Copyright © 2025 The Author(s)
Didaktika: Jurnal Kependidikan Vol. 14 No. 4 Nopember 2025

Assessing Akhlak Internalization in Islamic Religious Education: An Integrative Review of Indicators, Techniques, and Evidence Quality

Andi Riawarda[1], Sitti Maryam Gani[2], Aydi Syam[3], Suarni[4]
[1]Universitas Islam Negeri Palopo, Indonesia
[2,3]IAI DDI Mangkoso, Indonesia
[4]Universitas Muhammadiyah Kendari, Indonesia

Abstract

Assessment in Islamic Religious Education (IRE) should measure more than the acquisition of knowledge; it should also provide defensible information about students' developing attitudes and conduct. Assessing akhlak is challenging because it is an abstract construct expressed across varied contexts and cannot be inferred reliably from a single incident or instrument. This article synthesizes concepts, indicators, techniques, and evidence quality in the assessment of akhlak internalization in IRE learning. An integrative review was conducted across empirical studies, instrument-development research, conceptual reviews, and relevant methodological sources. Searches covered Google Scholar, GARUDA, ERIC, DOAJ, and journal and publisher websites using Indonesian and English terms related to IRE, akhlak, affective assessment, authentic assessment, and character assessment. Sources were selected for their relevance to Islamic education, clarity of method or argument, and contribution to indicators, techniques, validity, reliability, or assessment use. The synthesis of 24 sources—17 focused on IRE, akhlak, or character assessment and seven methodological or conceptual sources—yielded four findings: akhlak indicators should be translated into observable conduct without equating a single act with an inner state; assessment evidence is stronger when collected repeatedly and from multiple sources; rubrics, observation, self-reflection, portfolios, and family communication serve complementary functions; and assessment quality depends on the validity of interpretations, consistency of judgment, contextual fairness, and purposeful use of results. The article proposes a layered assessment framework connecting values, indicators, evidence, interpretation, feedback, and follow-up. The framework positions assessment as part of tarbiyah and learning rather than as a means of morally labeling students.

Keywords: affective assessment; akhlak; Islamic religious education; authentic assessment; validity

Introduction

Islamic Religious Education (IRE) pursues aims that extend beyond the acquisition of religious information. Students are expected to understand religious teachings, develop religious awareness, and cultivate practices that reflect Islamic values in their relationships with God, themselves, other people, and the environment. In Islamic educational thought, personal development encompasses interconnected intellectual, moral, social, and individual dimensions (Halstead, 2004). IRE assessment should therefore provide information about learning that is broader than written examination scores.

A challenge arises when schools need to translate the goal of akhlak formation into assessment indicators. Honesty, responsibility, discipline, care, adab, and respect are not always directly observable. Conduct may be shaped by circumstances, power relations,

opportunities, habits, and environmental support. A student who commits no violation while being observed has not necessarily internalized a value, while one mistake is insufficient to conclude that the student lacks good akhlak. When teachers assign scores without defining indicators and observation contexts, assessment can become a general impression or a judgment of personality.

This issue is evident in studies of attitude assessment in elementary IRE. Mahmudah and Shafrizal (2022) found that attitude assessment requires attention from both families and schools, while the assumption that moral formation is solely the school's responsibility may weaken the continuity of student development. Research on affective assessment in Akidah Akhlak has shown that self-assessment instruments can be developed and tested for content validity, construct structure, and reliability (Rohmad et al., 2023). These findings indicate that affective assessment can be designed more systematically, although an instrument validated in one context cannot automatically be used in another without adaptation.

