DOI: https://doi.org/10.21580/jieed.v6i1.30802

Journal of Integrated
Elementary Education

eISSN 2776-1657 | pISSN 2828-223X | Vol. 6 No. 1 (2026) | pp. 173-193 | OPEN ACCESS

The effect of AI-assisted authentic assessment on students' learning growth in a primary school science classroom

Ari Hasan Ansori[1]*(✉), Ghina Siti Fitria Heriyanti[2], M. Arie Firdaus Heriansyah[3]

[1]Sekolah Tinggi Agama Islam Syekh Manshur Pandeglang, Indonesia
[2]Universitas Sultan Ageng Tirtayasa, Indonesia
[3]Universitas Islam Negeri Sultan Maulana Hasanudin Banten, Indonesia

*Corresponding author's email: cep.arie@gmail.com

Abstract

This study examines the implementation of artificial intelligence (AI)-assisted authentic assessment and its effect on the learning growth of elementary school students. The study employed a nonequivalent control-group quasi-experimental design involving 72 fifth-grade students. The experimental group received AI-assisted authentic assessment through performance tasks, inquiry-based projects, and open-ended activities, whereas the control group used conventional assessment methods. The findings indicate that the AI-assisted assessment system operated effectively. The natural language processing (NLP)-based system mapped students' levels of conceptual mastery, identified dominant errors, and generated diagnostic information that teachers used to provide rapid, specific, and adaptive feedback. The system also accelerated response analysis, monitored students' individual understanding, and promoted active cognitive engagement, enabling students to explain, connect, and reflect on their observational findings related to scientific concepts. The results further show that the experimental group achieved a high N-Gain score (0.72), whereas the control group obtained a moderate N-Gain score (0.43), with $t(70) = 11.704$, $p < 0.001$, and Cohen's $d = 2.76$, indicating a very large practical effect of AI-assisted authentic assessment on students' learning growth. These findings confirm that AI-assisted authentic assessment is effective in enhancing learning growth, strengthening the formative function of assessment, and supporting more personalized, reflective, and evidence-based science learning.

ARTICLE HISTORY

Received:
9 January 2026
Accepted:
28 March 2026
Published:
31 March 2026

KEYWORDS

authentic assessment; artificial intelligence; science learning; learning growth; primary school

1. Introduction

Authentic assessment constitutes an important component of science learning at the Madrasah Ibtidaiyah level because it not only measures learning outcomes but also supports meaningful, reflective, and sustainable learning processes (Agir et al., 2023; Herianingtyas et al., 2023). At the primary education level, students need to engage in learning tasks that integrate conceptual understanding, scientific reasoning, and the ability to apply knowledge in realworld contexts (Galanti & Holincheck, 2022; Salsabila et al., 2024; Wibawa et al., 2024). However, traditional assessment approaches that rely mainly on memorization or multiple-choice tests are often insufficient to capture the complexity of students' competencies and reasoning processes (Kasmi & Anasse, 2023; Poernomo et al., 2018). Consequently, the implementation of authentic assessment has become increasingly important in science instruction because it enables teachers to evaluate students' understanding through contextualized and performance-based tasks.

The theoretical foundation of authentic assessment is closely related to constructivist learning theory, which posits that knowledge is actively constructed through learners' interactions with meaningful experiences and contextual learning activities. In science education, constructivism emphasizes observation, inquiry, analysis, and reflection as fundamental processes through which students develop conceptual understanding and scientific reasoning. Authentic assessment reflects this perspective by engaging students in tasks that resemble real scientific practices, such as conducting observations, analyzing data, and explaining natural phenomena. Through these processes, assessment functions not merely as a measurement tool but as an integral component of the learning process that supports conceptual development and meaningful understanding.

Another important theoretical perspective underlying this study is assessment for learning, which emphasizes the formative role of assessment in supporting students' learning processes. According to this perspective, assessment should provide feedback that helps students recognize their strengths and misconceptions while guiding teachers in adjusting instructional strategies (Becerra & Cobos, 2025; Shen, 2025). Within this framework, assessment serves as a mechanism for monitoring learning progress, diagnosing conceptual difficulties, and facilitating continuous improvement in students' understanding. Therefore, authentic assessment can function effectively when it generates informative feedback that supports reflection, conceptual reconstruction, and deeper learning.

The concept of deep learning further strengthens the theoretical foundation of this study. Deep learning emphasizes profound conceptual understanding, the ability to connect ideas, and the transfer of knowledge to new contexts (Rasheed et al., 2023). In science learning, deep learning occurs when students actively engage in analyzing phenomena, constructing explanations, and linking empirical observations with scientific concepts. Authentic assessment is considered an appropriate approach for fostering deep learning because it requires students to demonstrate reasoning, interpretation, and conceptual integration rather than merely recalling factual information (Khojasteh et al., 2025; Saputra et al., 2024). As a result, authentic assessment not only measures students' knowledge but also supports the development of higher-order thinking skills and meaningful scientific understanding.

Despite its pedagogical advantages, the practical implementation of authentic assessment in Madrasah Ibtidaiyah science classrooms remains challenging. Teachers frequently encounter constraints related to limited time, heavy administrative workloads, and difficulties in consistently evaluating complex student tasks in heterogeneous classrooms (Melinda & Tanjung, 2022; Xin & Nasri, 2024). Assessing open-ended responses, inquiry projects, and performance tasks requires substantial effort and time, which often leads teachers to rely on simpler forms of assessment. Consequently, although authentic assessment is widely recommended in theory, its implementation in practice has not yet been fully optimized to enhance students' learning growth in primary education (Ansori & Heriansyah, 2025b; Hamel & Lee, 2024; Ravi & Besharat, 2026).

The integration of Artificial Intelligence (AI) into educational assessment offers promising opportunities to address these practical challenges. AI technologies, particularly those based on Natural Language Processing (NLP) and machine learning, are capable of analyzing students' open-ended responses, identifying patterns of conceptual understanding, and generating diagnostic feedback that can support teaching and learning processes (Arellanes et al., 2025; Kamalov et al., 2023; Pacala, 2023). Through automated analysis of textual responses, AI systems can map levels of conceptual mastery, detect common misconceptions, and provide information that assists teachers in delivering more targeted feedback. Such capabilities align closely with the principles of formative assessment, in which feedback serves as a key mechanism for improving learning outcomes.

In recent years, international studies have reported that AI-assisted assessment systems can enhance the efficiency of assessment processes, improve the quality of feedback, and support student engagement in complex learning tasks (Karim et al., 2025; Nagaraj, 2023; Zhai et al., 2023). AI technologies applied in STEM education are capable of analyzing learning data, monitoring student progress, and generating insights that support data-driven instructional decisions (Mahligawati et al., 2023; Nagaraj, 2023; Nuangchalerm, 2023). These systems enable teachers to process large volumes of student responses more efficiently while maintaining the depth of analysis required for authentic assessment practices.

