Prisma Sains: Jurnal Pengkajian Ilmu dan Pembelajaran Matematika dan IPA IKIP Mataram
https://e-journal.undikma.ac.id/index.php/prismasains/index
e-mail: prismasains.pkpsm@gmail.com

July 2026. Vol. 14, No. 3
p-ISSN: 2338-4530
e-ISSN: 2540-7899
pp. 1851-1863

Evaluation of Science Learning through Artificial Intelligence-Based Formative Feedback among Primary School Teacher Education Students

[1]*Muh. Nasir, [2]Juryatina, [3]Mei Indra Jayanti, [4]M. Ekahidayatullah

[1,3,4] Biology Education Department, Universitas Nggusuwaru, Kota Bima, Indonesia
[2] Primary School Teacher Education Department, Universitas Nggusuwaru, Kota Bima, Indonesia

*Corresponding Author e-mail: nasir.unswa25@gmail.com

Received: April 2026; Revised: May 2026; Published: July 2026

Abstract

This study examines the use of Artificial Intelligence (AI)-based formative feedback in the evaluation of science (IPA) learning among Primary School Teacher Education (PGSD) students. It addresses the limitations of conventional evaluation, particularly delayed feedback and the limited opportunities for conceptual correction in science learning. The study employed a mixed-methods approach with an explanatory sequential design; the quantitative component used a one-group pretest–posttest pre-experimental design without a control group, involving 40 students across two classes. The intervention used a web-based AI-driven digital evaluation application that generated generative and personalized feedback. Quantitative data were collected through pretest–posttest scores and a questionnaire, while qualitative data were obtained through interviews and system log analysis. The results show a statistically significant increase in learning-outcome scores after implementation, with a moderate improvement (6–7 points) and a medium effect size (Cohen's d = 0.65). The quality of AI-based feedback was categorized as good (mean = 3.13), particularly in terms of speed and usefulness, with a response time of less than five seconds. Qualitative findings indicate that immediate feedback supported conceptual understanding, helped identify misconceptions, and encouraged reflective learning behavior; however, limitations were found in feedback clarity and network stability. Because of the absence of a control group, this improvement cannot be fully attributed to AI feedback. The findings are preliminary and context-bound, yet indicate a positive contribution of AI-based formative feedback as an adaptive evaluation tool that needs to be tested with stronger designs.

Keywords: Artificial intelligence; Formative feedback; Science; Learning evaluation; Primary school teacher education

How to Cite: Nasir, M., Juryatina, J., Jayanti, M. I., & Ekahidayatullah, M. (2026). Evaluation of Science Learning through Artificial Intelligence-Based Formative Feedback among Primary School Teacher Education Students. *Prisma Sains: Jurnal Pengkajian Ilmu Dan Pembelajaran Matematika Dan IPA IKIP Mataram*, *14*(3), 1851–1863. https://doi.org/10.33394/j-ps.v14i3.20462

https://doi.org/10.33394/j-ps.v14i3.20462

INTRODUCTION

The evaluation of science (IPA) learning is a key component in ensuring the attainment of learning objectives while also serving as a basis for the continuous improvement of the learning process. In higher education, particularly in Primary School Teacher Education (PGSD) programs, the evaluation of science learning not only functions to measure learning outcomes but also fosters students' reflective ability as prospective teachers in developing a deeper understanding of scientific concepts (Mahardika & Zainuddin, 2022; Fleckenstein et al., 2023). However, evaluation practices that are still dominated by conventional approaches often face a number of challenges, including delayed feedback, a lack of personalization, and limited opportunities for students to correct their mistakes directly, especially in science material that is conceptual in nature and prone to misconceptions (Nazaretsky et al., 2026; Mientus et al., 2026). This condition causes evaluation to function more as a measurement tool than as an integral part of the learning process.

The development of digital technology, particularly Artificial Intelligence (AI), has begun to shift the paradigm of learning evaluation toward more adaptive and responsive approaches (Verawati et al., 2025; Corbett & Tangen, 2026). A number of studies show that AI-based systems are able to provide fast, consistent, and data-driven feedback, thus having the potential to improve the quality of formative evaluation (Nazaretsky et al., 2026; du Plooy et al., 2024). Other studies also indicate that AI-based automated feedback can accelerate the learning cycle and increase student engagement in understanding scientific concepts (Wang et al., 2025; Geschwind et al., 2026). In addition, previous research has tended to emphasize the development of AI-based systems or media, focusing on validity, practicality, and product effectiveness, without deeply examining how such feedback works in real learning contexts (Bashraheel & Ghinea, 2026; Steiss et al., 2024).

This study is grounded in four interrelated theoretical pillars. First, the theory of formative feedback asserts that effective feedback must clarify the learning goals, the learner's current position, and the next steps needed to close the performance gap (Van der Kleij et al., 2015; Wisniewski et al., 2020). Second, feedback literacy positions students as active agents who make sense of and act upon feedback (Carless & Boud, 2018; Winstone et al., 2017). Third, the theory of self-regulated learning (SRL) explains that feedback supports the cycle of planning, monitoring, and reflection that drives iterative learning efforts (Panadero, 2017; Panadero et al., 2018). Fourth, the conceptual change perspective shows that changes in scientific understanding occur gradually and require the explicit confrontation of initial misconceptions, for example regarding the concept of the human respiratory system (Kusnadi et al., 2019). These four pillars form the basis for interpreting not only whether scores improve, but how AI feedback helps students recognize errors and revise their conceptual understanding.

