Edu Global: Jurnal Pendidikan Islam
E-ISSN: 2747-2442
URL: https://journal-stainmadina.web.id/edu-global/index

OPEN ACCESS
Vol. 7 No. 1
July 2026

Peer -Reviewed

Research Article

Is Content Validation Enough? Evidence for Aqidah Akhlak Test in Madrasah

Syaukani,[1] Nurmawati,[2] Muhammad Faisal,[3] ✉ Radinal Mukhtar Harahap,[4] Jamiah Hariyati.[5]

[1,2,3] UIN Sumatera Utara Medan
[4] STIT Ar-Raudlatul Hasanah Medan
[5] Universitas Tjut Nyak Dien, Medan

Article Info
Received : 02 May 2026
Revised: 29 May 2026
Accepted: 03 June 2026

Keywords:
content validation, item analysis, Aqidah Akhlak, evidence-based assessment, test quality

Corresponding Author
Muhammad Faisal
muhammad0335253011@gmail.com

ABSTRACT: This study aims to evaluate the quality of a teacher-made Aqidah Akhlak test by examining whether content validation alone is sufficient to ensure assessment quality. The study employed a quantitative evaluative design integrating logical and empirical analysis. Logical analysis was conducted through expert judgment assessing content relevance, construction, and language clarity, while empirical analysis was performed using Classical Test Theory based on response data from 30 Grade VIII students at MTsN 2 Madina. The instrument consisted of 30 multiple-choice items and 5 essay items. The findings indicate that most items were considered acceptable by expert validators in terms of content validity. However, empirical analysis revealed that the test demonstrated moderate reliability (0.654), was dominated by easy items, and included several items with low discrimination power and non-functioning distractors. On the other hand, pre-test and post-test data indicated that the instrument was capable of capturing general learning improvement, although its ability to differentiate variations in student ability remained limited due to the dominance of easy items and the imbalance of item difficulty levels. Furthermore, essay items were found to be more effective in measuring higher-order competencies such as moral reasoning and value reflection. These results confirm that content validation alone is insufficient to ensure test quality. This study proposes a multi-evidence validation model integrating expert judgment, item analysis, and learning outcome evaluation. Practically, the findings highlight the need to strengthen teachers' assessment literacy and to implement data-driven quality assurance systems in madrasah assessment practices.

INTRODUCTION

Assessment quality is a fundamental component of educational systems because assessment results are used to support instructional decision-making, determine levels of competency attainment, improve learning processes, and evaluate curriculum effectiveness [1], [2], [3]. Poor-quality assessment instruments may generate biased, inaccurate, and potentially harmful decisions for students. For this reason, issues of validity and reliability remain central concerns in contemporary educational evaluation research [4], [5] [6].

In many schools, particularly at the secondary level, assessment practices continue to rely heavily on teacher-made tests, namely instruments developed by teachers for daily quizzes, mid-semester examinations, and end-of-semester assessments [7]. Such instruments offer substantial practical value because they are designed in accordance with classroom instruction, curricular targets, and student characteristics. Nevertheless, the psychometric quality of teacher-developed tests is not always assured, as they are often constructed without systematic procedures of instrument development and validation [8], [9].

Within school practice, the most common procedure used to ensure test quality is content validation through expert review. Teachers typically submit test manuscripts to senior teachers, subject-teacher association leaders, principals, or supervisors for evaluation in terms of content relevance, item construction, and language clarity. Once the instrument is considered appropriate, it is

https://doi.org/10.56874/edb.v7i1.106

subsequently administered for formal assessment purposes. Although this procedure is important, content validation primarily demonstrates alignment between test items and curriculum objectives or learning indicators [10] [11]. However, such procedures do not automatically demonstrate that items function effectively in actual measurement contexts.

Educational measurement literature [12], [13], [14] has consistently shown that instrument quality cannot be established solely through content alignment. An item that appears substantively appropriate may still have excessively high or low difficulty, weak discrimination power, non-functioning distractors, or limited capacity to distinguish high-performing from low-performing students. Consequently, judgments regarding test quality require empirical evidence derived from actual student response data [15]. Furthermore, assessment instruments may suffer from construct underrepresentation when important dimensions of the intended construct are insufficiently measured, despite appearing substantively appropriate during expert review.