However, most existing studies on AI-assisted assessment have focused on secondary and higher education contexts as well as broader STEM disciplines. Empirical evidence examining the implementation of AI-assisted authentic assessment in primary education, particularly in Madrasah Ibtidaiyah science classrooms, remains limited. Preliminary studies suggest that AI-based assessment systems can assist teachers in mapping students' conceptual understanding, identifying misconceptions, and accelerating automated scoring processes (Hopfenbeck et al., 2023; Lee et al., 2023; Riantoni, 2024). Other studies also highlight the potential of AI to provide adaptive feedback that supports reflection and enhances critical thinking skills (Mosher et al., 2024; Mulaudzi & Hamilton, 2025; N. Yu et al., 2025). Nevertheless, these studies primarily emphasize technological capabilities and assessment efficiency, while empirical investigations examining how AI-assisted authentic assessment influences students' learning growth in primary science education remain scarce.

This research gap is particularly important because learning growth provides a meaningful indicator of instructional effectiveness. Learning growth reflects improvements in students' understanding between the initial and final stages of learning and is commonly measured using normalized gain scores (Coletta & Steinert, 2020; Hake, 1998; Planinic et al., 2019). By integrating AI into authentic assessment practices, it becomes possible to analyze learning progress more efficiently while maintaining detailed diagnostic information regarding students' conceptual development. Such integration may enable teachers to monitor learning growth more systematically and provide timely interventions that support students' conceptual improvement (Hong, 2025; Karim et al., 2025; Łodzikowski et al., 2024).

Therefore, this study aims to examine the implementation of AI-assisted authentic assessment in science learning at the Madrasah Ibtidaiyah level and to analyze its effect on students' learning growth. Specifically, the study investigates how AI systems can process students' open-ended responses, map levels of conceptual mastery, and generate diagnostic feedback that supports instructional decision-making. By integrating AI-based diagnostic analysis with authentic assessment practices in primary science learning, this study provides empirical evidence on how AI-supported assessment can enhance students' learning growth. Furthermore, the study evaluates the extent to which AI-assisted authentic assessment contributes to students' learning growth, as measured through normalized gain (N-Gain) scores, and explores its implications for deep learning and active cognitive engagement.

This study contributes theoretically, empirically, and practically to the field of educational assessment. Theoretically, it strengthens the integration of constructivist learning theory, assessment for learning, and authentic assessment within the context of AI-supported assessment systems in primary education. Empirically, the findings provide quantitative evidence regarding the impact of AI-assisted authentic assessment on students' learning growth in science classrooms. Practically, the results offer guidance for teachers and policymakers in designing innovative, efficient, and data-driven assessment practices that support meaningful and sustainable learning in Madrasah Ibtidaiyah.

2. Methods

The study was conducted at Public Elementary School 1 (MIN1) in Pandeglang, Indonesia. The participants consisted of two parallel fifth-grade classes, each comprising 36 students. Class VA served as the experimental group and received AI-assisted authentic assessment integrated into science learning, whereas Class VB functioned as the control group and followed conventional assessment practices. In total, 72 students participated in the study. Initial equivalence between the two groups was confirmed using school academic records and pretest results, which indicated comparable baseline abilities. All procedures adhered to ethical research standards, including obtaining informed consent from both parents and students and ensuring the confidentiality of participants' identities.

The treatment instrument consisted of an authentic assessment package designed to measure students' conceptual understanding, scientific reasoning, and ability to present observational findings. The package included performance tasks, inquiry-based projects, and open-ended worksheets. In the performance tasks, students observed simple natural phenomena (e.g., changes in states of matter or environmental events), recorded data, identified basic variables, and related their observations to relevant scientific concepts. In the inquiry projects, students produced written reports describing procedures, analyses, and conclusions that connected empirical findings with theoretical explanations. The open-ended worksheets were designed to assess students' ability to explain causal relationships between their observations and relevant scientific concepts.

All authentic tasks were evaluated using an analytic rubric that assessed conceptual accuracy, procedural appropriateness, quality of reasoning, and presentation skills. The instruments and rubrics were validated by two experts: one specializing in educational evaluation and the other in science education. The validation results confirmed that the assessment instruments were aligned with the learning objectives and with the conceptual characteristics of the science content.

Students' learning growth was measured using a science achievement test administered as both a pretest and a posttest. The test consisted of 14 items developed based on science learning indicators. A pilot study conducted at Public Elementary School 2 (MIN 2) in Pandeglang indicated that 11 items met the validity criteria, while 3 were discarded. Item discrimination analysis showed that 10 items had high discrimination, 1 item had moderate discrimination, and 3 items had low discrimination. The difficulty index indicated that 5 items were difficult, 7 items were of moderate difficulty, and 2 items were easy. The reliability coefficient was 0.78, indicating high reliability and confirming the suitability of the instrument for calculating N-Gain scores (Ansori, 2025). In addition, a classroom observation sheet was also used to monitor the fidelity of the AI-assisted authentic assessment implementation, student engagement during learning activities, and the consistency of the treatment throughout the instructional process.

The research procedure consisted of three stages. First, both groups completed the pretest to determine their initial ability levels and to confirm group equivalence. Second, the experimental group participated in science learning integrated with AI-assisted authentic assessment, whereas the control group followed conventional assessment practices. Classroom observations were conducted during this stage to ensure the consistency and fidelity of the instructional treatment. Third, both groups completed the posttest. The pretest and posttest scores were subsequently used to calculate students' N-Gain as an indicator of learning growth.

Data analysis involved calculating N-Gain to measure students' learning growth. Descriptive statistics, including the mean and standard deviation, were computed for both groups. Prior to hypothesis testing, the assumptions of normality (Shapiro–Wilk test) and homogeneity of variance (Levene's test) were examined (Ansori, 2021). An independent-samples t-test was conducted to compare the N-Gain scores between the experimental and control groups (Ansori, 2021). Statistical significance was determined based on the results of the t-test, and Cohen's d was calculated to determine the magnitude of the practical effect size.

3. Results

3.1 Implementation of AI-assisted authentic assessment in science learning

Authentic assessment was implemented through performance tasks and inquiry-based projects. The performance tasks involved observations of simple natural phenomena, such as changes in the states of matter (melting, evaporation, and freezing) and environmental events occurring around the students. During these activities, students were required to record observational data, identify basic variables (e.g., temperature and time), and relate their findings to previously learned scientific concepts.

The second form of assessment consisted of inquiry-based projects that required students to prepare written observational reports. These reports included descriptions of the observation procedures, simple analyses of the observed phenomena, and conclusions linking empirical findings with scientific concepts. Through this format, the assessment did not merely evaluate final answers but also examined students' reasoning processes in connecting observational data with scientific explanations.

