Taqrib : Journal Of Islamic Studies
Vol. 4 No. 2 2026, 322-332
Publisher: CV. Doki Course and Training
E-ISSN: 2988-4497 | P-ISSN: 2988-4160

Anderson-Krathwohl Revised Taxonomy Alignment in Islamic Religious Education Assessment at MTS Nurul Yaqin

Agustin Suci Rizkianti*, Muamar Sidik, Nana Suryana, Rudi Iskandar, Sohwatun Hanifah, Noor Azida Batubara
Sekolah Tinggi Agama Islam Haji Agus Salim Cikarang Bekasi
Corresponding email: rizkiantiagustinsuci@gmail.com

ARTICLE INFO

Article History
Submission : 12-04-2026
Received : 20-07-2026
Revised : 05-08-2026
Accepted : 11-08-2026

Keywords
Anderson-Krathwohl revised taxonomy
Islamic Religious Education
Learning assessment
Cognitive alignment
HOTS

ABSTRACT

This article analyzes the alignment of Islamic Religious Education assessment with Anderson-Krathwohl's revised taxonomy at MTs Nurul Yaqin, North Cikarang. It addresses the gap between teachers' recognition of C1-C6 terminology and their ability to translate the taxonomy into assessment items with clear cognitive demand. This qualitative single-case study involved three Islamic Religious Education teachers and one head of madrasah. Data were collected in the second semester of the 2025/2026 academic year through four interviews, two observation sessions, and document analysis of 14 assessment-related documents containing 160 items. The analysis used a coding matrix of cognitive process, knowledge dimension, stimulus, expected response, and scoring criteria. The findings show that teachers had sufficient declarative taxonomy knowledge but weaker operational assessment literacy. Document coding identified 62 items at C1, 50 at C2, 28 at C3, 12 at C4, 6 at C5, and 2 at C6. The main mismatch occurred when blueprints used higher-order labels, but final items only required definitions or textbook information. C6 tasks were also often treated as product assignments without analytic rubrics. The study contributes a four-layer mismatch typology: objective, stimulus, response, and scoring alignment.

Introduction

Islamic Religious Education assessment in madrasah should not only measure students' ability to recall verses, terms, rules, figures, and definitions. It also needs to examine how students interpret religious meaning, apply Islamic values, analyze moral problems, evaluate behavior according to Islamic principles, and create relevant religious responses. This expectation is important at the junior secondary level because students face digital communication, peer interaction, academic pressure, and diverse social norms. Anderson-Krathwohl's revised taxonomy offers a useful analytical framework because it separates six cognitive processes from four types of knowledge (Anderson & Krathwohl, 2001). The six processes are remembering, understanding, applying, analyzing, evaluating, and creating, while the knowledge dimensions include factual, conceptual, procedural, and metacognitive knowledge. In Islamic Religious Education, this matrix is relevant because the subject requires students to connect doctrinal knowledge with ethical reasoning and contextual decisions (Kartini et al., 2022).

Recent studies show that higher-order thinking skills (HOTS) in Islamic Religious Education depend strongly on teachers' assessment literacy. Teachers may know the labels C1 to C6, but they do not always translate those labels into indicators, contextual stimulus materials, answer options, and scoring rubrics. Studies on Islamic education teachers' perceptions and HOTS assessment show that teachers often agree with the value of reasoning-based assessment, yet their classroom tests remain close to memorization and literal textbook understanding (Alfaizinun & Lilawati, 2023; Kosasih et al., 2022). This condition indicates that the problem is not only conceptual knowledge of taxonomy. It is also an operational problem of aligning learning objectives, task design, and cognitive demand.

The concept of constructive alignment strengthens this argument. Assessment should match learning objectives and classroom activities so that students are asked to demonstrate the same level of thinking that the curriculum intends (Biggs, 1996). In Islamic Religious Education, such alignment becomes more complex because a task may involve doctrinal accuracy, moral reasoning, and contextual judgment at the same time. A question with the verb explain can remain at C2 if it asks students to repeat a definition, but it can move toward C4 or C5 if it requires students to interpret a moral case, compare possible actions, and justify a decision. Therefore, cognitive level cannot be judged only from an operational verb. It must be judged from the whole item structure: objective, stimulus, expected response, and scoring criteria (Jansen & Möller, 2022; Zhou et al., 2023).