This issue is particularly relevant in the subject of Aqidah Akhlak within Islamic schools and madrasahs. The subject is intended not only to assess mastery of faith-related concepts, but also to evaluate moral understanding, social attitudes, and the ability to apply Islamic values in everyday life. However, in practice, many Aqidah Akhlak tests continue to focus primarily on memorisation of terminology, definitions, or factual religious knowledge [16], [17]. As a result, high test scores do not necessarily indicate deep religious understanding or mature character development. From the perspective of contemporary validity theory, such conditions raise concerns regarding construct representation because important dimensions of moral reasoning, ethical reflection, and value application may remain insufficiently assessed.

Despite its practical significance, research directly comparing expert-based content validation with the empirical performance of teacher-made Aqidah Akhlak tests remains limited , particularly within the Indonesian madrasah context [11], [18], [19]. This gap is important because madrasahs routinely use internally developed examinations as the basis for evaluating student learning outcomes, yet limited evidence is available regarding the extent to which such instruments function effectively in practice [20], [21]. Addressing this issue is important for at least three reasons. *First*, teacher-made assessments remain the dominant form of classroom evaluation in many schools, especially in resource-constrained contexts where access to externally standardised instruments is limited. *Second*, religious education assessments carry broader implications than conventional academic subjects because they may influence judgments concerning students' values, discipline, and moral development. *Third*, if school-based instruments are accepted merely because they have passed expert review, significant measurement weaknesses may remain undetected. Inadequate assessment quality in religious education may therefore produce misleading interpretations regarding students' moral and spiritual development.

From a theoretical perspective, this study is informed by contemporary validity theory, which conceptualises validity not as an inherent property of a test itself, but as the degree to which evidence supports interpretations made from test scores [10]. According to Messick's unified validity framework, validity refers to an integrated evaluative judgment concerning the extent to which empirical evidence and theoretical rationales support the adequacy and appropriateness of interpretations and actions based on test scores. Under this perspective, content evidence is important but insufficient when standing alone. Additional evidence related to internal consistency, item functioning, score patterns, response processes, and decision consequences is needed to support defensible use of assessment results. Accordingly, comparing logical review and empirical item performance provides a practical framework for examining whether school-based validation practices are adequate [15] [12]. This perspective also implies that expert judgment should not be treated as final proof of validity, but rather as one component within a broader evidence-based validation process.

Based on this background, the present study aims to evaluate the quality of an Aqidah Akhlak test through two complementary approaches: logical analysis and empirical analysis. Specifically, the study seeks to examine: *first*, the extent to which test items meet content appropriateness standards according to expert judgment; *second*, the statistical quality of items based on student response data; *third*, the degree of consistency between expert judgments and empirical findings; and *fourth*, the implications of these findings for strengthening assessment systems in madrasahs.

This study is expected to contribute both theoretically and practically. Theoretically, it extends discussions of school-based assessment quality into the underexplored domain of Islamic education. More specifically, the study contributes to the application of contemporary validity theory within the context of teacher-made religious education assessments. Practically, the findings may assist teachers, principals, supervisors, and curriculum developers in strengthening evidence-based assessment practices and improving the quality of internally developed tests. More broadly, the study may encourage a shift from administrative approval models of test validation toward more comprehensive quality assurance practices.

Therefore, the central question guiding this study is: *Is content validation alone sufficient to guarantee the quality of Aqidah Akhlak tests?*

METHOD

This study employed a quantitative evaluative design within an instrument validation and item analysis framework to examine the quality of a teacher-made Aqidah Akhlak test administered in an Indonesian madrasah. The design was selected because the purpose of the study was not to measure instructional effectiveness, but rather to evaluate whether the assessment instrument already in operational use functioned adequately in terms of content appropriateness, item construction quality, and empirical measurement performance. To achieve this objective, the study integrated two complementary approaches, namely logical analysis and empirical analysis. In this study, logical analysis refers to expert-based qualitative review focusing on content relevance, construction quality, and linguistic clarity of test items, whereas empirical analysis examined the statistical functioning of items based on students' actual responses. The integration of these approaches enabled the study to obtain multiple sources of evidence concerning instrument quality. Within the framework of contemporary validity theory, empirical item analysis was treated as a source of validity evidence concerning internal structure and

item functioning rather than merely a technical procedure for score calculation.