In addition, students completed open-ended worksheets that required them to explain the relationships between their observations and relevant scientific concepts. Analysis of the worksheets indicated that 87% of students were able to accurately document their observations and appropriately relate them to the concepts studied, whereas 13% still required additional guidance, particularly in articulating scientifically grounded cause-and-effect relationships.

The distribution of conceptual mastery levels among students in the experimental class is presented in Figure 1. The graph illustrates that the majority of students fell into the high and moderate conceptual mastery categories. This distribution pattern indicates that the varied forms of authentic assessment revealed differences in students' levels of understanding in greater detail than conventional, closed-ended, test-based assessments.

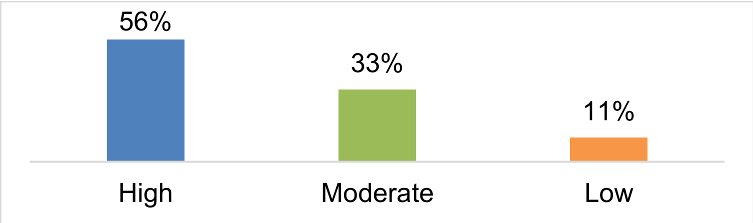

Fig. 1. Distribution of conceptual mastery levels of students in the experimental group

The distribution facilitated teachers in identifying students who had achieved strong conceptual mastery and those who still required reinforcement. Thus, AI-assisted authentic assessment did not merely generate scores but also mapped levels of conceptual mastery both individually and collectively.

The AI system utilized in this study was a web-based application equipped with Natural Language Processing (NLP) capabilities and learning analytics features. The system was used to process students' open-ended written responses and map their levels of conceptual mastery using rubric indicators.

The analysis process involved extracting conceptual keywords relevant to the instructional content, such as melting, evaporation, freezing, heat, and cold. After identifying these keywords and analyzing the structure of the explanations, the system categorized students' responses into three levels of conceptual mastery: high, moderate, and low.

In addition to categorical mapping, the AI system generated summaries of dominant conceptual errors appearing in students' responses. These summaries served as a basis for teachers to determine focal points for conceptual reinforcement in subsequent lessons. The mapping data indicated that the majority of students in the experimental class were categorized as having high to moderate levels of conceptual mastery, while only a small proportion fell into the low category.

The information generated by the AI system was used by teachers to adjust instructional strategies, for example, by reinforcing concepts that were frequently misunderstood and by developing additional examples tailored to students' learning needs. In this way, AI functioned as an assessment support system that provided rapid and structured diagnostic information.

The implementation of AI-assisted authentic assessment was carried out through systematic stages illustrated in Figure 2. The diagram depicts the sequence of activities, beginning with the planning of authentic tasks, followed by students' observational activities, data collection, analysis of responses using the AI system, and the provision of feedback by the teacher.

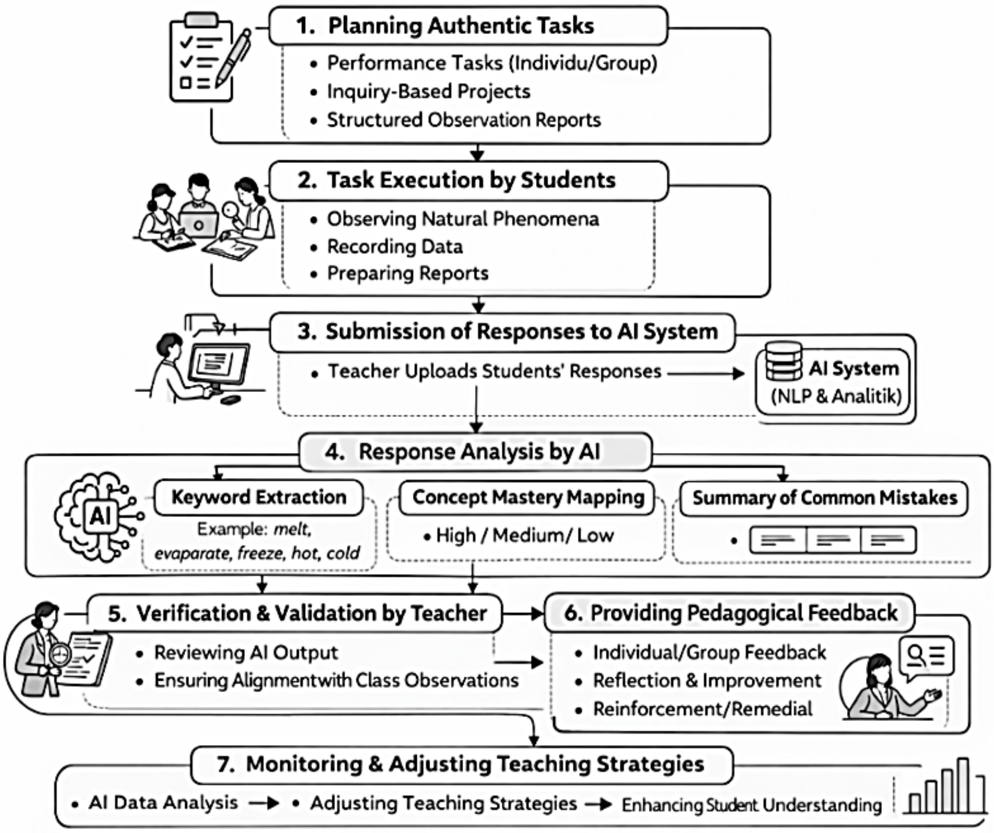

Fig. 2. Flow of the AI-assisted authentic assessment implementation procedure.

The first stage involved students implementing authentic tasks. Students conducted observations of natural phenomena according to the teacher's instructions, recorded their findings, and constructed explanations based on scientific concepts. In the diagram, this stage is represented as the Student Observation and Data Recording phase, which is directly connected to the Submission to AI System phase.

In the subsequent stage, students' responses were uploaded to the AI system and analyzed according to the assessment rubric encompassing conceptual accuracy, procedural appropriateness, quality of reasoning, and presentation skills. This stage is represented as the AI Assessment and Concept Mapping phase. The analysis results were then verified by the teacher during the Teacher Review and Feedback phase.

The final stage involved providing feedback to students. Teachers utilized the summarized AI analysis results to deliver specific feedback, while students engaged in reflection and revised their responses in subsequent tasks. The diagram illustrates a feedback loop connecting teacher feedback with student reflection, indicating that the assessment process operates continuously.

The synthesis of the findings indicates that AI-assisted authentic assessment was implemented through a combination of performance tasks, inquiry-based projects, and structured observational reports. The AI system, based on Natural Language Processing (NLP) and learning analytics, was capable of processing students' open-ended responses, mapping levels of conceptual mastery, and generating summaries of dominant errors.

The analysis data revealed that the majority of students were categorized within high to moderate levels of conceptual mastery. This information enabled teachers to target conceptual reinforcement more precisely. Furthermore, the use of AI accelerated the processes of response analysis and feedback provision, ensuring that the assessment not only generated scores but also functioned as a formative source of information that supported science learning in a measurable and evidence-based manner.