Previous studies have discussed HOTS-based Islamic Religious Education, problem-based learning, formative assessment, and teacher training. These studies explain that active learning, experiential learning, and formative feedback can help students move from memorization to reasoning (Adzidzah & Yudiawan, 2024; Liu & Pásztor, 2022). Other studies have examined HOTS assessment in Islamic Religious Education and found that analyzing and evaluating levels may appear in teacher-made tasks, but creating and metacognitive dimensions are still limited (Dewi et al., 2023; Prahesti et al., 2023). However, fewer studies audit the traceability between the target written in test blueprints, the actual cognitive demand of the item, the knowledge dimension being assessed, and the scoring criteria used by teachers. This traceability gap is important because an assessment can look HOTS in administrative documents while remaining low order in student response demand.

MTs Nurul Yaqin in North Cikarang, Bekasi Regency provides a relevant case for examining this issue. The madrasah serves students in an urban-industrial area where social diversity, digital media exposure, and differences in reading literacy shape classroom assessment practices. A preliminary reading of available assessment documents showed that many Islamic Religious Education items used commands such as mention, write, choose, and explain. These forms are useful for basic knowledge, but they do not always measure deeper reasoning. Some blueprints also stated C4 as the target level, while the final item required students only to restate a concept. Similar patterns have been found in recent studies on Islamic Religious Education item development, which stress the need for stimulus quality, rubrics, and validation of HOTS items (Putri et al., 2025; Saputri, 2025).

This study addresses the gap by examining assessment alignment in a single qualitative case. It asks three research questions: How do Islamic Religious Education teachers understand Anderson-Krathwohl's revised taxonomy? How are cognitive process levels and knowledge dimensions constructed in assessment instruments? What constraints and improvement strategies shape the alignment of objectives, stimuli, responses, and scoring?

Taqrib : Journal Of Islamic Studies
Vol. 4 No. 2 2026, 322-332

The contribution of this study is an empirical mismatch typology that shows how misalignment can occur at four layers: objective, stimulus, response, and scoring. This focus extends previous research by connecting teacher understanding, assessment documents, and institutional practices in one case study (Kistoro et al., 2023; Ningsi & Shaleh, 2024).

Method

This study used a qualitative single-case study design. A case study design was selected because the analysis focused on one bounded setting: Islamic Religious Education assessment practice at MTs Nurul Yaqin, North Cikarang, Bekasi Regency. The case was bounded by place, subject area, assessment documents, and the 2025/2026 academic year. This design allowed the study to compare what teachers said about HOTS-based assessment with the assessment instruments they actually prepared and used. Case study logic is suitable when a study aims to understand a contemporary educational practice in its real-life setting (Yin, 2018).

The participants were selected through purposive sampling. The main informants were three Islamic Religious Education teachers who taught grades VII and VIII and were involved in preparing daily assessments, mid-semester tests, and final-semester tests. One head of madrasah was included as a supporting informant because institutional support, teacher forums, and document review practices could not be understood from teacher interviews alone. The teachers had different teaching responsibilities in the Islamic Religious Education cluster, including Al-Qur'an Hadith, Akidah Akhlak, Fiqh, and Islamic Cultural History. Their teaching experience ranged from five to more than ten years.

Data collection was carried out in the second semester of the 2025/2026 academic year. The corpus consisted of four in-depth interviews, two observation sessions, and 14 assessment-related documents. The document corpus included three teaching modules, three test blueprints, three daily assessment sets, two mid-semester examination papers, two final-semester examination papers, and one internal assessment review note. The assessment documents contained 160 items or tasks. These items became the unit of document coding, while interview statements became the unit of thematic coding. The document audit did not function as statistical hypothesis testing. It provided a descriptive map of the assessment corpus so that claims such as dominant, limited, or rare had clear denominators.

Data were collected through interviews, observation, and documentation. The interviews explored teachers' understanding of the revised taxonomy, their experience in writing test blueprints, their use of operational verbs, their difficulty in designing Islamic contextual stimuli, and their views on students' reading literacy. Observation was used to record how teachers discussed assessment items, checked blueprints, and revised questions. Document analysis examined the written target in the blueprint, the operational verb, the stimulus, the expected student response, the knowledge dimension, and the scoring guide. This combination was used to compare declared understanding, observed practice, and written instruments.