Nevertheless, the effectiveness of feedback is determined not only by its speed but also by the quality of its content, its clarity, and its relevance to students' learning needs, particularly in helping them overcome misconceptions in science learning (Wang et al., 2025). This indicates that there is still a gap between the development of AI-based evaluation technology and the empirical understanding of its impact on the quality of the learning process and on student learning outcomes. More specifically, the existing evidence is still dominated by writing/language contexts and product development, while studies examining AI-based formative feedback in the evaluation of science learning among prospective primary school teachers (PGSD) remain very limited, and few integrate system log data to explain the link between technical performance and the learning experience.

Based on this gap, the novelty of this study lies in four aspects: (1) the population of PGSD students as prospective primary school teachers; (2) a focus on the empirical analysis of the implementation of AI-based formative feedback in science learning, rather than merely the development of a system; (3) a three-sided evaluative framework that integrates learning outcomes, feedback perception/quality, and learning experience; and (4) the integration of system log data to separate technical responsiveness from the pedagogical quality of feedback.

Accordingly, the research questions of this study are: (1) how is the quality of AI-based formative feedback in science learning; (2) how do students' learning outcomes change after its implementation; and (3) how does it contribute to students' learning experience. The objective of the study is to evaluate science (IPA) learning through the use of AI-based formative feedback among PGSD students by analyzing the quality of feedback, examining changes in learning outcomes, and evaluating its contribution to supporting a more reflective and responsive learning process.

METHOD

Research Design

This study employed a mixed-methods approach with an explanatory sequential design to evaluate science (IPA) learning through the use of AI-based formative feedback among

PGSD students. Quantitative analysis was conducted first to identify patterns in science learning outcomes and feedback quality, followed by qualitative analysis to explore and explain the quantitative findings related to students' learning experiences (Gierus et al., 2025). The quantitative component applied a one-group pretest–posttest pre-experimental design without a control/comparison group. Consequently, changes in scores cannot be fully attributed to AI feedback, as they may be influenced by regular classroom instruction, repeated-testing effects, lecturer explanations, peer discussion, and familiarity with the test format. This internal-validity limitation is openly acknowledged and serves as the basis for the use of non-causal language in the reporting. The qualitative phase was connected to the quantitative phase through the purposive selection of informants based on variation in academic ability and achievement (gain), as well as the juxtaposition of test, questionnaire, log, and interview results to explain the patterns of findings.

Subjects and Setting

The study was conducted in the PGSD Program of Universitas Nggusuwaru during one science learning cycle, namely from March to April 2025. The population consisted of all PGSD students enrolled in a course related to science learning, with a sample of 40 students divided into two classes. The sampling technique was total sampling. The limited sample size is appropriate for an exploratory classroom trial, so generalization is avoided.

The intervention used a web-based AI-driven digital evaluation application developed specifically for the PGSD context. After students answered the formative questions, the system generated feedback automatically. The feedback was generative and personalized rather than a mere right/wrong statement comprising three components: strengths, areas for improvement, and learning recommendations tailored to each student's answers. The application has student accounts and lecturer accounts and is integrated with a database for reporting. Before implementation, the teaching lecturers received training on using the application and then supervised the conduct of the evaluation in class.

The feedback engine utilized a commercial Large Language Model accessed through an API (Gemini), using standardized prompts containing the answer key, the scoring rubric, and indicators of the respiratory-system concept so that the feedback was consistent across students. In this implementation, the lecturers did not review each AI feedback item one by one in real time; rather, they performed periodic verification (spot-checks) of approximately 20% of the feedback to ensure accuracy and pedagogical appropriateness, and provided additional clarification in class for frequently missed items.

Instruments

The research instruments consisted of several main components. First, a formative test comprising 10 multiple-choice questions based on science material (score range 0–100), developed according to indicators of science competence covering conceptual understanding, application of concepts, and identification of misconceptions. The material focused on the concept of the human respiratory system, chosen for its high conceptual complexity and strong potential for misconceptions among PGSD students. Second, a Likert-scale questionnaire (1–4) (12 items) to assess feedback quality based on the aspects of speed, clarity, completeness, and usefulness. Third, a semi-structured interview guide to explore students' experiences. Fourth, system log data, including duration of use, response time, error rate, and system availability (uptime) (Cavalcanti et al., 2021).

Before use, the test instrument was first piloted on 25 students outside the research sample (a parallel class in the same cohort) to analyze the difficulty index and discrimination power of each item. The results of the item analysis are presented in Table 1 and Figure 1. Most items fell within the medium difficulty category, with one easy item and one difficult item, and adequate to good discrimination power; all items met the criteria and were therefore deemed suitable for use.

Table 1. Difficulty Index and Discrimination Index of the Formative Test Items (Respiratory System Material)

Item	Difficulty Index (P)	Category	Discrimination Index (D)	Category	Decision
1	0.78	Easy	0.33	Adequate	Accepted
2	0.65	Medium	0.37	Adequate	Accepted
3	0.55	Medium	0.46	Good	Accepted
4	0.48	Medium	0.38	Adequate	Accepted
5	0.27	Difficult	0.40	Good	Accepted
6	0.60	Medium	0.44	Good	Accepted
7	0.52	Medium	0.35	Adequate	Accepted
8	0.45	Medium	0.39	Adequate	Accepted
9	0.68	Medium	0.32	Adequate	Accepted
10	0.58	Medium	0.43	Good	Accepted
Mean	0.56	Medium	0.39	Adequate	

Category reference: Difficulty index P < 0.30 = difficult; 0.30–0.70 = medium; > 0.70 = easy. Discrimination D ≥ 0.40 = good; 0.30–0.39 = adequate; 0.20–0.29 = marginal; < 0.20 = poor.

