MATHE*dunesa*
Homepage: https://ejournal.unesa.ac.id/index.php/mathedunesa/index
Email: mathedunesa@unesa.ac.id
p-ISSN: 2301-9085; e-ISSN: 2685-7855
Vol. 15 No. 1 Tahun 2026
Halaman 127-133

Improving Islamic Boarding School Teacher' Competence in Designing AI-Based Assessment of Critical Thinking Skills

Masriyah[1*], Abdul Haris Rosyidi[1], Nurus Saadah[1], Abdul Hafidz[1], Umi Hanifah[2], Firda Hariyanti[2], Ayu Silvi Lisvian Sari[3]

[1]Mathematics Education Department, Universitas Negeri Surabaya, Surabaya, Indonesia
[2]Mathematics Education Department, Universitas Nahdlatul Ulama Pasuruan, Indonesia
[3]Mathematics Education Department, Universitas PGRI Adi Buana Surabaya (Blitar Campus), Indonesia

DOI: https://doi.org/10.26740/mathedunesa.v15n1.p127-133

Article History:
Received: 27 December 2025
Revised: 23 February 2026
Accepted: 23 February 2026
Published: 10 March 2026

Keywords:
AI in assessment; critical thinking; HOTS; teacher professional development; Islamic boarding school; *pesantren*

***Corresponding author:**
masriyah@unesa.ac.id

Abstract: This study reports a structured capacity-building program designed to improve Islamic boarding school teachers' competence in designing AI-based assessments for critical thinking skills. The intervention was carried out with 91 teachers across 14 school subjects at a large *pesantren* —a traditional Islamic educational institution that integrates religious and national curricula—in East Java, Indonesia. The program integrated three components: (i) foundational understanding of generative AI, (ii) explicit mapping of higher-order thinking indicators (analysis, inference, evaluation), and (iii) co-construction of assessment tasks and analytic rubrics through guided prompt engineering. A one-group pretest–posttest design was used to evaluate teacher learning gains. Results demonstrate a very large improvement in teachers' assessment design knowledge and skills, with mean scores increasing from 37.75 (pretest) to 90.49 (posttest), and a statistically significant t-value ($t = 29.49$, $p < .01$). Survey findings also indicate strong positive perceptions of AI usefulness, relevance, and feasibility. Teachers were able to generate authentic HOTS tasks contextualized to Islamic and local cultural values, accompanied by clear rubrics for evaluating evidence-based reasoning. The study concludes that AI-assisted design can meaningfully accelerate teacher capacity to design higher-order assessments in faith-based educational settings, provided that AI is used as a scaffold for human reasoning, not as an answer generator. Implications for scaling professional development and institutionalizing AI-embedded assessment innovation in *pesantren* contexts are discussed.

INTRODUCTION

Critical thinking is internationally recognized as a central twenty-first-century competence, enabling learners to analyze relationships, evaluate evidence, and justify decisions in authentic situations (Facione, 2020). Consequently, assessment for critical thinking must move beyond low-order recall to intentionally elicit reasoning processes, meta-cognition, and evidence-based argumentation (Brookhart, 2022). The shift from knowledge reproduction toward knowledge construction has triggered global pressure for assessment redesign in schools, including in faith-based education such as Islamic boarding schools (*pesantren*) in Indonesia.

In the last three years, generative Artificial Intelligence (AI)—particularly large language models—has emerged as a strategic partner for teacher work. Multiple studies show that AI can accelerate (a) generation of stimulus materials, (b) production of contextual problem variants, and (c) refinement of analytic rubric criteria for higher-order thinking

tasks (Zhang et al., 2024; Holmberg & Madigan, 2023; Waycott et al., 2024). Instead of replacing teachers, AI can function as a "cognitive amplifier" that helps educators iterate more efficiently toward better assessment designs (Sütfeld et al., 2023). However, this potential can only be realized when teachers have conceptual clarity about HOTS indicators and are able to direct AI with appropriate prompt engineering; otherwise, AI becomes merely a tool for answer-generation, not a scaffold for thinking (Chen & Tsai, 2024).

