved by reducing the criteria, rather than by designing a brief rubric that still includes achievement level descriptors (Hira et al., 2025). A concise instrument without a systematic recording format produces evidence in the form of teacher notes and final grades that are difficult to trace back. In other words, time determines how much can be recorded, technical ability determines what is chosen to be recorded, and the instrument determines how well the evidence can be accounted

for. Interestingly, the study groups in this madrasah are small (8–14 students), but the teachers still perceive the number of students as a constraint; the limited possibility is the observation time per student in one meeting per week (the duration of the meeting was not measured in this study). The simplification of instruments can affect the objectivity and reliability of assessments through four pathways. First, the rubric with criteria limited to only recording part of the competency dimensions means that the scores do not fully represent the intended achievements, as MM is aware. Second, vague affective indicators make the assessment rely on the teacher's impression. Krathwohl et al.'s (1964) affective taxonomy distinguishes five levels of value internalization (receiving, responding, valuing, organization, and characterization); indicators that only capture verbal activity or momentary compliance tend to stop at the levels of receiving and responding, whereas the intended PAI achievement is at the level of value judgment to characterization, which requires consistent behavior across time and situations. MM's awareness that a silent student is not necessarily behaving poorly can be interpreted as an intuitive effort to avoid that mistake. Third, gradual observation and alternating practice result in varying amounts and quality of evidence for each student, leading to unequal decisions among students. Fourth, without traceable documentation, decisions cannot be reexamined by another teacher or even by the same teacher later on. Repeated observations made by the teacher strengthen reliability, but are not sufficient without written criteria and systematic recording. This argument is analytical; the quality of the actual assessment decisions was not examined in this study. The most interesting finding is that all four teachers reported continuing to utilize the results of formative assessments despite facing the six issues, with concrete examples from each informant. This reflects the formative function of assessment, whose value lies in the use of information to adjust learning (Abd Halim et al., 2024). This finding differs from Rahayu and Istikomah (2024), who found that feedback and follow-up monitoring were not optimal, and is more in line with Syamsiah and Prastowo (2022) and Fadilah et al. (2025), who reported follow-up for feedback and remedial actions. In this location, the issue is not the absence of follow-up, but rather its informal nature: without structured recording of progress, continuity, and transfer of assessment information between periods or among teachers relying on the memory and individual capacity of the teachers. Fadilah et al. (2025) presented an alternative in the form of video portfolios, which in this location is hindered by perceived limited documentation facilities. This statement is the researchers' interpretation based on interview patterns and is not intended as an assessment of the entire madrasah assessment system.

Variation between subjects expands on Syamsiah and Prastowo (2022) and Fadilah et al. (2025) that psychomotor assessment techniques follow the competencies being measured: turn-taking and time constraints are more prominent in subjects oriented toward technical precision (Fiqh, Al-Qur'an Hadith), while group contribution and indicator ambiguity are more prominent in subjects oriented toward narrative and values (SKI, Akidah Akhlak). The empirical contributions of this research are thus threefold: separating the issues from the perceived contributing factors throughout the assessment cycle, demonstrating that the constraints in this area are more technical and structural than conceptual, and showing that the formative cycle can proceed informally without adequate documentation support. Practically, technical assistance cannot be standardized across subjects. The findings can be synthesized into a conceptual model based on the findings (Figure 1). Structural pressures, technical limitations, and limitations in documentation tools interact with each other and encourage the simplification of assessment practices (concise instruments, gradual observations, alternating practices, minimal documentation), which brings risks such as uneven learning evidence, low traceability of evidence, and decreased consistency and reliability of assessment decisions. The use of formative assessment continues, but informally (dashed line). This model organizes the relationships between themes in the qualitative data of this research and is not a causal model that is generally applicable or ready to be generalized.

ISSN 2776-5229

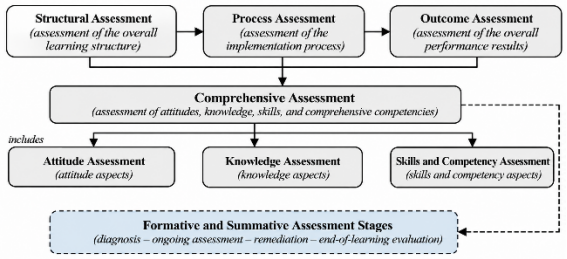


Figure 1. Conceptual Model Based on Research Findings (relationships identified from qualitative data, not universally applicable causal relationships)

The limitation of the research lies in the use of interviews as the sole data collection technique with four informants in one madrasa, without classroom observation, document review, or member checking; assessment facility data were also not specifically collected. The findings are therefore interpreted as the experiences and practices reported by teachers and intended as a basis for analytical generalization, not statistical. Further research can verify the findings through triangulation of techniques and sources, such as multisite studies that combine classroom observations and document analysis.