Figure 1. Difficulty Index (P) and Discrimination Index (D) per Item

Item 1 *(indicator: conceptual understanding; level C2)*. Where does the exchange of oxygen and carbon dioxide in the human respiratory system take place? (A) Trachea; (B) Bronchus; (C) Alveolus *(answer key)*; (D) Larynx; (E) Pharynx.

Item 2 *(indicator: application of concepts/identification of misconceptions; level C3)*. When a person exercises, their breathing rate increases. Which is the most accurate explanation of this phenomenon? (A) The body requires more oxygen and must expel more carbon dioxide *(answer key)*; (B) The lungs shrink, making breathing faster; (C) The oxygen level in the air decreases during exercise; (D) The heart stops pumping blood to the lungs; (E) The alveoli close to store reserve oxygen.

Ethical Procedures

Participation was voluntary and preceded by an explanation of the research objectives; students' identities were anonymized and the log data were processed without personal identifiers. All participants signed an informed consent form before data collection and could withdraw at any time without academic consequences. Students' identities were anonymized using codes (e.g., M-01 to M-40), and the log data were stored in encrypted form and accessible only to the research team. This study obtained ethical approval from the Research Ethics Committee of Universitas Nggusuwaru under number 020/KI.Ak.I/I/2026.

Data Collection and Analysis

Data collection was carried out in three stages: (1) administration of the pretest; (2) implementation of the AI-based science learning evaluation over two weeks (four sessions), during which students worked on the formative questions and received automated feedback directly (Naeem et al., 2023); and (3) the posttest and distribution of the perception questionnaire, followed by interviews with 10 students selected purposively to represent the variation in academic ability.

Quantitative data were analyzed using descriptive statistics and a paired-samples t-test at a significance level of 0.05 (Schäfer & Schwarz, 2019), reporting the t value, degrees of freedom (df), exact p-value, and 95% confidence interval (CI) for the mean difference. Effect size was calculated using Cohen's d for paired designs with the formula $d_z = t/\sqrt{n}$. The questionnaire data were analyzed by calculating the mean scores based on the Likert-scale intervals: 1.00–1.75 (poor), 1.76–2.50 (fair), 2.51–3.25 (good), and 3.26–4.00 (very good). Prior to the inferential analysis, the data were tested for normality using the Shapiro–Wilk test and showed a normal distribution ($p > 0.05$) (Nurjanah et al., 2023).

Qualitative data were analyzed using thematic analysis through the stages of data reduction, categorization, and conclusion drawing, with a search for negative cases. Coding was performed by two independent coders with an agreement level of Cohen's Kappa = 0.82 (very good category); coding discrepancies were resolved through discussion until consensus. Data trustworthiness was maintained through method and source triangulation by comparing the results of the test, questionnaire, interviews, and system log data (Braun & Clarke, 2019). The validity of the questionnaire instrument was tested using content validity through expert judgment with an Aiken's V index (0.82–0.91), while its reliability was tested using Cronbach's Alpha (0.87). The reliability of the science learning achievement test was tested using KR-20 (0.79), a high reliability category (Kusnadi et al., 2019)

RESULTS AND DISCUSSION

Students' Learning Outcomes in Science Learning (Pretest–Posttest)

Students' learning outcomes in science, particularly on the topic of the human respiratory system and the mechanism of gas exchange, were analyzed by comparing the pretest and posttest scores of the two classes. The test consisted of 10 multiple-choice questions (scale 0–100), designed to measure conceptual understanding, application, and simple analysis. A summary is presented in Table 2.

Table 2. Pretest and Posttest Results of PGSD Students in Science Learning (Scale 0–100)

Class	n	Mean Pretest (SD)	Mean Posttest (SD)	Difference	t	df	p-value	95% CI of difference*
Class A	20	67.50 (6.80)	74.50 (6.10)	7.00	3.05	19	0.006	[2.2; 11.8]
Class B	20	65.00 (7.20)	71.00 (6.50)	6.00	2.72	19	0.011	[1.4; 10.6]
Mean	40	66.25 (7.00)	72.75 (6.30)	6.50	4.10	39	0.0002	[3.3; 9.7]

The 95% CI was computed from the reported t value and mean difference (SE = difference/t).