In Indonesian *pesantren*, a unique challenge emerges. Teachers must design tasks that elicit modern critical thinking processes without losing the cultural, ethical, and religious identity of Islamic learning ecosystems (Nasir & Zainuddin, 2023). Recent works in the Indonesian context show that: (i) *pesantren* teacher digital adoption is increasing, but (ii) competency in AI-assisted assessment design is not yet systematically developed, and (iii) most professional development (PD) is still tool-oriented, not competence-oriented (Rahman & Fitri, 2024). There is therefore an urgent need for structured PD models that integrate AI literacy, critical-thinking assessment theory, and practical guided design.

This paper reports an intensive capacity-building program for 91 teachers across 14 subjects in an Islamic boarding school. The program combined short lectures, demonstration, and guided co-design using AI to generate HOTS-aligned items and rubrics. This study aims to: (1) enhance teachers' conceptual understanding of critical thinking assessment, (2) strengthen AI-supported assessment design skills, and (3) produce concrete teacher-generated assessment artifacts. The effectiveness of the program was examined using a pretest–posttest design and analyzed through paired-sample statistical comparison.

The contribution of this paper is twofold. First, it provides empirical evidence that a single-day, highly structured PD can generate significant teacher learning gains and produce quality assessment outputs in a *pesantren* context. Second, it formalizes a practical model for AI-supported assessment design that can be replicated by other institutions—particularly faith-based schools—seeking to integrate AI ethically and pedagogically into their assessment ecosystems.

METHOD

This study details the implementation of a professional development (PD) program at an Islamic boarding school. The activity followed a structured progression consisting of three main stages: preparation, implementation, and evaluation.

Context and Participants

This study was implemented in an Islamic boarding school (*pesantren*) in East Java, Indonesia. The school integrates Islamic sciences with the national curriculum. A total of 91 teachers voluntarily participated in the intervention. These participants represented 14 subject domains, including religion-based subjects (Aqidah Akhlak, Al-Qur'an Hadis, Fiqh, and History of Islamic Culture) and general subjects (Arabic, Bahasa Indonesia, English, Science, Social Studies, Mathematics, Physical Education, ICT, Guidance and Counseling, and Worship Practices). All participants taught at the lower secondary level, equivalent to

grades 7–9. While most teachers possessed basic ICT experience, they had not previously received structured training in AI-assisted assessment design.

Intervention and Prompt Engineering Protocol

The intervention was delivered as a one-day intensive program (8 hours) consisting of three sequential modules: AI Literacy, Critical Thinking & HOTS Assessment, and an AI-Supported Assessment Co-Design Workshop.

To strengthen methodological transparency and ensure the quality of AI outputs, the workshop utilized a structured prompt engineering protocol. Teachers were guided to construct prompts using a specific four-component framework: (1) Role: Assigning the AI a persona (e.g., "Act as an expert mathematics teacher"); (2) Task: Defining the specific assessment goal (e.g., "Create a scenario-based question to test the 'Inference' indicator"); (3) Context: Incorporating *pesantren* values or local social scenarios to ensure cultural relevance; (4) Constraint: Specifying that the output must include an analytic rubric with clear evidence-based reasoning criteria.

To mitigate risks of AI hallucination or logical errors, an iterative prompting strategy was employed where teachers refined AI-generated drafts through follow-up prompts. Furthermore, a "human-in-the-loop" verification process was implemented, where all AI-generated artifacts were subjected to peer reviews and expert facilitation to ensure pedagogical and content accuracy.

Instruments and Data Analysis

A 10-item multiple-choice test was utilized to measure teachers' fundamental conceptual knowledge regarding AI concepts and HOTS indicators. While a 10-item test is concise, it was designed specifically to establish a baseline of theoretical understanding; the more complex applied competencies (such as task design and rubric construction) were assessed separately through the qualitative analysis of generated artifacts, ensuring a comprehensive evaluation of both theory and practice. Additionally, a Likert-scale perception survey (1–5) measured four constructs: (i) perceived usefulness, (ii) confidence, (iii) relevance to curriculum, and (iv) intention to adopt. Descriptive statistics summarized these patterns. For qualitative output, two coders independently verified the alignment between teacher-created tasks and critical thinking indicators, as well as the clarity of rubric criteria, with discrepancies resolved through consensus.