The analysis used thematic analysis and document coding. Interview transcripts and observation notes were read repeatedly, coded, and grouped into themes related to declarative understanding, operational assessment literacy, stimulus design, student readiness, workload, and improvement efforts. Thematic analysis followed a staged process of familiarization, initial coding, theme construction, review, naming, and reporting (Braun & Clarke, 2021; Nowell et al., 2017). Document coding used a matrix based on Anderson-

Krathwohl's revised taxonomy. Each item was coded by cognitive process (C1-C6), knowledge dimension, stimulus type, response demand, and scoring availability. When the blueprint label differed from the actual response required by the item, the final classification followed the actual cognitive demand of the item rather than the label in the blueprint.

To strengthen trustworthiness, the study used source triangulation, technique triangulation, member reflection, and peer audit. Interview claims were compared with observation notes and assessment documents. Preliminary interpretations were discussed with the teachers to check whether the interpretation was reasonable in relation to their practice. Two authors independently coded the document corpus using the same codebook. Disagreements were discussed until agreement was reached, and the decision trail was retained in an audit matrix. This procedure was used because classifying C1-C6 involves judgment and requires transparent criteria. Ethical procedures included permission from the madrasah, voluntary participation, anonymization of informants, and removal of student names from documents. The study did not collect sensitive personal data from students.

Table 1
Participant Profile

Code	Role	Teaching area	Experience	Data source
T1	Islamic Religious Education teacher	Al-Qur'an Hadith and Fiqh	5-7 years	Interview and document clarification
T2	Islamic Religious Education teacher	Akidah Akhlak	8-10 years	Interview, observation, and document clarification
T3	Islamic Religious Education teacher	Islamic Cultural History and PAI cluster tasks	More than 10 years	Interview and document clarification
HM	Head of madrasah	Institutional supervision	More than 10 years	Supporting interview and institutional documents

Source: Field interview and document verification, 2026.

Table 2
Document Corpus and Coding Units

Document type	Number of documents	Number of items/tasks	Coding focus
Teaching modules	3	-	Learning objectives, assessment plan, and expected cognitive demand
Test blueprints	3	-	Target cognitive level, operational verb, and knowledge dimension
Daily assessment sets	3	60	Item stimulus, expected response, and cognitive level
Mid-semester examination papers	2	50	Item stimulus, answer format, and cognitive level
Final-semester examination papers	2	50	Item stimulus, answer format, and cognitive level
Internal assessment review note	1	-	Peer-review practice and improvement efforts
Total	14	160	Assessment alignment audit

Source: Assessment document corpus, 2026.

Taqrib : Journal Of Islamic Studies
Vol. 4 No. 2 2026, 322-332

Results and Discussion

Teachers' Conceptual Understanding of the Revised Taxonomy

The first finding shows a gap between declarative knowledge and operational assessment literacy. All three teachers could mention the sequence from C1 to C6 and identify Anderson-Krathwohl's revised taxonomy as a revision of Bloom's taxonomy. T1 stated, "We know the order from remembering to creating, but we still hesitate when we must turn it into questions for akidah or akhlak" (T1, interview). This statement shows that the teachers were not unfamiliar with the taxonomy. The difficulty appeared when they had to translate the taxonomy into indicators, stimuli, answer expectations, and scoring criteria.

The teachers were more confident with factual and conceptual knowledge than with procedural and metacognitive knowledge. Factual knowledge appeared in questions about names of angels, pillars of faith, meanings of verses, and Islamic historical figures. Conceptual knowledge appeared in questions about religious values, legal categories, and moral principles. Procedural knowledge appeared in worship materials, such as ablution, prayer, zakat, and pilgrimage. Metacognitive knowledge was the least visible. T2 explained, "Reflection questions are important, but we are unsure how to assess them fairly in a written test" (T2, interview). This finding supports the argument that teachers' assessment literacy involves more than knowing terms. It requires the ability to design evidence of student thinking (Xu & Brown, 2016).

Observation also showed that teachers used operational verbs as practical shortcuts. In one review session, teachers first checked whether the blueprint contained verbs such as analyze or evaluate before checking whether the item required students to analyze or evaluate. This practice explains why a blueprint may claim C4 while the final item remains C2. The problem is not the presence of operational verbs, but their use as labels instead of design tools. In Anderson-Krathwohl's matrix, the cognitive process must be read together with the knowledge dimension and the demanded response (Anderson & Krathwohl, 2001).