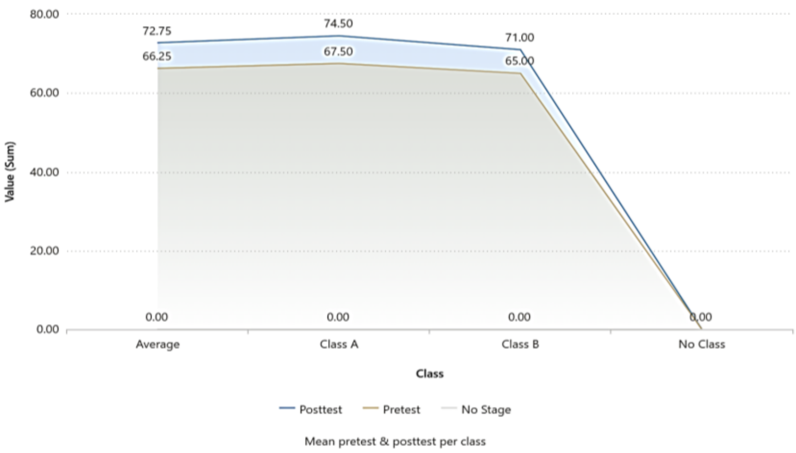

Figure 2. Comparison of Mean Pretest and Posttest Scores

The mean pretest scores in both classes were in the medium category, indicating that initial understanding of the human respiratory system concept was not yet optimal. This indicates initial misconceptions, particularly regarding the process of gas diffusion in the alveoli and the relationship between breathing rate and the body's oxygen needs. After implementation, the mean posttest scores increased in both classes, with an overall increase of 6.50 points. The paired-samples t-test showed that this increase was statistically significant ($p < 0.05$). However, because of the design without a control group, this improvement is interpreted as a change observed after implementation, not as evidence of a single causal effect of AI feedback (Van der Kleij et al., 2015).

Scientifically, this improvement is consistent with the role of formative feedback in enabling direct cognitive correction of misconceptions. In science learning, conceptual errors cannot be addressed merely by telling students whether they are right or wrong; the underlying process must be explained. The feedback in this study targeted the core of the error, for example in the process of gas exchange in the alveoli, so that students could revise their knowledge structures more quickly (Anand et al., 2026).

Nevertheless, the improvement was moderate (6–7 points). This shows that science learning is not an instant process; the characteristics of science material, which demand conceptual understanding and cause-and-effect relationships, mean that cognitive change occurs gradually through repeated practice and reinforcement (Wisniewski et al., 2020). The small decrease in standard deviation from pretest to posttest indicates that the variation in ability became more homogeneous, suggesting that feedback also helped to level out understanding, particularly for students with lower initial ability (Georgara et al., 2025). To strengthen the interpretation, the effect size was calculated using Cohen's d (formula $d_z = t/\sqrt{n}$), as presented in Table 3.

Table 3. Effect Size (Cohen's d) of Students' Learning Outcomes in Science Learning

Class	**Mean Pretest**	**Mean Posttest**	**Difference**	**Cohen's d (d_z)**	**Category**
Class A	67.50	74.50	7.00	0.68	Medium
Class B	65.00	71.00	6.00	0.61	Medium
Mean	66.25	72.75	6.50	0.65	Medium

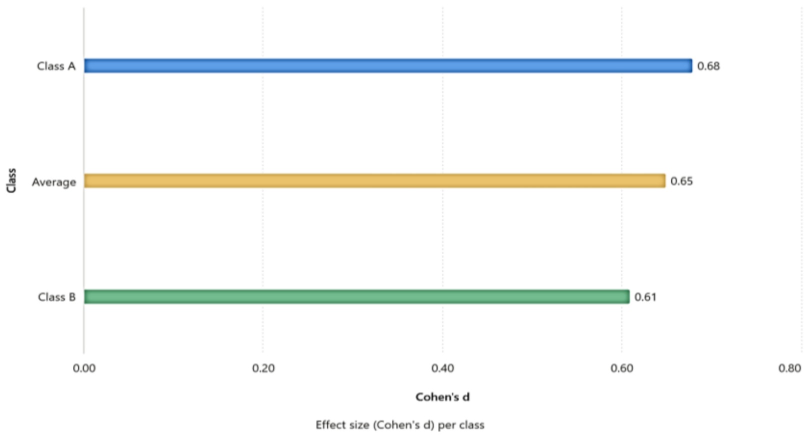

Figure 3. Effect Size (Cohen's d) per Class

The effect-size values in both classes were in the medium category, indicating a practically meaningful but not large impact. This medium category indicates that the intervention functioned as an augmentative factor rather than a single dominant factor, consistent with the absence of a control group. In science learning, conceptual understanding is built through the integration of explanation, learning experience, and reflection; AI-based feedback accelerates this reflective process but still depends on students' active engagement (Van der Kleij et al., 2015).

Quality of AI-Based Formative Feedback in Science Learning

The quality of feedback was analyzed based on four aspects: speed, clarity, completeness, and usefulness. Data were obtained through a Likert questionnaire (1–4) completed by 40 students and supported by system log data. A summary is presented in Table 4.

Table 4. Quality of AI-Based Formative Feedback in Science Learning

Aspect	Mean Time (seconds)	Perception Score (1–4)	Category
Speed	4.8	3.20	Good
Completeness	5.0	3.15	Good
Clarity	5.2	3.05	Good
Usefulness	5.1	3.10	Good
Mean	5.0	3.13	Good

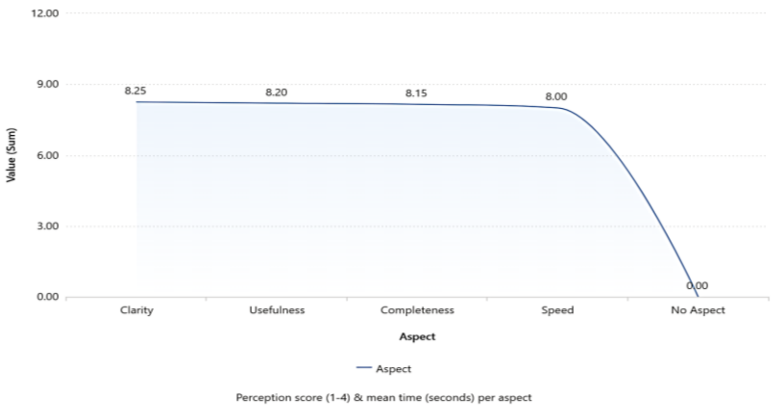

Figure 4. Feedback Quality Perception Score per Aspect

The feedback demonstrated a high response speed (average < 5 seconds), allowing cognitive reflection to occur immediately after the thinking process. In science learning, this speed helps students immediately identify and correct misconceptions in dynamic concepts such as the mechanism of gas exchange (Winstone et al., 2017; Mizumoto & Eguchi, 2023). All aspects were in the "good" category (mean 3.13), with the highest being speed (3.20) and the lowest being clarity (3.05).