RESULT AND DISCUSSION

Pretest–Posttest Gain Scores

The pretest scores indicated that teachers initially possessed limited conceptual understanding of AI-based assessment design and weak grasp of critical thinking indicators. The average pretest score was 37.75 on a scale of 0–100. After the intervention, the posttest mean increased dramatically to 90.49[2]. The mean gain score (Δ) was 52.75 points[3]. These summary results are presented in Table 1. A paired-samples t-test confirmed that the improvement was statistically significant ($t(90) = 29.49$, $p < .01$). This result provides strong evidence that the intervention had a large and meaningful impact on

teachers' knowledge and skill in AI-based HOTS assessment design. The interpretation of this significant performance leap and its specific attribution to AI use versus other pedagogical factors is elaborated in the Discussion section.

Table 1. Summary of Pretest and Posttest Scores (N = 91)

Variable	Mean
Pretest Score	37.75
Posttest Score	90.49
Difference (Post – Pre)	52.75

Gain Scores by Subject Cluster

When the data were analyzed by subject clusters, all groups showed substantial gain scores, as shown in Table 2. This finding indicates that the instructional model was domain-general, and that the core design logic (critical thinking indicators → task → rubric → AI-refinement) was successfully transferred across religious, linguistic, STEM, and social sciences domains.

Table 2. Mean Gain Scores by Subject Cluster

Subject Cluster	N Teacher	Mean Gain (Δ)
Islamic Sciences (Aqidah, Qur'an-Hadis, Fiqh, SKI)	32	51.98
Language (Arabic, Indonesian, English)	21	54.21
STEM (Science, Mathematics, ICT)	17	55.87
Social & Human Sciences (IPS, Guidance, etc.)	21	50.62

Visual Comparison

To further illustrate the change, the comparison of pretest and posttest mean scores is presented in Figure 1. The bar chart visually confirms the large improvement. Teachers more than doubled their performance after the intervention

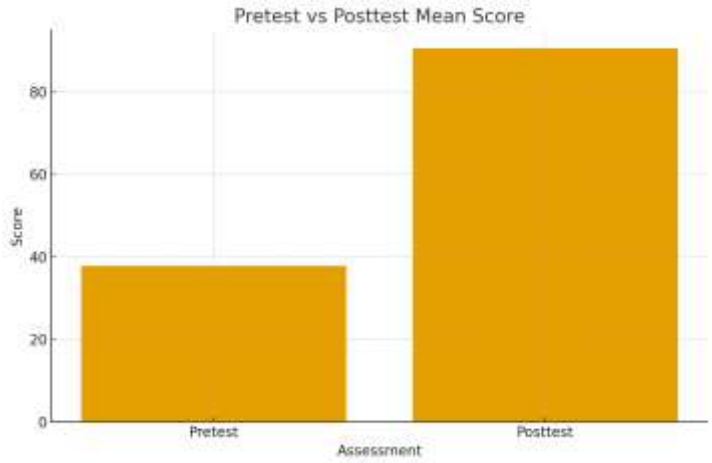

Figure 1. Comparison of Mean Pretest and Posttest Scores of Teachers (N = 91)

Teacher Perceptions

Survey responses (1–5 Likert Scale) indicated strong positive attitudes toward the intervention. Recapitulation of the three core constructs can be seen in Table 3. Table 3 indicates that "Agree + Strongly Agree" consistently exceeds 75% for the three constructs, confirming strong readiness toward continued use of AI in assessment.