3.2 The effect of AI-assisted authentic assessment on students' learning growth

Students' learning growth was analyzed using the N-Gain score, which represents the normalized increase in scores from the pre-test to the post-test relative to the maximum possible score. This analysis was conducted to examine changes in students' performance following instructional activities in the experimental and control groups.

Pre-test and post-test scores were first examined to describe students' baseline and final performance. As shown in Table 2, the pre-test results indicated that the experimental group (M = 45.00, SD = 8.50) and the control group (M = 44.50, SD = 8.70) had comparable initial scores.

Following the instructional treatment, post-test scores increased in both groups. The experimental group obtained a mean score of 85.00 (SD = 7.20), while the control group obtained a mean score of 69.50 (SD = 7.80).

To quantify learning growth, N-Gain scores were calculated for both groups. The experimental group achieved a mean N-Gain of 0.72 (SD = 0.10; median =

0.73), which falls within the high category. A total of 83% of students in this group were categorized as having high learning growth, while the remaining students were in the moderate category. No students were categorized in the low group.

In the control group, the mean N-Gain was 0.43 (SD = 0.11; median = 0.42), corresponding to the moderate category. In this group, 28% of students were categorized as having high learning growth, while the majority were in the moderate category.

Table 2.
Students' Pre-test, Post-test, and N-Gain by Group

Group	N	Pre-test		Post-test		N-Gain		Category
		Mean	SD	Mean	SD	Mean	SD	
Experimental	36	45.00	8.50	85.00	7.20	0.72	0.10	High
Control	36	44.50	8.70	69.50	7.80	0.43	0.11	Moderate

The data presented in Table 2 show that the experimental group had smaller standard deviations in post-test and N-Gain scores compared to the control group. In addition, observational data showed differences in assessment time between groups. The average time required for teachers to review students' work and provide feedback was approximately 20 minutes per session in the experimental group and approximately 45 minutes in the control group, representing a reduction of approximately 55%.

Analysis of students' assignments showed that approximately 75% of students in the experimental group revised their responses or observational reports in subsequent tasks following feedback. These revisions included changes in response structure, use of scientific terminology, and conceptual explanations in follow-up tasks.

Prior to testing differences in mean N-Gain between groups, the data were examined for normality and homogeneity of variance. The Shapiro–Wilk test indicated that the N-Gain distributions in both groups were normal ($p > 0.05$), while Levene's test indicated that the assumption of homogeneity of variances was met ($p = 0.32$). Based on these results, an independent-samples t-test was conducted.

The comparison of mean N-Gain scores between the experimental group ($M = 0.72$, $SD = 0.10$, $n = 36$) and the control group ($M = 0.43$, $SD = 0.11$, $n = 36$) yielded $t(70) = 11.704$, $p < 0.001$. The mean difference between the two groups was 0.29, with a 95% confidence interval ranging from 0.24 to 0.34.

The effect size for the difference in N-Gain scores between the experimental and control groups, calculated using Cohen's d, was 2.76, which falls within the very large category, indicating that the magnitude of the difference substantially exceeded the variability within each group.

4. Discussion

The findings of this study indicate that implementing AI-assisted authentic assessment in science learning at the Madrasah Ibtidaiyah level significantly contributes to students' learning growth. Authentic assessment was implemented through performance tasks, inquiry-based projects, and open-ended worksheets that required students to observe natural phenomena, analyze empirical data, and formulate explanations connecting their observations with scientific concepts. The integration of AI enabled the processing of students' responses, the mapping of conceptual mastery levels, and the generation of summaries of dominant conceptual errors that served as the basis for teacher feedback. In this context, AI functioned not merely as an assessment tool but also as a pedagogical mediator that provided diagnostic information to support instructional decision-making (Bond et al., 2024; Park & Doo, 2024; Wang et al., 2024).

These findings can be interpreted from the perspective of constructivist learning theory, which emphasizes that knowledge is actively constructed through learners' engagement in meaningful learning experiences. In science education, constructivism highlights the importance of observation, inquiry, and reflection in the development of conceptual understanding. In the present study, authentic assessment tasks required students to observe real phenomena, interpret data, and construct explanations that linked empirical findings with theoretical concepts. Such learning activities provided opportunities for students to actively construct knowledge rather than passively receive information. This process aligns with the constructivist perspective that meaningful learning occurs when students engage directly with contextualized tasks and reflect on their experiences (Akmam et al., 2025; Wodaj, 2020).

The results also support the principles of assessment for learning, which emphasize the formative role of assessment in supporting learning processes. According to this perspective, assessment should function not only as a tool for measuring achievement but also as a mechanism for generating feedback that helps students recognize misconceptions and improve their understanding. The AI system utilized in this study strengthened the formative function of assessment by analyzing students' open-ended responses and identifying patterns of conceptual understanding. Through Natural Language Processing (NLP) techniques, the system extracted conceptual keywords, categorized responses according to levels of conceptual mastery, and generated summaries of common misconceptions (Gao et al., 2024; Maity & Deroy, 2024; Riantoni, 2024; Somers et al., 2021).

The diagnostic information generated by the AI system enabled teachers to provide more targeted and timely feedback. This mechanism reflects the feedback model proposed by John Hattie and Helen Timperley (2007), which highlights the importance of feedback in improving learning outcomes. By identifying dominant conceptual errors and patterns of misunderstanding, teachers were able to focus instructional reinforcement on specific concepts that required clarification. Consequently, assessment became an integral component of the learning process rather than merely a tool for evaluating final performance.

The implementation of AI-assisted authentic assessment also contributed to the development of deep learning. Deep learning refers to the development of meaningful conceptual understanding through active engagement, analysis, and the integration of knowledge across contexts (Rasheed et al., 2023). In the present study, students were required to connect empirical observations with scientific concepts, explain causal relationships, and construct written reports based on their findings. These activities stimulated higher-order cognitive processes such as analysis, evaluation, and synthesis, which are central components of deep learning. The relatively high N-Gain observed in the experimental group (0.72) suggests that the integration of authentic tasks with AI-supported feedback strengthened students' conceptual understanding and promoted deeper engagement with scientific knowledge (Akolekar et al., 2025; Joseph et al., 2025; Liu et al., 2025; Sukmawati & Wahjusaputri, 2024).

Another important aspect revealed in this study is the role of AI-assisted assessment in supporting self-regulated learning. The conceptual mastery mapping generated by the AI system provided students with information about their learning progress, allowing them to reflect on their understanding and revise their responses accordingly. Observational data indicated that approximately 75% of students in the experimental group revised their assignments after receiving AI-based feedback. This process reflects metacognitive engagement, in which students actively monitor their learning processes and adjust their strategies to improve their understanding. Previous studies have reported that AI-based learning systems can strengthen self-regulated learning by providing personalized feedback and diagnostic information that help learners adapt their learning strategies (Halkiopoulos & Gkintoni, 2024; Su et al., 2025; Wei, 2023).