In response to the reviewers' request, the clarity aspect was analyzed in greater depth. The lowest score on clarity is consistent with the qualitative complaint that the feedback language was too general for complex questions, so that students had to re-read it. In science learning, which is replete with scientific terms and abstract processes, feedback that is insufficiently specific can hinder interpretation and even perpetuate misconceptions (Kasneci et al., 2023; Escalante et al., 2023). This pattern affirms an important distinction: technical speed does not automatically mean pedagogical clarity. The relatively high completeness and usefulness aspects show that the system does not merely assess right/wrong but provides explanations that function as scaffolding (Yan et al., 2024).

Students' Experiences and Responses to AI-Based Feedback

Students' experiences were analyzed through semi-structured interviews with 10 purposively selected students, using thematic analysis. A summary is presented in Table 5.

Table 5. Summary of Qualitative Findings on Students' Experiences

No	Main Theme	Finding Indicator	Representative Quote	Number of Informants (out of 10)	Tendency
1	Speed of feedback response	Feedback received immediately after completing the task	"The feedback appeared right away, so I immediately knew my mistake."	9 (90%)	Dominant

No	Main Theme	Finding Indicator	Representative Quote	Number of Informants (out of 10)	Tendency
2	Ease of understanding errors	Students could identify errors quickly	"I could tell which part was wrong without having to wait for the lecturer."	8 (80%)	Dominant
3	Support for conceptual understanding	Feedback helped connect concepts to answers	"The explanation helped, especially for questions that require deeper understanding."	6 (60%)	Fairly dominant
4	Clarity of feedback language	Language still too general for some complex questions	"Sometimes the explanation was still too general, so I had to read it again."	5 (50%)	Moderate
5	Technical constraints (network)	Delayed responses due to weak internet connection	"When the network was bad, the feedback took longer to appear."	3 (30%)	Minor
6	Change in learning behavior	Students attempted to redo questions after receiving feedback	"I usually tried again right after finding out my mistake."	6 (60%)	Fairly dominant

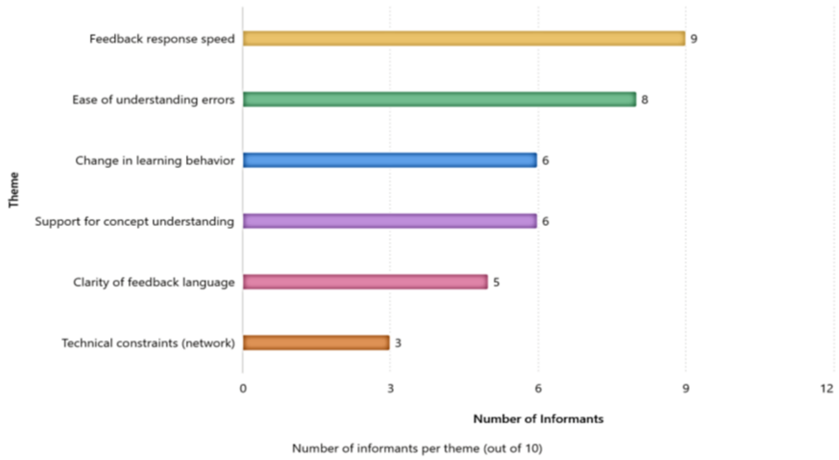

Figure 5. Number of Informants per Theme (out of 10)

Students' experiences were dominated by positive perceptions regarding the speed and ease of identifying errors, important indicators of formative evaluation. In science learning, the ability to recognize conceptual errors (e.g., in the mechanism of gas diffusion or the function of the respiratory organs) is the first step in developing accurate scientific understanding; AI feedback functions as a diagnostic tool that accelerates conceptual correction (Lim et al., 2023). Students also stated that the feedback helped connect their answers to the correct concepts, functioning as scaffolding (Aristovnik et al., 2020; Bond et al., 2020). Further analysis examined why the feedback helped some students but was less clear for others: students who were better prepared used it as a trigger for reflection and immediately revised their answers (supporting SRL; Panadero et al., 2018), while others were constrained by language that was too general (in line with the importance of feedback literacy; Carless & Boud, 2018; Panadero, 2017). As a negative case, 2 of the 10 informants stated that their scores did not change much despite perceiving the feedback positively; both reported difficulty understanding the feedback language for complex questions—consistent with the lowest clarity score. Delays when the network was weak were a minor factor but underscored the importance of infrastructure readiness. The change in learning behavior—reworking questions after receiving feedback—shows that AI feedback affects not only outcomes but also the learning process (Ryan et al., 2019; Braun & Clarke, 2019).

Usage Patterns of the AI-Based Evaluation System (Log Data)

Usage patterns were analyzed from the log data during the two weeks of implementation, including duration of use, response time, error rate, and uptime. A summary is presented in Table 6.