Table 3. Teacher Perceptions Regarding Usefulness, Capability, and Adoption Intention (N = 91)

Construct	Strongly Agree	Agree	Neutral	Disagree	Strongly Disagree
AI is useful for HOTS assessment design	62%	25%	11%	2%	0%
I feel capable of applying what I learned	48%	30%	18%	3%	1%
I intend to continue using AI after this workshop	67%	18%	11%	3%	1%

Quality of Teacher-Generated Artifacts

Finally, qualitative analysis of teacher-created products (tasks, prompts, and rubrics) confirmed that teachers also demonstrated applied competence and not only conceptual understanding. Empirical evidence shows that 82 out of 91 (90.1%) teacher-generated tasks successfully required higher-order thinking processes, such as analysis, justification, or evaluation, rather than simple recall. The quality of these artifacts was evidenced by the inclusion of explicit evaluative criteria, specifically requiring students to provide claims, evidence, and logical reasoning. Furthermore, the rubrics developed through AI-assisted co-design featured detailed performance level descriptors that allowed for a nuanced assessment of critical thinking. This high rate of successful output confirms that teachers were able to produce assessment tasks strictly consistent with HOTS principles. It demonstrates that AI was effectively utilized as a design amplifier to scaffold professional reasoning rather than functioning as a simple automated answer generator.

Discussion

The results demonstrate that a single–day, structured professional development (PD) integrating AI literacy, critical thinking indicators, and guided prompt engineering can rapidly improve teachers' competence in assessment design. The magnitude of the improvement (mean score increase from 37.75 to 90.49 on a scale of 0–100; $\Delta = 52.75$; $p < .01$) indicates that teacher learning was not incremental, but transformative. This significant statistical gain directly addresses the first two research aims stated in the Introduction: (1) enhancing teachers' conceptual understanding of critical thinking assessment and (2) strengthening AI-supported assessment design skills.

In the absence of longitudinal data, the term "transformative learning" is used here to describe a qualitative shift in teachers' procedural knowledge, moving from traditional knowledge reproduction toward complex, AI-assisted knowledge construction. This dramatic jump in performance is specifically attributed to the role of generative AI acting as a "cognitive amplifier" that scaffolded teachers' ability to bridge the gap between theoretical HOTS indicators and the practical generation of complex tasks (Sütfeld et al., 2023). However, the success of the intervention also stems from the workshop's pedagogical structure, which provided the necessary framework to direct the AI's output effectively. This corresponds with Zhang et al. (2024), who found that AI assistance is most effective when its use is anchored to explicit performance criteria rather than unconstrained output. This indicates that the success also stems from the workshop's pedagogical structure, which provided the necessary framework to direct the AI's output effectively.

The empirical evidence for the third research aim—producing concrete assessment artifacts—is found in the high success rate of the outputs. Qualitative analysis confirms that

82 out of 91 (90.1%) teacher-generated tasks successfully required analysis, justification, or evaluation instead of simple recall. These artifacts utilized rubrics with explicit criteria, including claims, evidence, and logical reasoning.

The positive perception data also confirms that teachers recognize AI as a legitimate partner in instructional design, with **87%** of participants (62% strongly agree; 25% agree) viewing AI as useful for HOTS assessment design. Furthermore, **85%** of teachers (67% strongly agree; 18% agree) expressed a firm intention to continue using AI after the workshop. This supports Holmberg & Madigan's (2023) claim that the pedagogical "alignment potential" of AI is unlocked only when teachers maintain epistemic control of what the AI is generating. Teachers in pesantren contexts in this study explicitly retained control: they contextualized assessment tasks within Islamic ethics and local social values. This is consistent with arguments that Islamic critical literacy is compatible with modern cognitive competencies when contextualized (Nasir & Zainuddin, 2023).