4. CONCLUSION

Based on the report of four PAI subject teachers at MTs Nahdlatul Ulum, authentic assessment in the affective and psychomotor domains is practiced through five stages, namely planning based on learning objectives, determining indicators according to the character of the material, selecting instruments and rubrics, observation or performance, and collecting learning evidence. The utilization of assessment results is the sixth stage that connects the evaluation process with the follow-up of learning. Six issues were reported throughout the cycle, namely the limited depth of the instruments, difficulties in concretizing affective indicators, limited scope of observation, unequal opportunities for practice and contribution, weak documentation of performance evidence, and potential disruptions to the consistency of assessment decisions. According to the informants' perception, these issues are related to the limitations of time and the number of students, the technical skills in developing instruments and rubrics, as well as documentation facilities. The assessment results are still reported to be used for feedback, remediation, enrichment, and improvement of learning, although the follow-up documentation remains informal. The main finding is that the issue of authentic assessment is not only related to teachers' understanding of the assessment concept but also to technical skills, time constraints, instruments, and documentation of learning evidence. The implication is that teachers need practical-based technical assistance in developing indicators and instruments for each subject, including the development of concise rubrics that still contain achievement descriptors. Madrasah administrators need to provide a simple yet structured assessment documentation system, such as an observation record format that can be accumulated across meetings and selective performance documentation, without adding administrative burden. Because the research only used interviews at one madrasah, further research needs to verify the findings through classroom observations and document reviews in a broader context.

5. REFERENCES

Abd Halim, H., Hamzah, M. I., & Zulkifli, H. (2024). A systematic review on the formative assessment practice in teaching and learning in secondary school. *International Journal of Evaluation and Research in Education, 13*(2), 1173–1183. https://doi.org/10.11591/ijere.v13i2.26507

Andrianto, S. R., Rahmi, Nur'hang, Setiawan, M. R., & Erniati. (2026). Readiness of PAI teachers to develop and use psychomotor assessment instruments in the Independent Curriculum at MTs Darul Ulum Toili. *Sosioedukasi: Jurnal Ilmiah Ilmu Pendidikan dan Sosial, 15*(3), 989–999. https://doi.org/10.36526/sosioedukasi.v15i3.9125

Azzahra, F., & Maksum, M. N. R. (2025). Implementation of evaluation of learning in the psychomotor domain of Islamic religious education and ethics at SMPIT Ar-Risalah Sukoharjo. Journal of Educational Sciences, 9(2), 827–834. https://doi.org/10.31258/jes.9.2.p.827-834

Badan Standar, Kurikulum, dan Asesmen Pendidikan. (2024). *Panduan pembelajaran dan asesmen pendidikan anak usia dini, pendidikan dasar, dan pendidikan menengah: Edisi revisi tahun 2024*. Kementerian Pendidikan, Kebudayaan, Riset, dan Teknologi.

Baroroh, U., & Hamani, T. (2022). Development of authentic assessment in Islamic religious education in elementary school. Nazhruna: Jurnal Pendidikan Islam, 5(3), 940–955. https://doi.org/10.31538/nzh.v5i3.2380

Fadilah, S. N. F., Anwar, S., & Nugraha, R. H. (2025). Implementation of psychomotor assessment and teacher readiness in Islamic religious education learning at senior high schools in Bandung City. *Ta'dib: Jurnal Pendidikan Islam, 30*(2), 251–270. https://doi.org/10.19109/9nm3pr27

Fadlillah, N., & Kusaeri. (2024). Optimizing assessment for learning in Islamic education through authentic and diagnostic assessment: A systematic literature review. *Jurnal Kependidikan: Jurnal Hasil Penelitian dan Kajian Kepustakaan di Bidang Pendidikan, Pengajaran dan Pembelajaran, 10*(2), 654–667. https://doi.org/10.33394/jk.v10i2.11555

Forisma, A., Ni'mah, Z., & Sukiman. (2023). Teknik dan instrumen asesmen keterampilan Pendidikan Agama Islam di Dikdasmen dan pendidikan tinggi. Jurnal Evaluasi Pendidikan, 14(1), 17–24. https://doi.org/10.21009/jep.v14i1.36741

Hazyimara, K., Nurasia, N., & Itsnainy, S. K. (2024). Implementation of affective domain assessment in Islamic religious education learning at SDN Rampal Celaket 2 Malang. *MANAZHIM, 6*(1), 64–77. https://doi.org/10.36088/manazhim.v6i1.4191

Hira, H. H., Ernawati, E., Praptama, R., Andhika, M. R., & Driana, E. (2025). Challenges and teacher strategies in implementing authentic assessment: A case study at a state Islamic junior high school. Al-Fikrah: Jurnal Manajemen Pendidikan, 13(2), 160–174. https://doi.org/10.31958/jaf.v13i2.16021

Imanudin, B. A. (2023). Penilaian autentik terhadap pembelajaran Pendidikan Agama Islam. Raudhah Proud To Be Professionals: Jurnal Tarbiyah Islamiyah, 8(2), 527–539. https://doi.org/10.48094/raudhah.v8i2.310