Table 6. Usage Patterns of the AI-Based Evaluation System

Week	Mean Session Duration (minutes)	Response Time (seconds)	Error Rate (%)	Uptime (%)
1	9.5	2.8	4.5	98.5
2	10.2	2.5	3.8	99.0
Mean	9.85	2.65	4.15	98.75

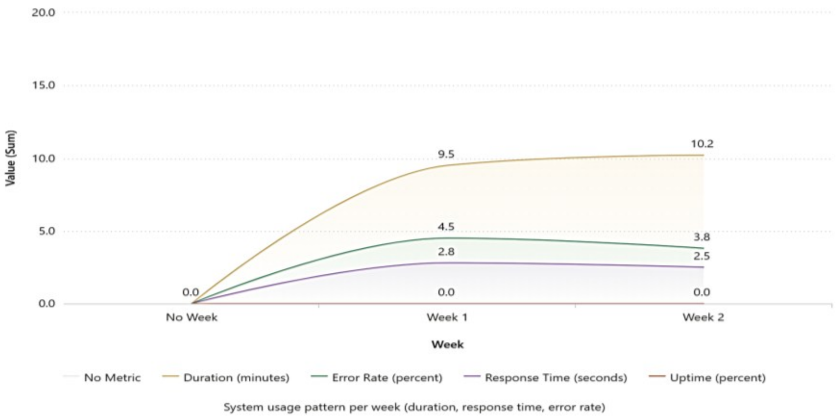

Figure 6. System Usage Patterns per Week

The average usage duration of 9–10 minutes per session reflects relatively good cognitive engagement, as students needed time to understand the conceptual explanations for process-based material such as the human respiratory system. The increase in duration in the second week indicates adaptation and a shift toward more reflective learning. A response time of under 3 seconds supports near-real-time formative feedback (Johar et al., 2023; Henrie et al., 2015). The decrease in error rate (4.5% → 3.8%) indicates improved stability, although the system was not yet entirely free of disruptions; an uptime above 98% indicates consistent access. Overall, good technical performance needs to be accompanied by user engagement in order to have an optimal impact on learning outcomes (Moturu & Nethi, 2023).

Integrative Discussion

The joint display of the data shows: log data = high technical responsiveness; questionnaire = good perceived usefulness; interviews = feedback-language clarity remains an obstacle for complex questions. This pattern affirms the distinction technical responsiveness ≠ pedagogical quality, which is the core novelty of this study. The findings support studies on the benefits of automated/AI feedback (Fleckenstein et al., 2023; Wang et al., 2025; Geschwind et al., 2026), extend the focus from the writing/language domain to the evaluation of science learning among prospective primary school teachers, and complement studies that highlight AI's speed by showing that speed is insufficient when conceptual clarity is weak (Bashraheel

& Ghinea, 2026; Steiss et al., 2024). The only moderate effect can be explained by the short duration, the limited number of items, the diversity of prior knowledge, the still-general feedback language, and network constraints—so that AI feedback should be positioned as a complement, not a replacement for the lecturer's role, accompanied by supervision and the alignment of content with concept indicators (Yan et al., 2024)

CONCLUSION

Following the three research questions: (1) Feedback quality was perceived as good (mean 3.13), highest in speed (3.20) and lowest in language clarity (3.05). (2) Learning outcomes increased moderately (66.25 → 72.75; difference of 6.50 points; Cohen's d = 0.65) and were statistically significant after implementation; because there was no control group, this improvement cannot be confirmed as solely the result of AI feedback. (3) The learning experience showed that students more easily identified misconceptions and were encouraged to correct them, while feedback-language clarity for complex questions and network stability remained obstacles.

Thus, the integration of AI-based formative feedback in the evaluation of science (IPA) learning functions not only as an assessment tool but also as a mechanism supporting adaptive and reflective learning. However, this evidence is preliminary and context-bound: a small sample ($n = 40$), a short duration (two weeks), a single science topic, and no control group. Therefore, the findings are not generalized.

RECOMMENDATION

Future research should focus on improving the pedagogical design of AI-based formative feedback, particularly the clarity, specificity, and contextual relevance of the feedback language in science learning. It is also recommended to use an experimental design with a control group and a larger sample, to extend the duration, to explore various science topics with differing levels of complexity, to develop adaptive feedback according to students' cognitive levels and misconceptions, and to improve system stability and reduce dependence on internet connectivity. Longitudinal research is needed to test the long-term impact on students' conceptual understanding and self-regulated learning.

ACKNOWLEDGMENTS

The authors wish to express their sincere gratitude to the students of the Primary School Teacher Education Study Program (PGSD) at Universitas Nggusuwaru for their valuable participation in this research. Special appreciation is also extended to our colleagues who provided constructive feedback throughout the research process. Furthermore, the authors would like to thank the Institute for Research and Community Services (LPPM) of Universitas Nggusuwaru for supporting this study.

FUNDING INFORMATION

This research received no external funding.

AUTHOR CONTRIBUTIONS STATEMENT

Name of Author	C	M	So	Va	Fo	I	R	D	O	E	Vi	Su	P	Fu
Muh. Nasir	✓	✓	✓	✓	✓	✓		✓	✓	✓		✓	✓	
Juryatina		✓		✓		✓				✓	✓	✓	✓	
Mei Indra Jayanti				✓		✓				✓			✓	
M. Ekahidayatullah	✓		✓	✓	✓	✓	✓			✓		✓	✓	

CONFLICT OF INTEREST STATEMENT

The authors have declared that there are no conflicts of interest regarding the publication of this paper. The authors have no financial, personal, or professional relationships that could have influenced the work reported in this study.

INFORMED CONSENT

We have obtained informed consent from all individuals included in this study.