Another notable finding is that the effect was domain–general: religion, languages, STEM, and social science clusters all experienced large gains. This suggests that the design logic (indicator → task → rubric → AI refinement) is transferable across disciplines. This parallels findings that rubric–mediated AI co-construction can support interdisciplinary assessment design (Waycott et al., 2024). The fact that teachers could produce HOTS–aligned items in a short time confirms that the key barrier was a lack of procedural know–how rather than content knowledge. When a clear structure was given through indicator mapping and prompt patterns, teachers could scale up quickly. Despite the positive outcomes, several limitations must be acknowledged. First, the use of a one-group pretest–posttest design without a control group limits the ability to isolate the intervention's effects from external variables. Second, the evaluation relied on an immediate posttest, reflecting short-term learning rather than long-term retention. Finally, survey responses may be subject to social desirability bias due to the supportive workshop environment. Future research should utilize longitudinal designs to evaluate the sustainability of AI-based assessment practices in the *pesantren* context.

CONCLUSION AND SUGGESTIONS

This study provides empirical evidence that a single–day, structured professional development integrating AI literacy, critical thinking indicators, and guided prompt engineering can significantly improve teachers' competence in designing AI-based assessments. The substantial gain in scores from pretest to posttest indicates that the intervention effectively shifted teachers' conceptual and procedural understanding of how to generate HOTS-oriented tasks and analytic rubrics. The qualitative artifacts also confirm that teachers were able to contextualize AI outputs into culturally and religiously relevant assessment tasks suitable for pesantren education.

The results highlight that AI can be used not to replace teacher reasoning but to strengthen it. When teachers maintain epistemic control, AI becomes a design amplifier that accelerates item generation, scenario contextualization, and rubric refinement. The domain-

general impact across 14 subjects indicates that this approach is scalable and transferable to different knowledge domains within Islamic boarding schools.

Further work should examine classroom enactment quality and student learning outcomes when these AI-generated assessment tasks are implemented. Longitudinal follow-up studies are needed to understand the sustainability of AI-based assessment practices and the institutional conditions required for continuous adoption.

Acknowledgment
The authors express their gratitude to the leadership and the 91 participating teachers of MTsN 3 Tambakberas Jombang for their active engagement throughout the program. Appreciation is also extended to Universitas Negeri Surabaya and Universitas Nahdlatul Ulama Pasuruan for institutional support that made this community–based research possible. This program was carried out with facilitation and resource support from the Lembaga Penelitian dan Pengabdian kepada Masyarakat (LPPM) Universitas Negeri Surabaya.

REFERENCES

Brookhart, S. M. (2022). *How to Assess Higher-Order Thinking Skills*. Alexandria, VA: ASCD.

Chen, K., & Tsai, C. (2024). Misalignment risks of generative AI in assessment. *Educational Technology Research and Development*, 72(2), 521–548. https://doi.org/10.1007/s11423-023-10332-6

Facione, P. A. (2020). *Critical Thinking: What It Is and Why It Counts*. Millbrae, CA: Insight Assessment.

Holmberg, J., & Madigan, D. (2023). Generative AI for pedagogical alignment. *British Journal of Educational Technology*, 55(1), 11–27. https://doi.org/10.1111/bjet.13364

Nasir, M., & Zainuddin, A. (2023). Islamic critical literacy and modern pedagogy in pesantren. *Journal of Islamic Education Studies*, 8(3), 301–320. https://doi.org/10.15575/jies.v8i3.24556

Rahman, F., & Fitri, A. (2024). Digital adoption and professional learning of pesantren teachers. *International Journal of Educational Development*, 103, 102986. https://doi.org/10.1016/j.ijedudev.2023.102986

Sütfeld, L., Meinel, M., & Hildebrandt, K. (2023). AI as cognitive amplifier in professional reasoning. *AI & Society*, 38, 547–568. https://doi.org/10.1007/s00146-021-01344-z

Waycott, J., Gray, A., & Cordova, M. (2024). Teachers' adoption of generative AI for assessment. *Assessment in Education*, 31(2), 233–254. https://doi.org/10.1080/0969594X.2024.2323537

Zhang, D., Li, R., & Fisher, B. (2024). AI-assisted assessment design for higher-order thinking. *Computers & Education*, 204, 104919. https://doi.org/10.1016/j.compedu.2023.104919