The distribution of N-Gain scores in the experimental group was also more concentrated in the high category compared with that of the control group. This pattern suggests that AI-assisted authentic assessment contributed not only to higher average learning gains but also to more equitable learning outcomes among students. Adaptive feedback enabled teachers to provide additional support to students who initially demonstrated lower levels of understanding, thereby reducing learning disparities within the classroom. These findings are consistent with previous research indicating that AI-supported assessment systems can facilitate personalized learning and reduce variability in students' academic achievement (Abrar et al., 2025; Ansori & Heriansyah, 2025a; Naseer & Khawaja, 2025).

The overall mechanism through which AI-assisted authentic assessment supports learning growth can be understood as an interaction between authentic learning tasks, AI-based diagnostic analysis, and teacher feedback. This mechanism is illustrated in Figure 3, which presents the conceptual model of how AI-assisted authentic assessment operates in science learning.

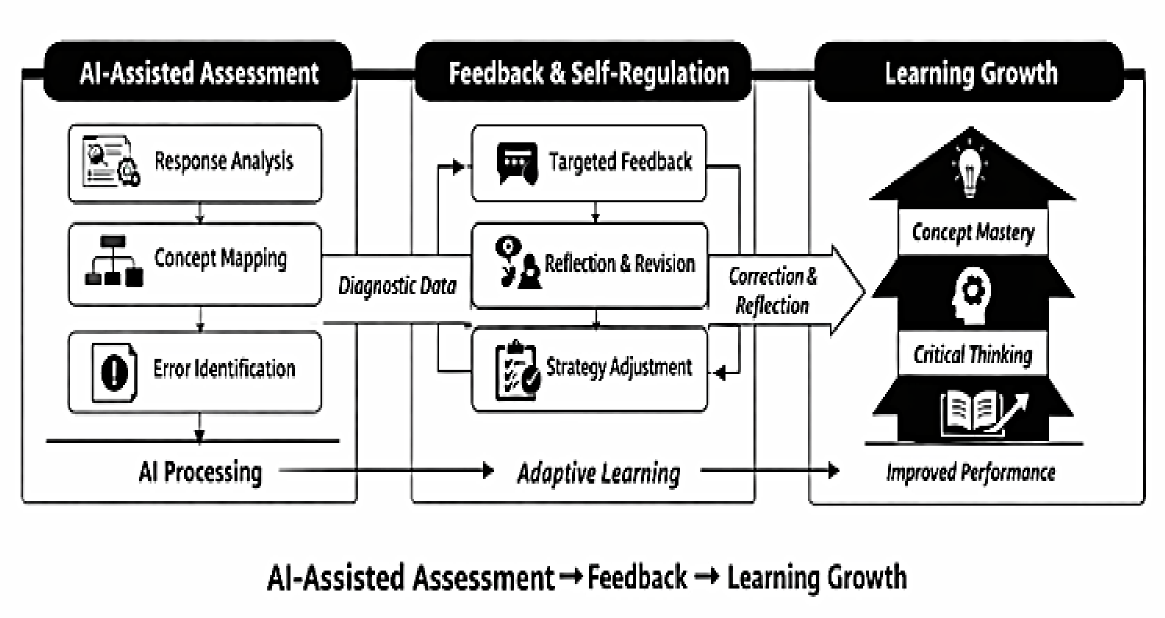

Fig 3. Conceptual mechanism of AI-assisted authentic assessment in promoting students' learning growth

As illustrated in Figure 3, the process begins with students' engagement in authentic tasks such as observation activities, inquiry-based projects, and open-ended responses. These tasks encourage students to analyze natural phenomena, interpret observational data, and construct explanations that connect empirical findings to scientific concepts. Students' responses are subsequently processed by the AI system, which analyzes conceptual keywords, evaluates the structure of explanations, and identifies patterns of conceptual understanding.

The AI system then generates diagnostic information in the form of conceptual mastery mapping and summaries of dominant misconceptions. This information provides teachers with insights into students' learning difficulties and enables them to deliver targeted feedback and conceptual reinforcement. Through this feedback loop, students are encouraged to reflect on their understanding, revise their explanations, and improve subsequent learning tasks. The interaction between authentic learning activities, AI-supported analysis, and feedback-driven instruction creates a continuous learning cycle that fosters conceptual reconstruction and deeper engagement with scientific knowledge.

The statistical results further confirm the effectiveness of this mechanism. The independent-samples t-test indicated a significant difference between the experimental and control groups, t (70) = 11.704, $p < 0.001$, while the effect size (Cohen's d = 2.76) indicated a very large practical impact. Such a large effect size suggests that AI-assisted authentic assessment substantially improved students' learning growth in science learning. These results are consistent with previous studies reporting that AI-supported assessment systems can enhance learning outcomes by providing rapid diagnostic feedback and enabling adaptive instructional decisions (Abrar et al., 2025; Khan et al., 2024; Song & Song, 2023).

In addition to improving students' learning growth, the use of AI-assisted authentic assessment also increased the efficiency of teachers' assessment practices. The time required for teachers to analyze student responses and provide feedback was reduced by approximately 55%, allowing teachers to allocate more time to instructional interaction and conceptual reinforcement. This efficiency highlights the dual role of AI as both an analytical tool and a pedagogical facilitator that supports the implementation of authentic assessment in classroom practice (Anaroua et al., 2025; Ifelebuegu, 2023; Karim et al., 2025).

From a broader pedagogical perspective, the findings of this study demonstrate that AI-assisted authentic assessment can create a learning ecosystem that integrates contextual learning tasks, diagnostic feedback, and adaptive instructional support. Authentic tasks provide meaningful learning contexts that stimulate inquiry and conceptual exploration, while AI technologies facilitate the analysis of students' responses and generate actionable information for teachers. The interaction between these components supports continuous improvement in students' understanding and promotes deeper engagement with scientific knowledge (Kotsis, 2024; Kusumawati et al., 2025; J. Yu, 2023).

The findings of this study suggest that AI-assisted authentic assessment constitutes a promising pedagogical innovation in science learning at the Madrasah Ibtidaiyah level. Situated within the expanding field of artificial intelligence in education, this study contributes empirical evidence at the primary education level by demonstrating how the integration of authentic assessment with AI-based diagnostic feedback enhances learning processes. It extends prior research by not only confirming improvements in learning outcomes but also elucidating the mechanisms through which AI facilitates conceptual mapping, accelerates feedback processes, and supports iterative student revision.