Research on authentic assessment in IRE has addressed the use of cognitive, affective, and psychomotor assessment, including written tests, assignments, observation, and projects (Amalia, 2023). Other studies have examined the combination of test and non-test methods in IRE evaluation and the need for more student-centered approaches (Musanna et al., 2025). In 2026, Mijdal et al. (2026) synthesized affective assessment instruments in IRE, while Falahain and Munip (2026) proposed triangulating performance evidence to reduce social desirability bias in self-assessment. These recent studies make clear that this article does not claim to be the first review of affective assessment or triangulation. Its contribution lies in bringing together the relationships among internalization indicators, the functions of assessment techniques, the strength of evidence, the limits of interpretation, and instructional follow-up in one analytical framework. Research on affective assessment under the Merdeka Curriculum has also reported challenges involving teachers' understanding, their skill in using instruments, limited time, and the availability of training (Susmiyati et al., 2023). A study in an Islamic elementary school documented the use of observation, incident records, and self-assessment through religious activity books, but identified constraints related to class size, time, and parental involvement (Azaliya, 2025). Meanwhile, research on assessment as, for, and of learning in IRE directs attention to assessment functions that go beyond measuring outcomes to support and involve students in the learning process (Nurlitasari & Hamami, 2023).

Although these studies provide an important foundation, research on IRE assessment often groups techniques without sufficiently explaining how behavioral evidence can support an inference about akhlak internalization. Assessing performance on a single task, recording recurring observable behavior, and making a claim about relatively enduring character are different activities. Each type of claim requires different evidence and a different degree of caution in interpretation. General literature on authentic assessment emphasizes alignment among the task, the conditions under which it is performed, the criteria, and the real-world practice being assessed (Gulikers et al., 2004). Formative assessment literature also treats evidence as a basis for feedback and instructional adjustment, rather than merely as a reporting score (Carless & Boud, 2018).

This article addresses a gap in the synthesis of four aspects: the conceptualization of akhlak internalization; the translation of values into behavioral indicators; the collection and combination of evidence; and the quality of assessment interpretation and follow-up. Its contribution is to place evidence quality and the limits of moral inference at the center of the discussion. Accordingly, the review asks: (1) How is akhlak internalization conceptualized and translated into IRE assessment indicators? (2) What techniques and sources of evidence are

Copyright © 2025 The Author(s)

used to assess this development? and (3) What principles should be upheld so that assessment results are valid, fair, and useful for learning? The article aims to synthesize findings from the literature and formulate an assessment framework that can inform the development of IRE practices and instruments.

Method

This article uses an integrative review to combine studies with diverse designs and conceptual sources that help explain the assessment of akhlak internalization. An integrative review is appropriate for questions requiring a comprehensive understanding through the synthesis of empirical evidence and theoretical thought, provided that the processes of searching, evaluation, analysis, and presentation are transparent (Whittemore & Knafl, 2005). Review writing should also produce a synthesis that advances a framework for understanding, rather than merely summarizing sources one by one (Torraco, 2005; Snyder, 2019).

A targeted literature search was conducted on October 22, 2026, using Google Scholar, GARUDA, ERIC, DOAJ, and journal and publisher websites. The final corpus comprised 24 sources: 17 studies directly addressing IRE, akhlak, affective assessment, or character assessment, and seven methodological or conceptual sources used to interpret evidence quality and the review process. Twenty sources were published between 2016 and 2026; four older sources were retained because they provide foundational references on integrative review methodology, Islamic educational thought, and authentic assessment. Indonesian search terms included “Pendidikan Agama Islam,” “asesmen akhlak,” “penilaian sikap PAI,” “asesmen afektif,” “asesmen autentik,” and “penilaian karakter.” English equivalents included “Islamic Religious Education,” “akhlak assessment,” “affective assessment,” “authentic assessment,” “attitude assessment,” and “character assessment.” Terms were combined using AND/OR operators, for example, (“Islamic Religious Education” OR PAI) AND (akhlak OR character OR attitude) AND (assessment OR evaluation). Reference lists and citations of relevant articles were also searched.

Sources were considered if they were identifiable journal articles or scholarly proceedings, addressed assessment in Islamic education or relevant character measurement, and provided information about concepts, indicators, techniques, validation processes, challenges, or the use of results. Classic methodological and theoretical sources were included as an analytical foundation, even when published more than ten years ago. Sources addressing cognitive achievement without a connection to affective assessment, opinions without traceable argumentation, and records with unverifiable metadata were not used as a basis for findings. Because the purpose of this review is conceptual synthesis, the article does not claim to provide a quantitative mapping of all publications; it does not report total search results or a flow diagram without auditable screening records.