The study was conducted at MTsN 2 Madina during the second semester of the ongoing academic year. The research data were derived from an Aqidah Akhlak assessment administered to Grade VIII students as part of the school assessment programme. The study was undertaken over one complete assessment cycle, including document collection, expert review, data coding, statistical processing, interpretation of findings, and verification of results. Because the study utilised naturally occurring school assessment data, the research context reflected routine classroom assessment practices within the participating madrasah.

The target population consisted of all Grade VIII students who participated in the Aqidah Akhlak examination at the madrasah. A total sampling technique was employed, whereby all students with complete response records were included in the analysis. Based on the available examination data, the final sample comprised 30 students. The use of total sampling was considered appropriate because the primary objective was to evaluate the functioning of the actual instrument used within the school rather than to generalise findings to a broader population. Accordingly, the empirical findings of this study should be interpreted as context-specific evidence regarding instrument quality in the participating madrasah.

The instrument analysed in this study was a teacher-developed test package consisting of 30 multiple-choice items and 5 essay items. The multiple-choice section employed five response alternatives (A–E), whereas the essay section required students to explain concepts, compare ideas, evaluate behaviours, and formulate responses to moral and religious situations. The content domains covered belief in revealed scriptures, belief in prophets, virtuous conduct, reprehensible behaviour, and social ethics. These domains were developed by the teacher based on the Grade VIII Aqidah Akhlak curriculum and semester learning targets.

Data were collected through three complementary procedures. First, document analysis was conducted on the test manuscript, answer key, scoring rubric, test blueprint, and official score records. This stage was intended to identify the structure of the instrument, content distribution, and scoring procedures. Second, expert judgment was conducted through validators consisting of senior teachers, madrasah supervisors, and academics with expertise in Islamic education or educational evaluation. The validators reviewed each item using a structured validation form covering content relevance, construction quality, language clarity, and alignment with learning objectives. Third, student response data were obtained from the official examination and subsequently coded for statistical analysis.

The principal analytical tool employed in this study was a spreadsheet-based Classical Test Theory (CTT) item analysis system designed to process student responses automatically and generate classical test statistics. The system produced indicators including test reliability, item difficulty, discrimination index, point-biserial correlation, option distribution, distractor effectiveness, and item-quality classification. The use of spreadsheet-assisted analysis improved analytical consistency, efficiency, and accuracy while reducing potential manual calculation errors. Because the study was conducted within an authentic school setting involving a relatively small sample, Classical Test Theory procedures were considered more appropriate and practically feasible than more complex psychometric models. Essay responses were scored using predetermined scoring criteria derived from the school rubric prior to descriptive analysis in order to maintain scoring consistency.

Prior to the study, the instrument had been developed through routine school procedures, including blueprint preparation, item writing, internal teacher review, answer-key preparation, and administration during the official examination schedule. The present study did not intervene in the original item construction process; rather, it evaluated the quality of the instrument after operational use. This approach was important because it reflected common school practice in which teacher-made tests are frequently administered before undergoing rigorous psychometric evaluation.

Logical analysis data were processed using inter-rater agreement procedures. Empirical analysis was conducted using Classical Test Theory procedures. *First*, test reliability was examined to estimate the internal consistency of the multiple-choice section. *Second*, item difficulty was calculated based on the proportion of students answering each item correctly. Very high values were interpreted as excessively easy items, whereas very low values indicated excessively difficult items. *Third*, item discrimination was analysed using point-biserial or biserial correlations to determine the extent to which each item differentiated between high- and low-performing students. Negative or near-zero values indicated problematic items. *Fourth*, distractor effectiveness was examined through the distribution and selection frequency of incorrect response options. Distractors were considered functional when they attracted responses from a meaningful proportion of students, particularly lower-performing students, indicating that the alternatives were sufficiently plausible within the cognitive context of the item. Conversely, distractors that were never selected or selected only by a negligible proportion of respondents were interpreted as evidence of weak option plausibility and limited item functioning. Distractor analysis was included because ineffective alternatives may reduce item discrimination and weaken the instrument's ability to differentiate varying levels of student understanding.