The mechanism behind the gap can be explained by three interacting factors. First, teachers want to introduce HOTS but lack concrete examples of Islamic contextual stimuli. Second, they worry that students with uneven reading literacy will misread long cases. Third, administrative workload makes it easier to reuse existing item banks. These factors make low-order items feel safer, faster, and easier to score. This finding extends previous studies by showing how declarative taxonomy knowledge can fail to become operational assessment literacy when stimulus design, student readiness, and institutional workload interact in daily assessment practice (Latifah, 2023; Otaya et al., 2025).

Table 3
Evidence of Teachers' Understanding and Assessment Literacy

Evidence code	Primary evidence	Analytical meaning
INT-T1	"We know the order from remembering to creating, but we still hesitate when we must turn it into questions for akidah or akhlak."	The teacher has declarative taxonomy knowledge but limited operational design skill.
INT-T2	"Reflection questions are important, but we are unsure how to assess them fairly in a written test."	Metacognitive knowledge is recognized but rarely operationalized.
OBS-01	Teachers checked the KKO label before examining stimulus and expected response.	Operational verbs tend to function as administrative markers.

DOC-Blue-02	Blueprint stated C4, but the final item asked students to restate a definition.	Mismatch between target level and actual response demand.

Source: Interview transcripts, observation notes, and document coding, 2026.

Cognitive Level Construction in Assessment Instruments

Document coding of 160 items shows that assessment instruments were dominated by remembering and understanding. The corpus contained 62 C1 items and 50 C2 items. Together, they represented 70 percent of all coded items. C3 appeared in 28 items, while C4, C5, and C6 appeared in 20 items combined. This distribution supports the reviewers' concern that the word dominated needs a denominator. The pattern indicates that the assessment system still relies heavily on lower-order thinking, especially in multiple-choice and short-answer formats.

The distribution varied by assessment type. Daily assessments contained more C3 and C4 items because teachers sometimes used classroom situations and short cases. Mid-semester and final-semester tests had more C1 and C2 items because teachers prioritized item clarity, scoring speed, and score stability. This tendency aligns with Jansen & Möller (2022), who show that judgments of higher-order skills can be influenced by lower-order performance. In this case, teachers lowered the cognitive demand because they expected students to struggle with long stimuli.

The most frequent mismatch appeared at the objective and response layers. For example, one blueprint classified an item as C4 because it used the verb analyze. However, the final item asked students to explain the meaning of honesty based on the textbook. This item was classified as C2 because the expected answer required meaning explanation, not analysis of a case. Another item was more aligned: students were given a short fiqh situation about purification before prayer and asked to determine the correct procedural steps. This item was classified as C3 procedural knowledge because it required applying a rule to a practical situation.

C5 and C6 required stricter judgment. An item about cyberbullying could be classified as C5 only when students had to judge behavior based on an Islamic moral criterion and justify the decision. If students only wrote a general opinion without criteria, the item could not be treated as valid evaluation. Similarly, a poster, short sermon, or video was not automatically classified as C6. It was classified as C6 only if the task required students to synthesize religious concepts, select relevant evidence, design a message for a target audience, and explain the reason for their product. Without these criteria, the product remained a performance assignment with unclear cognitive demand. This finding clarifies the religio-ethical specificity of Islamic Religious Education assessment: the task should assess not only product creation, but also the student's ability to reorganize religious knowledge into reasoned moral communication.

The analysis generated a four-layer mismatch typology. The first layer is objective mismatch, which occurs when the cognitive level in the blueprint does not match the real item. The second is stimulus mismatch, which occurs when an item claims HOTS but provides no case, data, conflict, or problem to analyze. The third is response mismatch, which occurs when the task asks for a short recall answer although the blueprint states analysis or evaluation. The fourth is scoring mismatch, which occurs when an item asks for reasoning or creation but has no analytic rubric. This typology is the main empirical

Taqrib : Journal Of Islamic Studies
Vol. 4 No. 2 2026, 322-332

contribution of the study because it shows how misalignment emerges within actual Islamic Religious Education assessment documents.