DATA AVAILABILITY

The data that support the findings of this study are available from the corresponding author upon reasonable request.

REFERENCES

Anand, S., Holzmann, U., & Payumo, A. Y. (2026). AI-assisted qualitative analysis of formative feedback to support student-centered learning in physiology. *Advances in Physiology Education, 50*(1), 244–248. https://doi.org/10.1152/advan.00252.2024

Aristovnik, A., Keržič, D., Ravšelj, D., Tomaževič, N., & Umek, L. (2020). Impacts of the COVID-19 pandemic on the life of higher education students: A global perspective. *Sustainability, 12*(20), Article 8438. https://doi.org/10.3390/su12208438

Bashraheel, M., & Ghinea, G. (2026). How does leveraging artificial intelligence in assessments impact student outcomes? A systematic review. *Computer Science Review, 61*, Article 100929. https://doi.org/10.1016/j.cosrev.2026.100929

Bond, M., Buntins, K., Bedenlier, S., Zawacki-Richter, O., & Kerres, M. (2020). Mapping research in student engagement and educational technology in higher education: A systematic evidence map. *International Journal of Educational Technology in Higher Education, 17*(1), Article 2. https://doi.org/10.1186/s41239-019-0176-8

Braun, V., & Clarke, V. (2019). Reflecting on reflexive thematic analysis. *Qualitative Research in Sport, Exercise and Health, 11*(4), 589–597. https://doi.org/10.1080/2159676X.2019.1628806

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

Cavalcanti, A. P., Barbosa, A., Carvalho, R., Freitas, F., Tsai, Y.-S., Gašević, D., & Mello, R. F. (2021). Automatic feedback in online learning environments: A systematic literature review. *Computers and Education: Artificial Intelligence, 2*, Article 100027. https://doi.org/10.1016/j.caeai.2021.100027

Corbett, B. J., & Tangen, J. M. (2026). AI tutors vs. tenacious myths: Evidence from personalised dialogue interventions in education. *Computers in Human Behavior, 175*, Article 108828. https://doi.org/10.1016/j.chb.2025.108828

du Plooy, E., Casteleijn, D., & Franzsen, D. (2024). Personalized adaptive learning in higher education: A scoping review of key characteristics and impact on academic performance and engagement. *Heliyon, 10*(21), Article e39630. https://doi.org/10.1016/j.heliyon.2024.e39630

Escalante, J., Pack, A., & Barrett, A. (2023). AI-generated feedback on writing: Insights into efficacy and ENL student preference. *International Journal of Educational Technology in Higher Education, 20*(1), Article 57. https://doi.org/10.1186/s41239-023-00425-2

Fleckenstein, J., Liebenow, L. W., & Meyer, J. (2023). Automated feedback and writing: A multilevel meta-analysis of effects on students' performance. *Frontiers in Artificial Intelligence, 6*, Article 1162454. https://doi.org/10.3389/frai.2023.1162454

Georgara, A., Santolini, M., Kokshagina, O., Haux, C. J. J., Jacobs, D., Biwott, G., Correa, M., Sierra, C., Fernandez-Marquez, J. L., & Rodriguez-Aguilar, J. A. (2025). Optimising team dynamics: The role of AI in enhancing challenge-based learning participation experience and outcomes. *Computers and Education: Artificial Intelligence, 8*, Article 100388. https://doi.org/10.1016/j.caeai.2025.100388

Geschwind, S., Graf, J., Voss, D., & Hackl, V. (2026). GPT-4 feedback increases student activation and learning outcomes in higher education. *International Journal of Artificial Intelligence in Education, 36*(1–2), Article 100014. https://doi.org/10.1016/j.ijaied.2026.100014

Gierus, B., Du, T., Maduforo, A. N., Gilbert, B., & Koh, K. (2025). Prevalence and quality of mixed methods research in educational subdisciplines: A systematic review. *SAGE Open, 15*(2), 1–20. https://doi.org/10.1177/21582440251335171

Henrie, C. R., Halverson, L. R., & Graham, C. R. (2015). Measuring student engagement in technology-mediated learning: A review. *Computers & Education, 90*, 36–53.

https://doi.org/10.1016/j.compedu.2015.09.005

Johar, N. A., Kew, S. N., Tasir, Z., & Koh, E. (2023). Learning analytics on student engagement to enhance students' learning performance: A systematic review. *Sustainability, 15*(10), Article 7849. https://doi.org/10.3390/su15107849

Kasneci, E., Sessler, K., Küchemann, S., Bannert, M., Dementieva, D., Fischer, F., Gasser, U., Groh, G., Günnemann, S., Hüllermeier, E., Krusche, S., Kutyniok, G., Michaeli, T., Nerdel, C., Pfeffer, J., Poquet, O., Sailer, M., Schmidt, A., Seidel, T., . . . Kasneci, G. (2023). ChatGPT for good? On opportunities and challenges of large language models for education. *Learning and Individual Differences, 103*, Article 102274. https://doi.org/10.1016/j.lindif.2023.102274

Kusnadi, K., Lazuardi, Z., & Surakusumah, W. (2019). The conceptual change of the human respiratory system through POE-based learning. *Journal of Physics: Conference Series, 1280*(3), Article 032009. https://doi.org/10.1088/1742-6596/1280/3/032009