The essay section was analysed using descriptive statistics including mean scores, score range, standard deviation, and distribution of achievement across items. In addition, students' written responses were reviewed to determine whether the essay prompts successfully elicited conceptual explanation, moral reasoning, and value application in real-life contexts, which represent important competencies in Aqidah Akhlak learning.

To address the objectives of the study, findings from logical and empirical analyses were compared at the item level. Items judged acceptable by experts but demonstrating weak statistical performance were treated as evidence that content validation alone may be insufficient. Conversely, items showing satisfactory statistical functioning but receiving substantive criticism from experts were considered in need of content revision. Through this evidence-based comparative procedure, decisions regarding instrument quality were based on multiple sources of validity evidence rather than a single form of validation.

Several procedures were implemented to strengthen the methodological rigor and credibility of the findings. The credibility of the findings was strengthened

through the integration of multiple sources of evidence, including document analysis, expert judgment, and student response analysis. Additional rigor was established through the involvement of validators from different professional backgrounds. All student identities were anonymised prior to analysis, and examination data were used solely for academic purposes while maintaining institutional confidentiality.

Given the relatively small sample size, the empirical findings of this study should be interpreted as exploratory and context-bound evidence rather than population-level psychometric generalisations. Nevertheless, the study provides important insight into how teacher-made religious education tests function in authentic school settings and offers a practical model of evidence-based quality assurance for assessment systems in madrasahs.

RESULTS

This section presents the findings of the study strengthened by pre-test and post-test data, together with their interpretation in addressing the central research question: *Is content validation alone sufficient to guarantee the quality of Aqidah Akhlak tests?* The analysis integrates logical evidence obtained from expert judgment, item analysis results, and evidence of student learning outcomes. The instrument consisted of 30 multiple-choice items and 5 essay items administered to 30 Grade VIII students at MTsN 2 Madina.

Overall Instrument Quality

The analysis indicated that the test obtained a reliability coefficient of 0.654**,** which can be classified as moderate. This value suggests that the instrument was sufficiently consistent for routine classroom assessment, although it was not yet ideal for more consequential academic decisions. Moderate reliability commonly reflects uneven item quality, such as overly easy items, weak discrimination power, or ineffective distractors.

From a practical perspective, the instrument was usable within the context of classroom evaluation in madrasahs, but still required systematic revision in order to produce more stable and accurate scores. This finding strengthens the argument that tests passing content validation are not automatically high-quality measurement instruments.

From the perspective of contemporary validity theory, the moderate reliability coefficient also indicates that the internal structure of the instrument had not yet fully supported consistent score interpretation. In other words, although the instrument demonstrated acceptable curricular alignment according to expert judgment, the empirical consistency of student responses remained only partially satisfactory. This finding supports the argument that validity claims require empirical evidence beyond content appropriateness alone.

Table 1. General Test Statistics

Indicator	Value
Number of Students	30
Multiple-Choice Items	30
Essay Items	5
Reliability	0.654
Interpretation	Moderate

Comparison of Pre-Test and Post-Test Results

The data demonstrated a positive improvement in student achievement after the instructional process. In the pre-test, most students achieved scores within the middle range, whereas in the post-test a larger number of students reached high scores. Several students showed substantial gains, for example from 86 to 98, from 84 to 96, and from 90 to 98.

Students who initially obtained lower scores also showed progress, although not as strongly as higher-performing students. This pattern indicates that the instrument was capable of detecting changes in student competence following instruction.

This finding is important because test quality should not only be judged by content relevance, but also by the ability of the instrument to reflect variations in student achievement after learning activities. A useful classroom assessment instrument should provide score patterns that meaningfully correspond to instructional outcomes.