From a practical perspective, these findings highlight several actionable implications for classroom assessment. The integration of AI enables teachers to analyze students' open-ended responses more efficiently while maintaining the depth and quality of feedback. AI-generated diagnostic information supports more targeted instructional interventions, particularly in addressing persistent misconceptions. Moreover, the use of iterative feedback cycles encourages students to revise their work, thereby fostering self-regulated learning and deeper conceptual understanding. These findings indicate that AI-assisted authentic assessment can serve as a viable strategy for strengthening formative assessment practices in primary science education.

However, several challenges and limitations should be acknowledged. The effectiveness of AI-assisted assessment depends on the availability of adequate technological infrastructure, which may be uneven across educational contexts. In addition, teachers' digital literacy and their ability to interpret and apply AI-generated feedback are critical for successful implementation. The study is also limited by its specific context and reliance on written response data, which may not fully capture other dimensions of learning, such as oral communication, collaboration, and affective engagement. Future research should therefore explore multimodal assessment approaches and broader applications across subjects and educational levels.

Despite these limitations, this study underscores the role of AI as a pedagogical support system rather than a replacement for teachers. By enhancing the provision of timely, data-informed, and adaptive feedback, AI-assisted authentic assessment contributes to the development of a more responsive and evidence-based learning environment, ultimately supporting continuous improvement in students' conceptual understanding and overall learning growth.

5. Conclusion

This study indicates that implementing AI-assisted authentic assessment in science learning at the elementary school level enhances students' learning growth and conceptual understanding. The integration of authentic tasks with AI-supported diagnostic feedback facilitated learning processes in which students actively constructed knowledge, engaged in higher-order thinking, and refined their understanding through feedback-driven revision.

The findings suggest that AI functions as a pedagogical support system rather than a replacement for teachers. By enabling rapid analysis of student responses and generating structured diagnostic information, AI strengthened the formative role of assessment and supported more targeted, timely, and adaptive feedback. Assessment practices shifted from a predominantly summative orientation toward a continuous, learning-centered process that fosters self-regulated learning and cognitive development.

This study contributes empirical evidence to the field of artificial intelligence in primary education. Its findings are limited by the specific research context and focus on written responses, which may not capture broader dimensions of student learning. Future research should examine AI-assisted authentic assessment across diverse educational settings, integrate multimodal assessment approaches, and investigate long-term effects through longitudinal studies.

AI-assisted authentic assessment has potential as a viable and scalable approach to improving formative assessment practices in primary education. When implemented with adequate infrastructure and supported by teachers' professional competence, AI can contribute to more responsive, data-informed learning environments that support continuous improvement in students' learning outcomes.

References

Abrar, M., Aboraya, W., Khaliq, R. A., Subramanian, K. P., Husaini, Y. A., & Husaini, M. A. (2025). AI-Powered Learning Pathways: Personalized Learning and Dynamic Assessments. *International Journal of Advanced Computer Science and Applications*, *16(1).* 454-462. https://doi.org/10.14569/IJACSA.2025.0160145

Agir, N., Effendi, M., Matore, E. M., Eteuati, N. F., & Marquez, N. (2023). Outcome-Based Assessment in The Evaluation of Education Programs Through a Systematic Literature Review. *International Journal of Academic Research in Progressive Education and Development*, *12*(2), 2662-2677. https://doi.org/10.6007/IJARPED/v12-i2/18095

Akmam, A., Ahzari, S., Emiliannur, E., Anshari, R., & Setiawan, D. (2025). Enhancing Science Literacy Through Cognitive Conflict-Based Generative Learning Model: An Experimental Study in Physics Learning. *Social Science and Humanities Journal*, *9*(07), 8507–8522. https://doi.org/10.18535/sshj.v9i07.1934

Akolekar, H., Jhamnani, P., Kumar, V., Tailor, V., Pote, A., Meena, A., Kumar, K., Challa, J. S., & Kumar, D. (2025). The role of generative AI tools in shaping mechanical engineering education from an undergraduate perspective. *Scientific Reports*, *15*(1), 9214. https://doi.org/10.1038/s41598-025-93871-z

Anaroua, F. I., Li, Q., Tang, Y., & Liu, H. P. (2025). *AI-driven formative assessment and adaptive learning in data-science education: Evaluating an LLM-powered virtual teaching assistant.* arXiv. https://doi.org/10.48550/ARXIV.2509.20369

Ansori, A. H. (2021). *Statistika Penelitian*. Staisman Press.

Ansori, A. H. (2025). *Desain & Analisi Instrumen Penelitian Teori, Praktik Lapangan, dan Aplikasi SPSS.* Yayasan Putra Adi Dharma.

Ansori, A. H., & Heriansyah, M. A. F. (2025a). Evaluasi Pendidikan Islam dalam Teori Modern: Integrasi Teknologi dan Nilai Karakter. In *Transformasi Pendidikan Islam: Filosofi, Nilai, dan Inovasi Menuju Masyarakat Berkeadaban*. Zahir Publishing.

Ansori, A. H., & Heriansyah, M. A. F. (2025b). *Transformasi Pembelajaran Abad 21: Sinergi Proyek Kontekstual dan Penilaian Autentik Mewujudkan Pembelajaran Mendalam.* Goresan Pena.

Arellanes, F. E. T., Strycker, L., Alvez, G. G., Miller, B., & Vargas, K. (2025). Promoting Agency Among Upper Elementary School Teachers and Students with an Artificial Intelligence Machine Learning System to Score Performance-Based Science Assessments. *Education Sciences*, *15*(1), 54. https://doi.org/10.3390/educsci15010054

Becerra, A., & Cobos, R. (2025). *Leveraging Peer, Self, and Teacher Assessments for Generative AI-Enhanced Feedback.* arXiv. https://doi.org/10.48550/ARXIV.2512.18306

Bond, M., Khosravi, H., Laat, M. D., Bergdahl, N., Negrea, V., Oxley, E., Pham, P., Chong, S. W., & Siemens, G. (2024). A meta systematic review of artificial intelligence in higher education: A call for increased ethics, collaboration, and rigour. *International Journal of Educational Technology in Higher Education*, *21*(1), 4. https://doi.org/10.1186/s41239-023-00436-z

Coletta, V. P., & Steinert, J. J. (2020). Why normalized gain should continue to be used in analyzing preinstruction and postinstruction scores on concept inventories. *Physical Review Physics Education Research*, *16*(1), 010108. https://doi.org/10.1103/PhysRevPhysEducRes.16.010108

Galanti, T. M., & Holincheck, N. (2022). Beyond content and curriculum in elementary classrooms: Conceptualizing the cultivation of integrated STEM teacher identity. *International Journal of STEM Education*, *9*(1), 43. https://doi.org/10.1186/s40594-022-00358-8

Gao, R., Guo, X., Li, X., Narayanan, A. B. L., Thomas, N., & Srinivasa, A. R. (2024). *Towards Scalable Automated Grading: Leveraging Large Language Models for Conceptual Question Evaluation in Engineering.* arXiv. https://doi.org/10.48550/ARXIV.2411.03659