Table 4
Distribution of Cognitive Levels in the Assessment Corpus

Assessment type	Items	C1	C2	C3	C4	C5	C6	Dominant pattern
Daily assessment	60	20	20	12	5	2	1	C1-C2 with some C3-C4 cases
Mid-semester test	50	22	16	8	3	1	0	C1-C2 and short explanation
Final-semester test	50	20	14	8	4	3	1	C1-C2 with limited C4-C5
Total	160	62	50	28	12	6	2	C1-C2 = 70%; C4-C6 = 12.5%

Source: Document coding matrix, 2026.

Table 5
Examples of Alignment and Mismatch in Assessment Items

Code	Blueprint target	Item or task pattern	Actual classification	Reason for classification
DOC-04	C4 analyzing	Students were asked to explain the meaning of honesty.	C2 conceptual	The expected response was a definition or explanation, not analysis of a case.
DOC-07	C3 applying	Students read a short fiqh situation about impurity before prayer and chose the correct purification steps.	C3 procedural	The item required applying a rule to a practical worship situation.
DOC-11	C5 evaluating	Students judged cyberbullying based on hifz al-lisan and gave a reason.	C5 conceptual	The item required judgment using a religious moral criterion, but scoring needed clearer rubric.
DOC-13	C6 creating	Students made a religious poster with a verse and message for peers.	Conditional C6	The task measured C6 only when the rubric assessed synthesis, audience relevance, and reasoning, not appearance only.

Source: Document coding matrix and peer-audit notes, 2026.

Constraints and Improvement Efforts

The first constraint was limited practical training in HOTS item writing. The teachers had received general socialization about HOTS, but they had limited experience with clinical item-writing practice. In this revised manuscript, the term clinical training refers to guided training that includes drafting, reviewing, revising, and scoring assessment items with direct feedback. The absence of such training explains why teachers could state the C1-C6 sequence but struggled to design stimuli, distractors, and rubrics.

The second constraint was heterogeneous student reading literacy. T2 noted, "When the text is long, some students do not read it completely and then guess the answer" (T2, interview). This problem encouraged teachers to shorten the stimulus or lower the cognitive level of the item. The issue is not that HOTS is impossible for MTs students. Rather, students need graded exposure to stimulus-based questions before formal tests. This supports the role of formative assessment and problem-based learning in preparing students for higher-order tasks (Adzidzah & Yudiawan, 2024; Nuraini, 2025).

The third constraint was administrative workload and examination culture. Teachers had to prepare teaching documents, teach, score, and report results. As a result, old item banks were often reused because they were familiar and easy to score. HM stated, "The teacher forum already reviews questions, but usually it checks completeness first, not always the depth of thinking" (HM, interview). This statement shows that institutional review exists, but its focus needs to move from document completeness to cognitive alignment.

The improvement efforts should be read as emerging institutional strategies, not as tested interventions. Teachers used internal subject-teacher forums to share items, revise blueprints, and discuss student difficulty. They also participated in HOTS workshops when available. However, the findings indicate that workshops need to be followed by peer moderation, a codebook for item audit, analytic rubrics, and a shared item bank. The item bank should classify each item by topic, cognitive process, knowledge dimension, stimulus type, response demand, and scoring criteria. This structure can help teachers avoid repeated C1-C2 patterns and make the distribution of cognitive demand more visible.

These findings offer a more precise explanation than simply saying that teachers lack competence. The teachers had conceptual willingness and basic taxonomy knowledge. The misalignment emerged because assessment literacy requires design judgment under real constraints. In Islamic Religious Education, this judgment includes doctrinal accuracy, ethical reasoning, student literacy, and feasible scoring. Therefore, improvement should focus on practical moderation of actual items. A codebook-based review can help teachers decide whether an item truly measures remembering, understanding, applying, analyzing, evaluating, or creating.

Table 6
Constraints, Evidence, and Improvement Priorities

Constraint	Evidence code	Impact on assessment	Priority for improvement
Limited clinical item-writing training	INT-T1, OBS-01	Teachers know C1-C6 but struggle to write stimulus, distractors, and rubrics.	Guided item-writing workshop with direct feedback.
Heterogeneous student reading literacy	INT-T2, DOC-05	Teachers shorten cases or lower cognitive demand to avoid student confusion.	Graded formative stimulus practice before formal examinations.
Administrative workload	INT-HM, OBS-02	Old item banks are reused and blueprint review is reduced.	Internal peer moderation using a simple codebook.
Weak scoring criteria	DOC-11, DOC-13	C5 and C6 tasks risk being scored subjectively.	Analytic rubrics for reasoning, criteria use, and synthesis.