However, the interpretation of score improvement in this study should be approached cautiously. The increase in post-test scores may indeed reflect successful learning outcomes, yet the dominance of easy items potentially inflated score gains and reduced score dispersion among higher-performing students. Consequently, although the instrument demonstrated practical responsiveness to classroom learning, its capacity to measure nuanced differences in higher-order mastery remained limited.

Table 2. Examples of Pre-Test and Post-Test Changes

Student	Pre-Test	Post-Test	Change
Student 1	86	98	+12
Student 3	90	90	High stable performance
Student 5	84	96	+12
Student 9	86	98	+12

Overall, the upward trend in scores suggests that the instrument was sufficiently responsive to learning improvement.

Item Difficulty Distribution

The difficulty analysis showed that most items were classified as easy**.** Many items had proportions correct above 0.85, and several approached 1.00. For example, some post-test items were answered correctly by 96.7% of students.

The dominance of easy items helps explain why student scores were generally high in the post-test. Although this may partly reflect successful learning, from a psychometric perspective an excessive number of easy items reduces the test's ability to differentiate high-performing students from average-performing students.

In the context of Aqidah Akhlak, easy items often focus on factual recall, conceptual definitions, or identification of religious terms. If the aim is to assess broader religious competencies, a greater proportion of moderate and difficult items requiring moral analysis and value application is needed.

The graphical distribution of correct responses further confirmed this tendency. Several content areas, such as "Kitab Taurat Nabi Musa," "Sifat tabligh rasul," and "Pentingnya amanah," showed correct-response percentages approaching 100%. While these results may indicate successful classroom instruction, they

simultaneously suggest limited item challenge and restricted measurement variance.

Table 3. Item Difficulty Distribution

Category	Interpretation
Easy	Dominant
Moderate	Some items
Difficult	Very few

Item Discrimination Power

Several items demonstrated strong discrimination power and successfully distinguished high-achieving from low-achieving students. However, some items showed zero or very weak discrimination. For instance, an item with a facility index of 0.967 but a point-biserial correlation of 0.000 indicated that nearly all students answered correctly, making the item unable to differentiate student ability.

Such items may appear acceptable in terms of content, yet function weakly statistically. This provides direct evidence that expert validation alone is insufficient to guarantee item quality. Questions that seem substantively sound must still be tested using actual student response data.

Conversely, some moderately difficult items showed stronger discrimination indices, confirming that medium-difficulty items tend to function more effectively as measurement tools.

These findings are important because discrimination indices represent empirical evidence concerning the internal functioning of test items. Within modern validity frameworks, items that fail to differentiate between students with varying levels of mastery weaken the interpretability of total test scores. Thus, statistical item functioning becomes an essential complement to expert-based logical review.

Distractor Effectiveness

The analysis of response options revealed that several distractors were not selected at all. In some items, almost all students selected the correct answer, while alternative responses recorded frequencies of zero. For example, in Item 18, all students selected option B as the correct answer, whereas options A, C, D, and E received zero responses.

Non-functioning distractors make items excessively easy and increase the probability of guessing the correct answer. This weakens test quality because not all options operate meaningfully from a psychometric standpoint.

For future item development, teachers should construct distractors that are more plausible, closely related to common misconceptions, and attractive to students who have not yet mastered the content.

The presence of non-functioning distractors also indicates that some items were constructed with insufficient consideration of students' potential misunderstanding patterns. In classroom assessment contexts, distractors should not merely serve as formal alternatives, but should function diagnostically by revealing different levels of conceptual understanding.

Performance of Essay Items

The essay section made an important contribution in measuring higher-order competencies. The five essay questions covered topics such as differences between scriptures and suhuf, the wisdom of believing in Allah's books, miracles and karomah, the negative effects of ghibah, and programmes for developing noble character.

Unlike multiple-choice items, essay questions required students to explain concepts, justify answers, and relate religious content to real-life contexts. In the post-test, several students achieved high essay scores, indicating improvement in elaborative thinking skills.

These findings suggest that, for Aqidah Akhlak, essay formats are more suitable for assessing affective-cognitive dimensions such as moral reasoning, value reflection, and religious argumentation.