Hake, R. R. (1998). Interactive-engagement versus traditional methods: A six-thousand-student survey of mechanics test data for introductory physics courses. *American Journal of Physics*, *66*(1), 64–74. https://doi.org/10.1119/1.18809

Halkiopoulos, C., & Gkintoni, E. (2024). Leveraging AI in E-Learning: Personalized Learning and Adaptive Assessment through Cognitive Neuropsychology—A Systematic Analysis. *Electronics*, *13*(18), 3762. https://doi.org/10.3390/electronics13183762

Hamel, P., & Lee, W. K. (2024). Supporting the evaluation of authentic assessment in environmental sciences: A case study. *Cogent Education*, *11*(1), 2399433. https://doi.org/10.1080/2331186X.2024.2399433

Hattie, J., & Timperley, H. (2007). The Power of Feedback. *Review of Educational Research*, *77*(1), 81–112. https://doi.org/10.3102/003465430298487

Herianingtyas, N. L. R., Rohman, C., Widiyanto, R., & Amarulloh, R. R. (2023). Evaluation of the Implementation of Science Literacy-Based Learning in Madrasah Ibtidaiyah. *JMIE (Journal of Madrasah Ibtidaiyah Education)*, *7*(2), 92. https://doi.org/10.32934/jmie.v7i2.602

Hong, L. (2025). Development and validation of a competency-based ladder pathway for AI literacy enhancement among higher vocational students. *Scientific Reports*, *15*(1), 29898. https://doi.org/10.1038/s41598-025-15202-6

Hopfenbeck, T. N., Zhang, Z., Sun, S. Z., Robertson, P., & McGrane, J. A. (2023). Challenges and opportunities for classroom-based formative assessment and AI: A perspective article. *Frontiers in Education*, *8*, 1270700. https://doi.org/10.3389/feduc.2023.1270700

Ifelebuegu, A. (2023). Rethinking online assessment strategies: Authenticity versus AI chatbot intervention. *Journal of Applied Learning and Teaching*, *6*(2), 385–392. https://doi.org/10.37074/jalt.2023.6.2.2

Joseph, T. S., Gowrie, S., Montalbano, M. J., Bandelow, S., Clunes, M., Dumont, A. S., Iwanaga, J., Tubbs, R. S., & Loukas, M. (2025). The Roles of Artificial Intelligence in Teaching Anatomy: A Systematic Review. *Clinical Anatomy*, *38*(5), 552–567. https://doi.org/10.1002/ca.24272

Kamalov, F., Calonge, D. S., & Gurrib, I. (2023). New Era of Artificial Intelligence in Education: Towards a Sustainable Multifaceted Revolution. *Sustainability*, *15*(16), 12451. https://doi.org/10.3390/su151612451

Karim, A. N., Isa, C. M. M., & Noor, S. M. (2025). A Methodological Framework for AI-Integrated Alternative Assessments in Engineering Education. *Jurnal Kejuruteraan*, *37*(2), 807–819. https://doi.org/10.17576/jkukm-2025-37(2)-20

Kasmi, H., & Anasse, K. (2023). The Status of Alternative Assessment in Morocco: Teachers' Attitudes and Obstacles. *International Journal of Language and Literary Studies*, *5*(1), 300–311. https://doi.org/10.36892/ijlls.v5i1.1189

Khan, M. A., Kurbonova, O., Abdullaev, D., Radie, A. H., & Basim, N. (2024). Is AI-assisted assessment liable to evaluate young learners? Parents support, teacher support, immunity, and resilience are in focus in testing vocabulary learning. *Language Testing in Asia*, *14*(1), 48. https://doi.org/10.1186/s40468-024-00324-x

Khojasteh, L., Kafipour, R., Pakdel, F., & Mukundan, J. (2025). Empowering medical students with AI writing co-pilots: Design and validation of AI self-assessment toolkit. *BMC Medical Education*, *25*(1), 159. https://doi.org/10.1186/s12909-025-06753-3

Kotsis, K. T. (2024). Chatgpt in Teaching Physics Hands-On Experiments in Primary School. *European Journal of Education Studies*, *11*(10). https://doi.org/10.46827/ejes.v11i10.5549

Kusumawati, E., Suswandari, & Umam, K. (2025). Strengthening teacher competence for leading and sustaining the implementation of the Merdeka Curriculum. *Cogent Education*, *12*(1), 2501458. https://doi.org/10.1080/2331186X.2025.2501458

Lee, C.-A., Huang, N.-F., Tzeng, J.-W., & Tsai, P.-H. (2023). AI-Based Diagnostic Assessment System: Integrated with Knowledge Map in MOOCs. *IEEE Transactions on Learning Technologies*, *16*(5), 873–886. https://doi.org/10.1109/TLT.2023.3308338

Liu, A., Esbenshade, L., Sarkar, S., Tian, Z., Sun, M., Zhang, Z., Han, T., Lapicus, Y., & He, K. (2025). *AI as a Teaching Partner: Early Lessons from Classroom Codesign with Secondary Teachers*. arXiv. https://doi.org/10.48550/ARXIV.2512.12045

Łodzikowski, K., Foltz, P. W., & Behrens, J. T. (2024). *Generative AI and Its Educational Implications*. arXiv. https://doi.org/10.48550/ARXIV.2401.08659

Mahligawati, F., Allanas, E., Butarbutar, M. H., & Nordin, N. A. N. (2023). Artificial intelligence in Physics Education: A comprehensive literature review. *Journal of Physics: Conference Series*, *2596*(1), 012080. https://doi.org/10.1088/1742-6596/2596/1/012080

Maity, S., & Deroy, A. (2024). *The Future of Learning in the Age of Generative AI: Automated Question Generation and Assessment with Large Language Models*. arXiv. https://doi.org/10.48550/ARXIV.2410.09576

Melinda, N. E., & Tanjung, I. F. (2022). Analysis of Impelementing Authentic Assessment of Curriculum 2013 by High School Biologi Teachers. *Jurnal Pelita Pendidikan*, *10*(2). 28-35. https://doi.org/10.24114/jpp.v10i2.34500

Mosher, M., Dieker, L., & Hines, R. (2024). The Past, Present, and Future Use of Artificial Intelligence in Teacher Education. *Journal of Special Education Preparation*, *4*(2), 6–17. https://doi.org/10.33043/8aa9855b

Mulaudzi, L. V., & Hamilton, J. (2025). Lecturer's Perspective on the Role of AI in Personalized Learning: Benefits, Challenges, and Ethical Considerations in Higher Education. *Journal of Academic Ethics*, *23*(4), 1571–1591. https://doi.org/10.1007/s10805-025-09615-1

Nagaraj, B. K. (2023). The Emerging Role of Artificial Intelligence in STEM Higher Education: A Critical Review. *International Research Journal of Multidisciplinary Technovation*, 1–19. https://doi.org/10.54392/irjmt2351