Source: Interview transcripts, observation notes, and document audit, 2026.

Conclusion

This study shows that the main issue in Islamic Religious Education assessment at MTs Nurul Yaqin is not the absence of taxonomy knowledge, but the gap between declarative knowledge of Anderson-Krathwohl's revised taxonomy and operational assessment literacy. Teachers can name the C1-C6 sequence and understand the idea of HOTS, yet they still face difficulty converting this knowledge into aligned objectives, stimuli, response demands, and scoring criteria.

The document audit found that C1 and C2 dominated the assessment corpus, while C4, C5, and C6 appeared only in limited forms. The strongest contribution of this study is the identification of four layers of assessment mismatch: objective mismatch, stimulus mismatch, response mismatch, and scoring mismatch. This typology explains why an item can be labeled HOTS in a blueprint but remain low order in actual student work.

The practical implication is that the madrasah needs a structured assessment review system. The most urgent actions are a codebook for cognitive-level classification, peer moderation of test items, analytic rubrics for essay and project tasks, and gradual use of stimulus-based formative assessment. These steps are recommendations derived from the observed constraints and document audit. They should be treated as improvement priorities, not as evidence that validity or fairness has already increased.

This study has limitations. It examined one madrasah, three teachers, one supporting informant, and a limited document corpus. The classification of cognitive levels involved researcher judgment, although peer audit was used to reduce bias. The study did not test the psychometric validity of items, scoring reliability, or the effect of an item bank on student learning outcomes. Future studies should compare several madrasahs, validate the coding rubric with multiple raters, examine inter-rater agreement, and test whether an Anderson-Krathwohl-based item bank improves item quality and student performance.

Declarations

Author contribution. Agustin Suci Rizkianti developed the study concept, coordinated manuscript revision, and reviewed the final manuscript. Muamar Sidik and Nana Suryana supported data organization and literature synthesis. Rudi Iskandar, Sohwatun Hanifah, and Noor Azida Batubara supported manuscript preparation, formatting, and reference checking. All authors approved the final manuscript.
Funding statement. This research received no external funding.
Conflict of interest. The authors declare no conflict of interest.
Ethical statement. Institutional permission was obtained from the madrasah. Participation was voluntary. Informant names were anonymized, and student names were removed from the document corpus.
Additional information. No additional information is available for this paper.

References

Adzidzah, N., & Yudiawan, A. (2024). HOTS-based formative assessment: The key to improving the quality of learning. *Journal of Quality Assurance in Islamic Education (JQAIE)*, *4*(2), 109–120. https://doi.org/10.47945/jqaie.v4i2.1670

Alfaizinun, C. N., & Lilawati, E. (2023). Penerapan pembelajaran Pendidikan Agama Islam berbasis Higher Order Thinking Skill (HOTS) di SMA Negeri Jogoroto. *ISLAMIKA*, *5*(4), 1551–1565. https://doi.org/10.36088/islamika.v5i4.3877

Anderson, L. W., & Krathwohl, D. R. (2001). *A taxonomy for learning, teaching, and assessing: A revision of Bloom's taxonomy of educational objectives*. Addison Wesley Longman, Inc.

Biggs, J. (1996). Enhancing teaching through constructive alignment. *Higher Education*, *32*, 347–364. https://doi.org/10.1007/BF00138871

Braun, V., & Clarke, V. (2021). *Thematic analysis: A practical guide*. SAGE Publications.

Dewi, Y. S., In'ami, M., Audia, A., & Masrukhin, M. (2023). Assessment based on higher

order thinking skill on SKI learning. *AL-ISHLAH: Jurnal Pendidikan*, *15*(1), 443–458. https://doi.org/10.35445/alishlah.v15i1.2586

Jansen, T., & Möller, J. (2022). Teacher judgments in school exams: Influences of students' lower-order-thinking skills on assessing students' higher-order-thinking skills. *Teaching and Teacher Education*, *111*, 103616. https://doi.org/10.1016/j.tate.2021.103616

Kartini, N. E., Nurdin, E. S., Hakam, K. A., & Syihabuddin, S. (2022). Telaah revisi teori domain kognitif Taksonomi Bloom dan keterkaitannya dalam kurikulum Pendidikan Agama Islam. *Jurnal Basicedu*, *6*(4), 7292–7302. https://doi.org/10.31004/basicedu.v6i4.3478