This finding is theoretically important because the construct of Aqidah Akhlak extends beyond factual religious knowledge toward ethical reasoning and value internalisation. Consequently, assessment formats that allow explanation, interpretation, and contextual judgement may provide richer validity evidence than objective items focused primarily on recognition and recall.

Consistency Between Logical Validation and Empirical Evidence

Expert validators previously judged most items acceptable in terms of content, construction, and language. However, empirical results showed that several items remained too easy, unable to discriminate students, or supported by weak distractors.

Therefore, the consistency between logical validation and empirical evidence was partial rather than absolute. Content validation successfully ensured curricular alignment, but did not guarantee satisfactory statistical performance when the test was implemented in practice.

This finding supports modern validity theory, which argues that instrument quality must be established through multiple sources of evidence, rather than relying on a single type of validation.

More importantly, the findings indicate that school-based validation practices in madrasahs still tend to operate within an administrative conception of validity, where expert approval is often treated as sufficient evidence of instrument quality. The present study demonstrates that such practices may overlook important psychometric weaknesses that only emerge through empirical analysis of student responses.

Main Implications of the Study

Overall, the findings indicate that the Aqidah Akhlak test possessed acceptable quality for classroom use, as reflected in moderate reliability and its ability to reflect observable learning gains through pre-test and post-test comparison. Nevertheless, the psychometric quality of several items still requires improvement.

These findings directly answer the research question: *Content validation alone is not enough.*

Expert review is important for ensuring content relevance, but it cannot independently detect:

1. overly easy items,
2. weak discrimination power,
3. ineffective distractors, and
4. the sensitivity of the instrument to learning improvement.

Therefore, quality assurance systems in madrasahs should integrate:

1. expert validation before test administration,
2. post-administration item analysis, and

3. gain-score evaluation through pre-test and post-test procedures.

The study therefore proposes a shift from validation practices centred primarily on administrative approval toward evidence-based assessment culture in madrasahs. Under such a framework, assessment quality should be established through the integration of substantive review, empirical evidence, and continuous item revision. This approach would support the development of more accurate, fair, meaningful, and defensible assessment systems within Islamic education.

CONCLUSION

This study demonstrates that content validation alone is insufficient to guarantee the quality of teacher-made Aqidah Akhlak assessment instruments. Although expert validators generally judged the test items to be appropriate in terms of content relevance, construction, and language clarity, empirical analysis revealed several psychometric weaknesses that remained undetected during logical review. The findings showed that the instrument achieved only moderate reliability, contained a substantial proportion of overly easy items, included several weak or non-discriminating items, and displayed multiple distractors that failed to function effectively. These results indicate that curricular alignment and expert approval do not automatically ensure adequate measurement performance when instruments are implemented in authentic classroom contexts.

The study further found that empirical item analysis provides important sources of validity evidence beyond content representation alone. Item difficulty, discrimination indices, distractor effectiveness, and score consistency offered critical insights into how the instrument actually functioned in measuring student ability. Several items that appeared acceptable during expert review were empirically unable to differentiate high-performing and low-performing students, demonstrating the limitations of relying exclusively on logical validation procedures. These findings support contemporary validity theory, which conceptualises validity as an evidence-based argument derived from multiple sources rather than as an inherent property of a test.

An additional contribution of this study lies in the integration of pre-test and post-test evidence into the evaluation of assessment quality. The instrument was able to detect general improvement in student learning outcomes after instruction, indicating a degree of responsiveness to learning change. However, the dominance of easy items reduced the sensitivity of the test in distinguishing variations in higher-level understanding. This finding suggests that assessment quality in religious education should not only be evaluated in terms of content appropriateness, but also in terms of the instrument's capacity to capture meaningful differences and developmental progress in student learning.

The study also highlights the important role of essay items in Aqidah Akhlak assessment. Compared with multiple-choice items, essay questions were more capable of eliciting moral reasoning, conceptual explanation, value reflection, and contextual application of Islamic teachings. This suggests that complex competencies in Islamic education cannot be adequately represented through factual recall items alone. Therefore, balanced integration between objective and constructed-response formats is necessary to support more authentic assessment practices.