Each relevant source was read using an extraction form covering its purpose, educational context, design, concepts or indicators of akhlak, evidence-collection techniques, assessors, validity or reliability information, challenges, and implications. Source evaluation considered the alignment between the research question and design, clarity of the subject or context, transparency of procedures, sufficiency of evidence for the conclusions, and limitations acknowledged by the authors. Instrument-development studies were examined for content and construct validity, reliability, respondent characteristics, and limits on instrument use. Qualitative studies were assessed for contextual clarity, triangulation, and the fit between conclusions and data. Conceptual studies were assessed for conceptual clarity and argumentative consistency.

The analysis was thematic. First, sources were read and information was grouped by construct, indicators, techniques, evidence, quality, challenges, and use of results. Second, similarities and differences across sources were compared, and empirical evidence was distinguished from conceptual recommendations. Third, themes addressing the review questions were developed and relationships among them were formulated. This process follows principles of integrative analysis: data from different designs should be reduced, displayed, compared, and interpreted without assuming that every type of evidence has the same strength (Whittemore & Knafl, 2005).

Results

Conceptualization and Indicators of Akhlak Internalization

The literature shows that akhlak assessment lies between two important considerations. On the one hand, moral and spiritual values are central aims of Islamic education. On the other hand, inner qualities cannot be observed directly or fully measured through a single response or event. Therefore, the most defensible object of assessment is not a student's "level of faith" or "personal quality," but evidence of learning relevant to behavioral indicators and reflection that have been explained in advance.

Islamic educational thought connects moral development with personal growth, social life, and the pursuit of knowledge (Halstead, 2004). In IRE learning, this relationship can be translated into indicators close to students' learning experiences: communicating information honestly, fulfilling task responsibilities, respecting others, honoring group agreements, responding to disagreement with adab, and reflecting on the consequences of one's actions. Such indicators do not replace spiritual aims; rather, they limit assessment claims to aspects that can be observed and developed through learning.

Research on Akidah Akhlak instrument development shows that spiritual and social attitude dimensions can be organized into indicators and examined psychometrically. Rohmad et al. (2023) developed a self-assessment instrument and reported support for content and construct validity, along with relationships among indicators of akhlak toward God and the Prophet and akhlak toward parents, teachers, oneself, and others. This finding suggests that indicators can be designed as related constructs rather than as a set of disconnected behaviors. However, construct validity does not establish that a score can be used to make a comprehensive conclusion about a student's personality.

Studies of attitude assessment in elementary schools draw on observations and information from several parties, including teachers, parents, and peers (Mahmudah & Shafrizal, 2022). This multi-source design reflects the fact that student behavior does not occur only in the classroom. Character-instrument research outside the specific IRE context has also used school culture as a basis for developing indicators, allowing assessment to account for behavioral expectations and routines in educational settings (Hidayah & Setiawati, 2022). These sources support the need to distinguish general indicators—such as responsibility—from specific behaviors shaped by age, task, and context.

Thematically, indicators in the literature can be grouped into four areas. First is responsibility toward oneself and one's tasks, such as honoring agreements, safeguarding entrusted responsibilities, and acknowledging mistakes. Second is adab in social relationships, including respect, courtesy, empathy, cooperation, and attitudes toward difference. Third is discipline and consistency of practice, which should be assessed proportionately to learning objectives and developmental stage. Fourth is reflective capacity: explaining moral reasoning, recognizing the consequences of actions, and planning

Copyright © 2025 The Author(s)

improvement. This grouping is an analytical synthesis offered by this article, not a standardized taxonomy validated across all educational levels.