Kistoro, H. C. A., Latipah, E., & Burhan, N. M. (2023). Probing experiential learning approach in Islamic religious education. *Jurnal Pendidikan Islam*, *9*(2), 157–168. https://doi.org/10.15575/jpi.v9i2.24374

Kosasih, A., Supriyadi, T., Firmansyah, M. I., & Rahminawati, N. (2022). Higher-order thinking skills in primary school: Teachers' perceptions of Islamic education. *Journal of Ethnic and Cultural Studies*, *9*(1), 56–76. https://doi.org/10.29333/ejecs/994

Latifah, E. (2023). Peningkatan kompetensi guru Pendidikan Agama Islam dalam mengembangkan soal higher order thinking skills melalui bimbingan individu. *JIIP - Jurnal Ilmiah Ilmu Pendidikan*, *6*(5), 3396–3404. https://doi.org/10.54371/jiip.v6i5.2051

Liu, Y., & Pásztor, A. (2022). Effects of problem-based learning instructional intervention on critical thinking in higher education: A meta-analysis. *Thinking Skills and Creativity*, *45*, 101069. https://doi.org/10.1016/j.tsc.2022.101069

Ningsi, A., & Shaleh, S. (2024). Problematika penerapan asesmen berbasis HOTS (Higher Order Thinking Skills) di sekolah dasar. *SAP (Susunan Artikel Pendidikan)*, *8*(3), 447–455. https://doi.org/10.30998/sap.v8i3.20998

Nowell, L. S., Norris, J. M., White, D. E., & Moules, N. J. (2017). Thematic analysis: Striving to meet the trustworthiness criteria. *International Journal of Qualitative Methods*, *16*(1), 1–13. https://doi.org/10.1177/1609406917733847

Nuraini, R. (2025). Implementation of the problem based learning strategy in Islamic Religious Education learning in improving students' higher order thinking skills. *Global Education Journal*, *3*(1), 45–52. https://doi.org/10.59525/gej.v3i1.660

Otaya, L. G., Yahiji, K., Rasimin, R., Rahmawati, R., Imtihan, N., & Muhamad, N. B. (2025). Assessment of teachers' ability to develop HOTS-based test items: A GRM analysis. *Jurnal Ilmiah Peuradeun*, *13*(3), 1995–2018. https://doi.org/10.26811/peuradeun.v13i3.1555

Prahesti, V. D., Fatonah, S., & Maisarah, A. (2023). Implementation of HOTS assessment in Islamic religion lesson in 3rd grade Islamic elementary school East Java. *MUDARRISA: Jurnal Kajian Pendidikan Islam*, *15*(1), 42–59. https://doi.org/10.18326/mdr.v15i1.42-59

Putri, Y., Amra, A., Yoma, A. R., Amri, I., & Wince, W. (2025). Pengembangan sistem evaluasi pembelajaran Pendidikan Agama Islam berbasis Taksonomi Bloom Revisi untuk meningkatkan keterampilan berpikir tingkat tinggi peserta didik di SMAN 1 Gunung Talang. *Jurnal Penelitian Ilmu Pendidikan Indonesia*, *4*(4), 2815–2822. https://doi.org/10.31004/jpion.v4i4.849

Saputri, R. (2025). Development of higher order thinking skills questions on Kahoot media

Taqrib : Journal Of Islamic Studies
Vol. 4 No. 2 2026, 322-332

in Islamic Religious Education and character education. *Al-Liqo: Jurnal Pendidikan Islam*, *10*(2), 293–307. https://doi.org/10.46963/alliqo.v10i2.2501

Xu, Y., & Brown, G. T. L. (2016). Teacher assessment literacy in practice: A reconceptualization. *Teaching and Teacher Education*, *58*, 149–162. https://doi.org/10.1016/j.tate.2016.05.010

Yin, R. K. (2018). *Case Study Research and Applications: Design and Methods*. Sage.

Zhou, Y., Gan, L., Chen, J., Wijaya, T. T., & Li, Y. (2023). Development and validation of a higher-order thinking skills assessment scale for preservice teachers. *Thinking Skills and Creativity*, *48*, 101272. https://doi.org/10.1016/j.tsc.2023.101272