Theoretically, this research extends the application of contemporary validity theory into the relatively underexplored field of Islamic education assessment. The study contributes empirical evidence showing that school-based religious education tests require multi-source validation procedures similar to those recommended in mainstream educational measurement literature. Practically, the findings provide a realistic model for strengthening assessment quality assurance in madrasahs through the integration of expert judgment, empirical item analysis, and evaluation of learning responsiveness.

Based on these findings, several recommendations may be proposed. *First*, teachers should be encouraged to conduct post-administration item analysis routinely rather than relying solely on expert validation before test implementation. *Second*, professional development programmes for madrasah teachers should strengthen assessment literacy, particularly in relation to Classical Test Theory, distractor construction, and evidence-based validation practices. *Third*, school supervisors and curriculum developers should promote institutional systems that integrate logical and empirical analysis as complementary components of assessment quality assurance. *Finally*, future studies may expand this research using larger samples, broader school contexts, or more advanced psychometric approaches such as Item Response Theory to obtain stronger generalisability and deeper evidence regarding the quality of Islamic education assessments.

Ultimately, this study shows that ensuring the quality of Aqidah Akhlak assessment is not merely a technical matter of preparing curriculum-aligned questions, but a broader responsibility to ensure that instruments used to evaluate faith, morality, and character development genuinely produce accurate, meaningful, and defensible interpretations of student learning.

REFERENCES

[1] C. M. Callahan, "Evaluation for decision-making: The practitioner's guide to program evaluation," in *Systems and models for developing programs for the gifted and talented*, Routledge, 2023, pp. 119–142, https://doi.org/10.4324/9781003419426.

[2] D. Summers, "Teachers' use of assessment data to improve student achievement," *Literature Reviews in Education and Human Services*, vol. 2, no. 2, pp. 21–49, 2023.

[3] C. Wyatt-Smith, L. Adie, and L. Harris, "Supporting teacher judgement and decision-making: Using focused analysis to help teachers see students, learning, and quality in assessment data," *British Educational Research Journal*, vol. 50, no. 3, pp. 1420–1448, 2024, https://doi.org/10.1002/berj.3984Digital Object Identifier (DOI).

[4] E. Buldu and Ç. Ö. Şendil, "A critical point about early childhood assessments: Validity and reliability issues in teachers' formative assessment," *Yaşadıkça Eğitim*, vol. 37, no. 1, pp. 253–268, 2023, https://doi.org/10.33308/26674874.2023371507.

[5] R. Meylani, "A comparative analysis of traditional and modern approaches to assessment and evaluation in education," *Batı anadolu eğitim bilimleri dergisi*, vol.