Assessment Techniques and Sources of Evidence

The reviewed sources indicate the use of combined test and non-test techniques. Written or oral tests can assess understanding of moral concepts, religious arguments, and the reasoning students use when facing dilemmas. However, test results alone do not demonstrate that knowledge has become habitual conduct. Research on authentic assessment in IRE identifies written and oral assessments, assignments, projects, observation, and performance assessment as forms that may be combined according to the competencies being assessed (Amalia, 2023). Studies of IRE evaluation also recommend balancing test and non-test techniques so that assessment does not stop at cognitive achievement (Musanna et al., 2025).

Observation is an important technique because it allows teachers to record behavior during learning activities. For observation to be useful, it requires defined indicators, adequate observation contexts and time spans, and specific behavioral notes. Labels such as “good,” “undisciplined,” or “rude,” without examples of incidents, do not provide sufficient evidence for later review. Research on IRE teachers’ experiences reports the use of observation checklists, self-assessment scales, and portfolios, while also noting limitations in the models used (Maemunah et al., 2023). This indicates that multimodal practices exist, but that the quality of their design and interpretation needs further development.

Self-assessment can help students compare their conduct with criteria, recognize their development, and set goals for improvement. Rohmad et al. (2023) show that a self-assessment instrument for Akidah Akhlak can be developed and tested. However, self-assessment should not be treated as a single objective report. Responses may be influenced by language comprehension, the desire to make a favorable impression, psychological safety, and reflective ability. General educational literature recommends that self-assessment focus on the quality of work or action in relation to known criteria and be followed by an opportunity to improve (Panadero et al., 2017). Self-reflection is therefore better understood as a learning resource and one form of evidence, not as a tool for measuring inner states.

Portfolios and reflective journals can collect evidence over time. They may contain records of practice, reflections on situations, feedback received, and plans for follow-up. Performance tasks and projects can provide authentic contexts for observing responsibility, cooperation, honesty in reporting processes, and ways of resolving disagreement. Frameworks for authentic assessment emphasize that authenticity involves more than the format of a task; it also concerns the alignment of the task, context, process, outcome, and assessment criteria with meaningful practice (Gulikers et al., 2004). A project with an Islamic theme, therefore, is not necessarily an authentic assessment of akhlak if only its final product is assessed.

Information from parents and peers can broaden the picture of behavior, as illustrated by multi-source approaches in studies of IRE attitude assessment in elementary schools (Mahmudah & Shafrizal, 2022). Such information requires a clear purpose, student consent and protection, and caution regarding relational conflicts. Family assessment should not turn the school into a monitor of household behavior, and peer assessment should not become a vote on a student’s character. Multi-source information is used to enrich context and examine the consistency of patterns, not to count majority opinions.

Evidence Quality and Limits of Interpretation

Assessment quality depends on the accuracy of the relationships among constructs, indicators, tasks, evidence, and conclusions. Instrument-development studies show that

content validity, construct structure, and reliability should be examined before an instrument is used (Rohmad et al., 2023). The development of an honest-affect assessment instrument for junior high school students also involved expert validation, readability testing, and reliability testing, and reported that some initial items did not meet the criteria (Hikmawati & Hasnawati, 2022). Such findings affirm that the fit of an assessed value cannot be established merely because a statement "sounds Islamic."

Validity does not reside in an instrument as an absolute property; it concerns the interpretation and use of scores. An instrument suitable for prompting classroom reflection may not be appropriate for ranking students or determining sanctions. Sari et al. (2026) emphasize the need to connect empirical validity with interpretive and ethical dimensions in the context of IRE. Mijdal et al. (2026) also position justice, honesty, and responsibility as considerations in developing affective assessment instruments. These requirements are especially important in akhlak assessment because scores may affect students' self-image, teacher–student relationships, and school decisions.

Consistency in assessment requires criteria that teachers and students can understand. Descriptive rubrics are more informative than moral labels or single scores because they explain the behavior underlying each level of attainment. Teachers need to distinguish between having had no opportunity to demonstrate a behavior, not yet having observed it, and not yet having achieved it. These categories have different meanings; combining them as a low score may produce an unfair conclusion. Assessing rare or sensitive behaviors requires more opportunities for observation and adequate contextual support.