Indriani, A., Anwar, S., & Budiyanti, N. (2025). The implementation of affective assessment in Islamic religious education at elementary schools: Between concept, practice, and challenges. *Tarbiyah Wa Ta'lim: Jurnal Penelitian Pendidikan dan Pembelajaran, 12*(3), 289–305. https://doi.org/10.21093/twt.v12i3.11904

Ismail, F., Astuti, M., Nasrudin, D., Wigati, I., & Zuhdiyah. (2022). Assessment model of Islamic religious education in the psychomotor domain during the COVID-19 pandemic in Indonesia. Eurasian Journal of Educational Research, 101, 190–203. https://doi.org/10.14689/ejer.2022.101.011

Ismail, R., Erawadi, E., & Zulhammi, Z. (2023). Implementasi penilaian autentik pada mata pelajaran Al-Qur'an Hadits di MAN 1 Tapanuli Tengah. Islamika, 5(3), 1293–1310. https://doi.org/10.36088/islamika.v5i3.3683

Krathwohl, D. R., Bloom, B. S., & Masia, B. B. (1964). *Taxonomy of educational objectives: The classification of educational goals. Handbook II: Affective domain*. David McKay.

Nawali, J., Zuhriyah, I. A., Susilawati, S., & Yaqin, A. Z. N. (2024). Implementasi penilaian autentik di SDI Surya Buana Malang untuk meningkatkan kualitas pembelajaran. Pendas: Jurnal Ilmiah Pendidikan Dasar, 9(4), 232–245. https://doi.org/10.23969/jp.v9i04.20886

Prihantoro, A. (2021). Kegagalan pelaksanaan asesmen autentik pada pelajaran pendidikan agama Islam di sekolah dasar dan menengah. *Ulumuddin: Jurnal Ilmu-Ilmu Keislaman, 11*(1), 29–54. https://doi.org/10.47200/ulumuddin.v11i1.688

Rahayu, F., & Istikomah. (2024). Implementation of Merdeka Curriculum in Islamic religious education learning: Analysis of principles and assessment models. *Ta'dib: Jurnal Pendidikan Islam, 29*(1), 149–161. https://doi.org/10.19109/td.v29i1.23688

Rosidah, C. T., Pramulia, P., & Susiloningsih, W. (2021). Analisis kesiapan guru mengimplementasikan asesmen autentik dalam Kurikulum Merdeka Belajar. *Jurnal Pendidikan Dasar, 12*(1), 87–103. https://doi.org/10.21009/jpd.v12i01.21159

Rozikin, K., Suffandi, Y., Amrullah, M. B., & Khoiroh, H. (2026). Developing an integrated model of authentic assessment in Islamic religious education (PAI): Cognitive, affective, and psychomotor dimensions. *Al Mikraj: Jurnal Studi Islam dan Humaniora, 6*(1), 306–318. https://doi.org/10.37680/almikraj.v6i01.10510

Rustam, R., Luthfiyah, L., Ilham, I., & Syarifuddin, S. (2025). Harmonisasi asesmen kognitif, afektif, dan psikomotorik dalam pembelajaran pendidikan agama Islam: Analisis kesenjangan antara regulasi kurikulum dan praktik di kelas. *Pendas: Jurnal Ilmiah Pendidikan Dasar, 10*(4). https://doi.org/10.23969/jp.v10i04.39102

Sahlan, M., Abqoriy, M., & Mawidha, R. F. (2023). Implementation of authentic assessment in Islamic cultural history subject within the independent curriculum. *Scaffolding: Jurnal Pendidikan Islam dan Multikulturalisme, 5*(3), 831–845. https://doi.org/10.37680/scaffolding.v5i3.3865

Sirinding, A., Kania, D., Muzaqih, M. F., Ahmad, N. A., & Khoiriyah, S. N. (2023). Analisis kesulitan guru dalam penerapan penilaian autentik pada Kurikulum Merdeka di sekolah dasar. Cendikia: Jurnal Pendidikan dan Pengajaran, 1(3), 160–167. https://doi.org/10.572349/cendikia.v1i3.392

Syamsiah, P. N., & Prastowo, A. (2022). Implementasi penilaian autentik pada ranah psikomotorik dalam pembelajaran pendidikan agama Islam di sekolah dasar. *Journal of Research and Thought on Islamic Education (JRTIE), 5*(2), 130–143. https://doi.org/10.24260/jrtie.v5i2.2204

Wijaya, A., Andriani, R. T., Ismady, M. W., Suwarni, W. S. D., & Ahzim, R. (2023). The implementation of affective and psychomotor assessment in Islamic religious education learning in junior high schools. TARBAWY: Indonesian Journal of Islamic Education, 10(2), 195–206. https://doi.org/10.17509/e1mq4j87