Naseer, F., & Khawaja, S. (2025). Mitigating Conceptual Learning Gaps in Mixed-Ability Classrooms: A Learning Analytics-Based Evaluation of AI-Driven Adaptive Feedback for Struggling Learners. *Applied Sciences*, *15*(8), 4473. https://doi.org/10.3390/app15084473

Nuangchalerm, P. (2023). AI-Driven Learning Analytics in STEM Education. *International Journal of Research in STEM Education*, *5*(2), 77–84. https://doi.org/10.33830/ijrse.v5i2.1596

Pacala, F. A. A. (2023). Artificial intelligence in a modernizing science and technology education: A textual narrative synthesis in the COVID-19 era. *Journal of Physics: Conference Series*, *2611*(1), 012028. https://doi.org/10.1088/1742-6596/2611/1/012028

Park, Y., & Doo, M. Y. (2024). Role of AI in Blended Learning: A Systematic Literature Review. *The International Review of Research in Open and Distributed Learning*, *25*(1), 164–196. https://doi.org/10.19173/irrodl.v25i1.7566

Planinic, M., Boone, W. J., Susac, A., & Ivanjek, L. (2019). Rasch analysis in physics education research: Why measurement matters. *Physical Review Physics Education Research*, *15*(2), 020111. https://doi.org/10.1103/PhysRevPhysEducRes.15.020111

Poernomo, J. B., Wiyanto, M., Rusilowati, A., & Saptono, S. (2018). The Development of Integrated Science Learning Instrument Based on Project-Based Learning to Measure Critical Thinking Skills. *Proceedings of the International Conference on Science and Education and Technology 2018 (ISET 2018)*. Proceedings of the International Conference on Science and Education and Technology 2018 (ISET 2018). https://doi.org/10.2991/iset-18.2018.57

Rasheed, H., A., Weber, C., & Fathi, M. (2023). Context based learning: A survey of contextual indicators for personalized and adaptive learning recommendations – a pedagogical and technical perspective. *Frontiers in Education*, *8*, 1210968. https://doi.org/10.3389/feduc.2023.1210968

Ravi, M., & Besharat, M. (2026). A holistic consideration of authentic assessments: Student perception of assessment design, delivery, flexibility and creativity. *European Journal of Engineering Education*, *51*(1), 25–42. https://doi.org/10.1080/03043797.2025.2480116

Riantoni, C. (2024). A Hybrid Automatic Scoring System: Artificial Intelligence-Based Evaluation of Physics Concept Comprehension Essay Test. *International Journal of Information and Education Technology*, *14*(6), 876–882. https://doi.org/10.18178/ijiet.2024.14.6.2113

Salsabila, S., Setiyono, Y. C. P., Damayani, A. T., & Azizah, M. (2024). Implementation of Science Literacy Through Eclipse Diorama Project in

Grade VI at Supriyadi Elementary School Semarang. *Jurnal Sains Sosio Humaniora*, *8*(1), 11–23. https://doi.org/10.22437/jssh.v8i1.33955

Saputra, I., Kurniawan, A., Yanita, M., Yeni Putri, E., & Mahniza, M. (2024). The Evolution of Educational Assessment: How Artificial Intelligence is Shaping the Trends and Future of Learning Evaluation. *The Indonesian Journal of Computer Science*, *13*(6). https://doi.org/10.33022/ijcs.v13i6.4465

Shen, H. (2025). *Bidirectional Human-AI Alignment in Education for Trustworthy Learning Environments*. arXiv. https://doi.org/10.48550/ARXIV.2512.21552

Somers, R., Nelson, S. C., & Boles, W. (2021). Applying natural language processing to automatically assess student conceptual understanding from textual responses. *Australasian Journal of Educational Technology*, *37*(5), 98–115. https://doi.org/10.14742/ajet.7121

Song, C., & Song, Y. (2023). Enhancing academic writing skills and motivation: Assessing the efficacy of ChatGPT in AI-assisted language learning for EFL students. *Frontiers in Psychology*, *14*, 1260843. https://doi.org/10.3389/fpsyg.2023.1260843

Su, M., Dang, B., Nguyen, A., & Nagashima, T. (2025). Choice-making in an adaptive learning system with motivational pedagogical agents. *Npj Science of Learning*, *10*(1), 77. https://doi.org/10.1038/s41539-025-00366-7

Sukmawati, W., & Wahjusaputri, S. (2024). Integrating RADEC Model and AI to Enhance Science Literacy: Student Perspectives. *Jurnal Penelitian Pendidikan IPA*, *10*(6), 3080–3089. https://doi.org/10.29303/jppipa.v10i6.7557

Wang, Y., Wu, J., Chen, F., Wang, Z., Li, J., & Wang, L. (2024). Empirical Assessment of AI-Powered Tools for Vocabulary Acquisition in EFL Instruction. *IEEE Access*, *12*, 131892–131905. https://doi.org/10.1109/ACCESS.2024.3446657

Wei, L. (2023). Artificial intelligence in language instruction: Impact on English learning achievement, L2 motivation, and self-regulated learning. *Frontiers in Psychology*, *14*, 1261955. https://doi.org/10.3389/fpsyg.2023.1261955

Wibawa, I. M. C., Widiana, I. W., & Jampel, N. (2024). How EtnoEduction is Essential and Linked to the Science Learning in the 21st Century Paradigm? *Jurnal Edutech Undiksha*, *12*(1), 11–19. https://doi.org/10.23887/jeu.v12i1.82441

Wodaj, H. (2020). Effects of 7E Instructional Model with Metacognitive Scaffolding on Students' Conceptual Understanding in Biology. *Journal of Education in Science, Environment and Health*. https://doi.org/10.21891/jeseh.770794

Xin, K. K., & Nasri, N. B. M. (2024). The Readiness Level of Primary School Teachers Towards the Implementation of Classroom-Based Assessment

(PBD). *International Journal of Academic Research in Progressive Education and Development*, *13*(4), Pages 1788-1799. https://doi.org/10.6007/IJARPED/v13-i4/23224

Yu, J. (2023). Exam Culture and Formative Assessment in China: The Gaokao Reform and Its Sociocultural Hindrance. *Journal of Education, Humanities and Social Sciences*, *23*, 291–301. https://doi.org/10.54097/ehss.v23i.12900

Yu, N., Zhang, J., Mitra, S., Smith, R., & Rich, A. (2025). *AI-Educational Development Loop (AI-EDL): A Conceptual Framework to Bridge AI Capabilities with Classical Educational Theories*. https://doi.org/10.48550/ARXIV.2508.00970

Zhai, X., Neumann, K., & Krajcik, J. (2023). Editorial: AI for tackling STEM education challenges. *Frontiers in Education*, *8*, 1183030. https://doi.org/10.3389/feduc.2023.1183030