Studies of challenges in affective IRE assessment under the Merdeka Curriculum report limitations in teachers' understanding, instrument-design skills, time, training, and resources (Susmiyati et al., 2023). An implementation study in elementary schools also highlights the need for simple and adaptable instruments, consistent use, stronger teacher capacity, and policy support (Indriani et al., 2025). Assessment challenges therefore concern not only the choice of technique but also the working conditions that determine whether teachers can routinely collect evidence and provide meaningful feedback.

Using Assessment Results to Support Development

Formative assessment literature treats evidence as information for determining the next steps in learning. Effective feedback helps students understand goals, recognize the gap between their performance and the criteria, and identify actions they can take. Carless and Boud (2018) emphasize that students need the capacity to understand feedback, make judgments, manage emotional responses, and take action. Self-assessment can support learning monitoring when criteria and reflective processes are taught, rather than when students are simply asked to score themselves (Panadero et al., 2017). Research on assessment as, for, and of learning in IRE reinforces the need to connect assessment with learning, reflection, and reporting (Nurlitasari & Hamami, 2023).

Useful akhlak assessment produces actionable information: which behaviors have been consistent, in which situations students still need support, and what learning experiences may help their further development. Reports should identify evidence and context rather than assign labels such as "problem student" or "not religious enough." Summative assessment may summarize development, but numerical scores or categories should not replace descriptions explaining the basis for conclusions. Research on summative IRE assessment under the Merdeka Curriculum emphasizes the need to plan competencies, instruments, scales, result processing, reporting, and evaluation (Fitria et al., 2024).

Copyright © 2025 The Author(s)

Discussion

The synthesis indicates that assessment of akhlak internalization should be understood as evidence-based reasoning that takes place throughout learning. Teachers do not observe internalization directly; they interpret students' conduct, reflection, and responses to situations using relevant indicators. Because such inferences are uncertain, the language of assessment results should limit the claims being made. Stating that a student "demonstrated responsibility in three observed activities" is more defensible than describing the student generally as "a responsible person." This distinction does not diminish the goal of character formation; it protects the integrity of assessment.

Within Islamic educational thought, limiting claims can coexist with the aims of tarbiyah. Islamic education regards moral development as part of individual growth and social life, rather than as an outcome of a single subject that can be fully assessed at the end of a semester (Halstead, 2004). IRE assessment should therefore be positioned as one space for guidance and reflection. It can observe actions related to learning, provide feedback, and encourage improvement, but it cannot measure a person's entire religious quality. The value of amanah in assessment requires teachers not to claim more than the available evidence supports.

Findings on affective instruments and self-assessment offer an important methodological lesson. The development study by Rohmad et al. (2023) shows that conduct and attitudes can be operationalized and tested, while Hikmawati and Hasnawati (2022) show that instrument items need to be examined and may fail during testing. An instrument is a tool for clarifying observation; it is not a machine that automatically converts a moral concept into an objective score. Its quality depends on construct definitions, language appropriate to students' ages, opportunities for the behavior to occur, administration procedures, and the interpretation of results. Even a strong reliability coefficient does not guarantee that the measured construct is appropriate or that the use of scores is fair.

The use of multiple techniques should be based on the function of evidence, not on a desire to collect as many forms as possible. Falahain and Munip (2026) specifically propose performance-evidence triangulation to examine socially desirable responses in religious self-assessment. This proposal is relevant, but it should be applied proportionately: documenting personal habits must not become surveillance that exceeds learning purposes. The framework in this article extends this discussion by positioning triangulation as one layer of evidence, alongside clear indicators, task context, interpretation, and follow-up. Oral or written tests can show students' understanding of moral reasoning. Observation can show conduct in real activities. Self-reflection can reveal how students understand their choices and consequences. Portfolios can show change over time. Family input may provide information about other contexts, provided that it is collected ethically. Combining these sources allows for triangulation, but does not require every source to yield the same score. Differences between students' reflections and teachers' observations can become a basis for dialogue and contextual understanding, rather than proof that one party must be wrong.