15, no. 1, pp. 520–555, 2024, https://doi.org/10.51460/baebd.1386737.
[6] A. C. Amalia, "Analysis of Reliability and Validity of Islamic Cultural History (SKI) PAT Items in Madrasah Ibtidaiyah Surabaya," *JURNAL ILMU PENDIDIKAN & SOSIAL (SINOVA)*, vol. 3, no. 3, pp. 195–206, 2025, https://doi.org/10.71382/sinova.v3i3.303.
[7] J. Subando, E. Muslimin, M. Fatimah, and A. E. Rochmawan, "Pemetaan Kemampuan Siswa dan Kualitas Butir Soal Aqidah di Madrasah Aliyah," *Wiyata Dharma: Jurnal Penelitian dan Evaluasi Pendidikan*, vol. 13, no. 1, pp. 74–95, 2025, https://doi.org/10.30738/wd.v13i1.18231
[8] L. Judijanto *et al.*, *Assessment, Testing dan Evaluasi*. PT. Sonpedia Publishing Indonesia, 2025.
[9] I. W. Widiana, I. K. Gading, I. M. Tegeh, and P. A. Antara, *Validasi penyusunan instrumen penelitian pendidikan*. PT. RajaGrafindo Persada-Rajawali Pers, 2023.
[10] M. Ulfah, D. Darmansyah, and R. Rehani, "Instrumen Pengujian Produk Pembelajaran (Pengujian Validitas, Praktikalitas, Efektivitas)," *At-Tarbiyah: Jurnal Penelitian Dan Pendidikan Agama Islam*, vol. 3, no. 1, pp. 43–51, 2025. [access]
[11] A. Hasna, "Analisis Pendidikan Islam terhadap Evaluasi Pembelajaran Akidah Akhlak," *Jurnal Manajemen Islam*, vol. 1, no. 2, pp. 216–231, 2024. [access]
[12] W. Arbeni, A. Windiani, D. S. B. Sihotang, N. Anggraini, S. Wulandari, and A. Nugroho, "Test reliability analysis in educational evaluation: a quantitative approach to consistency and validity," *Holistic Science*, vol. 5, no. 1, pp. 59–64, 2025, https://doi.org/10.56495/hs.v5i1.838
[13] J. Fischer, M. Bearman, D. Boud, and J. Tai, "How does assessment drive learning? A focus on students' development of evaluative judgement," *Assessment & Evaluation in Higher Education*, vol. 49, no. 2, pp. 233–245, 2024.
[14] N. Musfirah, N. Nurbaya, N. Nursalam, and A. Rasyid, "Pengembangan instrumen hasil penilaian belajar tes dan non tes," *Socius: Jurnal Penelitian Ilmu-Ilmu Sosial*, vol. 2, no. 11, 2025, https://doi.org/10.5281/zenodo.15534085.
[15] T. Siregar, "Mengapa Validitas dan Reliabilitas Penting dalam L&D (Pembelajaran & Pengembangan): Apa yang Dapat Anda Lakukan Tentang Hal Ini," *JURNAL PEMIKIRAN DAN PENGEMBANGAN PEMBELAJARAN*, vol. 7, no. 1, pp. 1–12, 2025. https://doi.org/10.31970/pendidikan.v7i1.463
[16] S. Hamdi and M. Muslimah, "The Dilemma of Applying Authentic Assessment to Aqidah Akhlak Subjects," *Nazhruna: Jurnal Pendidikan Islam*, vol. 5, no. 3, pp. 1091–1104, 2022. https://doi.org/10.31538/nzh.v5i3.2211
[17] A. M. Afifah, "Analysis of the Akidah Akhlak Textbook to Strengthen Students' Moral Character and Spiritual Values," *Journal of Islamic Education*, vol. 10, no. 2, pp. 640–662, 2025. https://doi.org/10.35723/jie.v10i2.643
[18] S. Anggraini, "Pengembangan Butir Soal HOT's dalam Kegiatan Evaluasi Pembelajaran Mata Pelajaran Akidah Akhlak di MTS Al-Ahsan Bogor," Skripsi, UNUSIA, Bogor, 2022. https://repository.unusia.ac.id/id/eprint/543/
[19] D. Darodjat and D. Zuchdi, "Model evaluasi pembelajaran akidah dan akhlak di Madrasah Tsanawiyah (MTs)," *Jurnal Penelitian dan Evaluasi Pendidikan*, vol. 20, no. 1, pp. 11–26, 2016.
[20] A. Maria and N. Indriyani, "Pengaruh Penggunaan Alat Evaluasi Sikap Terhadap Hasil Belajar Afektif Siswa Pada Mata Pelajaran Akidah Akhlak," *MASAGI: Jurnal Pendidikan Agama Islam*, vol. 2, no. 1, pp. 303–309, 2023, https://doi.org/10.37968/masagi.v2i1.570
[21] M. Nashihah and M. P. I. Eliyanto, "Evaluasi Pembelajaran Akidah Akhlak Di MTs Negeri 2 Kebumen," 2024, *Tesis*
[22] J. Hill, K. Ogle, M. Gottlieb, S. A. Santen, and A. R. Artino Jr, "Educator's blueprint: A how-to guide for collecting validity evidence in survey- based research," *AEM Education and Training*, vol. 6, no. 6, p. e10835, Dec. 2022, https://doi.org/10.1002/aet2.10835Digital Object Identifier (DOI)