Lim, L., Bannert, M., van der Graaf, J., Singh, S., Fan, Y., Surendrannair, S., Rakovic, M., Molenaar, I., Moore, J., & Gašević, D. (2023). Effects of real-time analytics-based personalized scaffolds on students' self-regulated learning. *Computers in Human Behavior, 139*, Article 107547. https://doi.org/10.1016/j.chb.2022.107547

Mahardika, N. I., & Zainuddin, A. (2022). Application of formative assessment to analyze students' problem-solving skills. *Prisma Sains: Jurnal Pengkajian Ilmu dan Pembelajaran Matematika dan IPA IKIP Mataram, 10*(2), 226. https://doi.org/10.33394/j-ps.v10i2.4947

Mientus, L., Wulff, P., Nowak, A., & Borowski, A. (2026). When AI feedback was still in its infancy: An exploratory comparison of early AI feedback attempts on preservice physics teachers' reflective writing. *Education Sciences, 16*(2), Article 301. https://doi.org/10.3390/educsci16020301

Mizumoto, A., & Eguchi, M. (2023). Exploring the potential of using an AI language model for automated essay scoring. *Research Methods in Applied Linguistics, 2*(2), Article 100050. https://doi.org/10.1016/j.rmal.2023.100050

Moturu, V. R., & Nethi, S. D. (2023). Artificial intelligence in education. In *Lecture notes in networks and systems* (Vol. 478, pp. 233–244). https://doi.org/10.1007/978-981-19-2940-3_16

Naeem, M., Ozuem, W., Howell, K., & Ranfagni, S. (2023). A step-by-step process of thematic analysis to develop a conceptual model in qualitative research. *International Journal of Qualitative Methods, 22*, 1–18. https://doi.org/10.1177/16094069231205789

Nazaretsky, T., Gabbay, H., & Käser, T. (2026). Can students judge like experts? A large-scale study on the pedagogical quality of AI and human personalized formative feedback. *Computers and Education: Artificial Intelligence, 10*, Article 100533. https://doi.org/10.1016/j.caeai.2025.100533

Nurjanah, S., Istiyono, E., Widihastuti, W., Iqbal, M., & Kamal, S. (2023). The application of Aiken's V method for evaluating the content validity of instruments that measure the implementation of formative assessments. *Journal of Research and Educational Research Evaluation, 12*(2), 2023–2125. http://journal.unnes.ac.id/sju/index.php/jere

Panadero, E. (2017). A review of self-regulated learning: Six models and four directions for research. *Frontiers in Psychology, 8*, Article 422. https://doi.org/10.3389/fpsyg.2017.00422

Panadero, E., Andrade, H., & Brookhart, S. (2018). Fusing self-regulated learning and formative assessment: A roadmap of where we are, how we got here, and where we are going. *The Australian Educational Researcher, 45*(1), 13–31. https://doi.org/10.1007/s13384-018-0258-y

Ryan, T., Henderson, M., & Phillips, M. (2019). Feedback modes matter: Comparing student perceptions of digital and nondigital feedback modes in higher education. *British*

Journal of Educational Technology, 50(3), 1507–1523. https://doi.org/10.1111/bjet.12749

Schäfer, T., & Schwarz, M. A. (2019). The meaningfulness of effect sizes in psychological research: Differences between subdisciplines and the impact of potential biases. *Frontiers in Psychology, 10*, Article 813. https://doi.org/10.3389/fpsyg.2019.00813

Steiss, J., Tate, T., Graham, S., Cruz, J., Hebert, M., Wang, J., Moon, Y., Tseng, W., Warschauer, M., & Olson, C. B. (2024). Comparing the quality of human and ChatGPT feedback on students' writing. *Learning and Instruction, 91*, Article 101894. https://doi.org/10.1016/j.learninstruc.2024.101894

Van der Kleij, F. M., Feskens, R. C. W., & Eggen, T. J. H. M. (2015). Effects of feedback in a computer-based learning environment on students' learning outcomes: A meta-analysis. *Review of Educational Research, 85*(4), 475–511. https://doi.org/10.3102/0034654314564881

Verawati, N. N. S. P., Wahyudi, W., Nisrina, N., & Asy'ari, M. (2025). ChatGPT in physics education: A content-based analysis of Newtonian force problems. *Prisma Sains: Jurnal Pengkajian Ilmu dan Pembelajaran Matematika dan IPA IKIP Mataram, 13*(2), 335. https://doi.org/10.33394/j-ps.v13i2.15824

Wang, W., Wang, Y., Chen, J., Wang, X., Zhang, H., Guo, C., & Peng, Y. (2025). The effectiveness of AI-supported personalized feedback on students' learning outcomes and motivation: A meta-analysis. *Journal of Educational Computing Research*. Advance online publication. https://doi.org/10.1177/07356331251410020

Winstone, N. E., Nash, R. A., Parker, M., & Rowntree, J. (2017). Supporting learners' agentic engagement with feedback: A systematic review and a taxonomy of recipience processes. *Educational Psychologist, 52*(1), 17–37. https://doi.org/10.1080/00461520.2016.1207538

Wisniewski, B., Zierer, K., & Hattie, J. (2020). The power of feedback revisited: A meta-analysis of educational feedback research. *Frontiers in Psychology, 10*, Article 3087. https://doi.org/10.3389/fpsyg.2019.03087

Yan, L., Sha, L., Zhao, L., Li, Y., Martinez-Maldonado, R., Chen, G., Li, X., Jin, Y., & Gašević, D. (2024). Practical and ethical challenges of large language models in education: A systematic scoping review. *British Journal of Educational Technology, 55*(1), 90–112. https://doi.org/10.1111/bjet.13370