Principles of authentic assessment extend this argument. Authenticity is not merely the use of projects or real-life activities. Tasks must provide a reasonable opportunity to demonstrate the intended competence, criteria must relate to the quality of actions, and assessment situations must not demand irrelevant behavior. In IRE, a social project may assess responsibility and care if its process is documented and individual roles are visible. If only the group product is assessed, each member's conduct may be obscured. The five-

dimensional framework of Gulikers et al. (2004) helps teachers examine tasks, context, form, criteria, and assessment standards in an integrated way.

Formative assessment also changes the purpose for which data are used. Behavioral records should not be collected only to complete report cards. Specific feedback can help students understand their choices and determine steps for improvement. For example, instead of writing "undisciplined," a teacher could explain that a student has repeatedly failed to bring the materials agreed upon for a discussion, then agree on a reminder strategy and review progress during the next activity. This approach is consistent with formative assessment, which uses evidence to determine the next steps in learning. Carless and Boud (2018) emphasize students' capacity to understand feedback, make judgments, manage responses, and take action. The meta-analysis by Panadero et al. (2017) also indicates that self-assessment can support self-regulated learning and self-efficacy, particularly when it is designed as part of the learning process and accompanied by criteria students can use.

Fairness is a fundamental concern because opportunities to demonstrate behavior are shaped by conditions that are not always under students' control. Differences in special educational needs, family background, social experiences, language, culture, and access to activities may affect the evidence that becomes visible. Teachers need to ensure that indicators do not penalize different communication styles or cultural practices when learning objectives are still being met. Assessment of religious practices, for example, requires sensitivity to privacy and students' circumstances; not every personally meaningful practice is appropriate for direct observation in the classroom. The principle of 'adl requires transparent criteria and comparable opportunities, while amanah requires confidentiality and limited use of data.

Families and communities can strengthen continuity in student development, but their roles and authority must be clearly bounded. Mahmudah and Shafrizal (2022) show that families' perceptions of responsibility for moral development affect school–home relationships. However, involving families does not transfer the burden of assessment to parents. Schools may request limited information about relevant behavior, share developmental goals, and invite dialogue where contexts differ. Peer assessment should likewise be designed as feedback on specific cooperation or interaction, not as a ranking of moral character. Protection against stigma and social conflict must be part of the design.

Based on these themes, this article proposes a layered assessment framework. The first layer is the value and learning objective, such as amanah or care as a focus of learning. The second is the behavioral and reflective indicators specified for students' ages and task contexts. The third is evidence collected through an appropriate combination of techniques. The fourth is interpretation, which considers frequency, context, opportunity, consistency, and evidence sources. The fifth is feedback and follow-up in the form of learning experiences or needed support. This framework is not a standardized instrument; it is a conceptual tool to help teachers maintain coherence among objectives, evidence, and decisions.

The framework has implications for instrument design. Each indicator should use observable verbs, specify a context or opportunity, and avoid overly broad trait labels. Rubrics should describe behavior at each level rather than provide only numerical ranges. Observation forms should make room for incident notes and dates, while self-assessment can prompt reflection on examples of conduct and plans for improvement. Instruments should be tested for readability and content fit by IRE experts, assessment specialists, teachers, and—depending on age—students. When scores are used for high-stakes decisions, stronger validity and consistency evidence is required than for formative classroom use.

Copyright © 2025 The Author(s)

Implications for teacher professional development extend beyond providing a list of techniques. Teachers need opportunities to practice developing indicators, collecting factual records, distinguishing observation from interpretation, providing feedback, and discussing potentially biased assessment cases. Challenges reported by Susmiyati et al. (2023) and Indriani et al. (2025) show that assessment quality is related to time, class size, instrument availability, and policy support. Schools can reduce workload by selecting a small number of priority indicators for each learning unit, using simple digital records that protect data, and integrating evidence collection into activities already taking place. Additional paperwork without an instructional purpose may turn assessment into a formality.

The theoretical contribution of this review is to shift the question from "Which technique is used?" to "What conclusions are supported by the evidence collected?" This shift matters because affective assessment is often treated as if an abstract construct could be measured directly with a scale. The article connects Islamic education literature with theories of authentic and formative assessment and validity, while emphasizing that values must be interpreted in context. Its practical contribution is an assessment sequence that teachers and researchers can use to design assessment that is informative, non-punitive, and actionable.

This review has limitations. The search was targeted to Indonesian- and English-language sources and relied on metadata and texts available on journal or publisher websites. The article does not conduct a meta-analysis, estimate the causal effects of any technique, or calculate the prevalence of instrument use in schools. The range of terms—akhlak, character, spiritual attitude, social attitude, morality, and affective development—also means that the boundaries of the corpus depend on conceptual decisions. In addition, evidence from instrument-development studies, case studies, and conceptual reviews varies in strength. Future research should report search and screening strategies in greater detail, test the layered framework across educational levels, and examine the validity and fairness of assessment use among diverse student groups.

Conclusion

Assessment of akhlak internalization in IRE is most defensible when treated as evidence-based developmental assessment, rather than direct measurement of inner states or labeling of personality. The synthesis indicates that values should be translated into contextual behavioral and reflective indicators; evidence should be collected repeatedly through complementary techniques; and conclusions should account for validity, consistency, opportunity, fairness, and the limits of inference. Assessment results should inform feedback and student development, not only numerical reporting. A layered framework linking values, indicators, evidence, interpretation, and follow-up can help teachers maintain this alignment. Because this review is a conceptual synthesis and does not test the framework in the field, empirical research is needed to examine its usability, validity, and impact across diverse school and madrasah contexts.

References

Amalia, R. (2023). Authentic assessment dalam pembelajaran Pendidikan Agama Islam. *Risalah: Jurnal Pendidikan dan Studi Islam, 9*(4), 1467–1477. https://doi.org/10.31943/jurnal_risalah.v9i4.568

Azaliya, V. K. (2025). Affective assessment in Islamic Religious Education at MI Al-Huda Ploso Nganjuk. *Proceeding International Conference on Islam, Law, and Society, 4*(1). https://doi.org/10.70062/incoils.v4i1.276

Carless, D., & Boud, D. (2018). The development of student feedback literacy: Enabling uptake of feedback. *Assessment & Evaluation in Higher Education, 43*(8), 1315–1325. https://doi.org/10.1080/02602938.2018.1463354

Falahain, M., & Munip, A. (2026). Addressing social bias and enhancing the validity of student self-assessment with a performance evidence-based triangulation model in Islamic Religious Education learning. *Journal of Islamic Education and Ethics, 4*(2), 160–175. https://doi.org/10.18196/jiee.v4i2.170

Faizah, U., Zuchdi, D., & Alsamiri, Y. (2019). An authentic assessment model to assess kindergarten students' character. *Research and Evaluation in Education, 5*(2), 103–119. https://doi.org/10.21831/reid.v5i2.24588

Fitria, N. N., Mufidah, L. L. N., & Setiawati, P. (2024). Summative assessment of Islamic Education subject in Merdeka Curriculum. *Journal of Educational Research and Practice, 2*(3), 328–338. https://doi.org/10.70376/jerp.v2i3.157

Gulikers, J. T. M., Bastiaens, T. J., & Kirschner, P. A. (2004). A five-dimensional framework for authentic assessment. *Educational Technology Research and Development, 52*(3), 67–86. https://doi.org/10.1007/BF02504676

Halstead, J. M. (2004). An Islamic concept of education. *Comparative Education, 40*(4), 517–529. https://doi.org/10.1080/0305006042000284510

Hidayah, M. A., & Setiawati, F. A. (2022). Developing a character assessment instrument based on the school culture. *Research and Evaluation in Education, 8*(2), 90–99. https://doi.org/10.21831/reid.v8i2.46802

Hikmawati, N., & Hasnawati, H. (2022). Development of students' honest affective assessment instruments at the junior high school level as the implementation of character education. *Arisen: Assessment and Research on Education, 4*(2). https://doi.org/10.33292/arisen.v4i2.237

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). https://doi.org/10.21093/twt.v12i3.11904

Maemunah, M., Muttaqin, S., Rahim, H., Rosyada, D., Zamhari, A., & Kultsum, U. (2023). Insights and experiences of Indonesian senior high school teachers in implementing affective assessment model in Islamic Education subject. In *Proceedings of the 5th International Graduate Conference in Islam and Interdisciplinary Studies (IGCIIS 2022)*. EAI. https://doi.org/10.4108/eai.19-10-2022.2329045

Mahmudah, S., & Shafrizal, A. (2022). Analysis of attitude assessment in Islamic Religious Education in elementary schools. *Pedagogik Journal of Islamic Elementary School, 5*(2), 171–184. https://doi.org/10.24256/pijies.v5i2.2940

Mardiah, D., & Hamami, T. (2026). Pengembangan assessment as, for, dan of learning dalam pembelajaran Pendidikan Agama Islam: Studi literatur. *Edu Cendikia: Jurnal Ilmiah Kependidikan, 6*(1), 381–390. https://doi.org/10.47709/educendikia.v6i01.8123

Mijdal, B. S., Hamami, T., Aula, I. A., Anggraini, A., & Sulvaningsih, E. (2026). Developing affective assessment instruments in Islamic religious education. *Curricula: Journal of Curriculum Development, 5*(2), 847–866. https://doi.org/10.17509/curricula.v5i2.681

Musanna, S., Jamaluddin, & Duskri, M. (2025). Test and non-test evaluation methods in Islamic Religious Education learning in senior high school: A critical analysis based on literature review to strengthen student-centered learning. *Darussalam: Scientific Journal of Islamic Education, 2*(1), 8–13. https://doi.org/10.62945/darussalam.v2i1.752

Copyright © 2025 The Author(s)

Nurlitasari, A., & Hamami, T. (2023). Assessment as, for, dan of learning dalam pembelajaran Pendidikan Agama Islam tingkat menengah atas. *Risalah: Jurnal Pendidikan dan Studi Islam, 9*(4), 1556–1567. https://doi.org/10.31943/jurnal_risalah.v9i4.597

Panadero, E., Jonsson, A., & Botella, J. (2017). Effects of self-assessment on self-regulated learning and self-efficacy: Four meta-analyses. *Educational Research Review, 22*, 74–98. https://doi.org/10.1016/j.edurev.2017.08.004

Rohmad, R., Dharin, A., & Azis, D. K. (2023). Developing self-assessment instruments of affective domain on belief and morality (Aqidah Akhlak) subject in Madrasah Tsanawiyah. *Pegem Journal of Education and Instruction, 13*(1), 193–201. https://doi.org/10.47750/pegegog.13.01.21

Snyder, H. (2019). Literature review as a research methodology: An overview and guidelines. *Journal of Business Research, 104*, 333–339. https://doi.org/10.1016/j.jbusres.2019.07.039

Sari, E. Y. P., Shohib, M. W., & Anshori, A. (2026). The validity of authentic Islamic value-based assessment instruments in contemporary PAI education: A conceptual analysis. In *Proceeding of the International Summit on Science, Technology, and Humanity (ISETH)*. https://doi.org/10.23917/iseth.6920

Susmiyati, S., Zurqoni, Z., Abdillah, M. H., & Saugi, W. (2023). Challenges of affective assessment of Islamic Religious Education learning in Merdeka Curriculum. *Al-Hayat: Journal of Islamic Education, 7*(2), 710–731. https://doi.org/10.35723/ajie.v7i2.675

Torraco, R. J. (2005). Writing integrative literature reviews: Guidelines and examples. *Human Resource Development Review, 4*(3), 356–367. https://doi.org/10.1177/1534484305278283

Whittemore, R., & Knafl, K. (2005). The integrative review: Updated methodology. *Journal of Advanced Nursing, 52*(5), 546–553. https://doi.org/10.1111/j.1365-2648.2005.03621.x

---This page has been intentionally left blank---