International Journal of Evaluation and Research in Education (IJERE)
Vol. 15, No. 4, August 2026, pp. 2741~2749
ISSN: 2252-8822, DOI: 10.11591/ijere.v15i4.38648

Improving students' scientific argumentation through AI-supported feedback

Joelash R. Honra[1,2], John Lorence A. Villamin[1]
[1]General Education Department, School of Arts and Sciences, National University, Manila, Philippines
[2]Division of Curriculum and Instruction, College of Education, University of the Philippines Diliman, Quezon, Philippines

Article Info

Article history:

Received Jan 11, 2026
Revised May 14, 2026
Accepted Jun 12, 2026

Keywords:

Artificial intelligence
Formative feedback
Mixed-methods research
Scientific argumentation
Toulmin's argument pattern

ABSTRACT

Scientific argumentation is a core practice in science education, yet many students struggle to construct arguments that effectively integrate claims, evidence, and reasoning. With the growing use of artificial intelligence (AI) in education, AI-supported feedback has emerged as a potential tool to scaffold students' argumentation processes. This study examined the effects of AI-supported feedback on students' scientific argumentation using a quasi-experimental, explanatory sequential mixed-methods design. Two intact groups participated: an experimental group receiving AI-supported formative feedback on written arguments and a control group receiving conventional teacher feedback. Quantitative data were collected using a validated rubric based on the claim–evidence–reasoning (CER) framework and Toulmin's argument pattern (TAP), while qualitative data from student interviews and written responses provided contextual insights. Results showed that the experimental group achieved greater improvements in overall argumentation quality, particularly in evidence use and reasoning. Qualitative findings further indicated that AI feedback supported iterative revision and strengthened students' understanding of evidence–claim relationships.

Corresponding Author:

Joelash R. Honra
General Education Department, School of Arts and Sciences, National University
Coral Way, Mall of Asia (MOA) Complex, Pasay City, Metro Manila, 1300 Philippines
Email: jrhonra@nu-moa.edu.ph

1. INTRODUCTION

Scientific argumentation is widely recognized as a central practice in science education, essential for developing students' scientific literacy, critical thinking, and epistemic understanding of science [1]–[3]. Engaging in argumentation enables students to construct explanations, justify claims with empirical evidence, and evaluate competing explanations. International science education standards emphasize argumentation as a core scientific practice; however, empirical studies consistently report that students' arguments often lack coherent reasoning, sufficient evidence, or explicit connections between claims and supporting data [4], [5]. These persistent difficulties underscore the need for instructional approaches that effectively scaffold students' capacity to construct and refine evidence-based arguments.

A major challenge in teaching scientific argumentation lies in providing timely, specific, and formative feedback. High-quality feedback requires careful analysis of students' claims, evidence, and reasoning—processes that are cognitively demanding and time-intensive for instructors, particularly in large classes [6], [7]. Consequently, feedback is often limited to surface-level comments or summative evaluations that offer limited guidance for improvement. From a formative assessment perspective, effective feedback should support iterative revision and enable learners to bridge the gap between their current understanding and desired performance.

Recent advances in artificial intelligence (AI) present new opportunities to address these challenges. AI-mediated feedback systems can analyze student writing, identify structural components of arguments, and generate immediate, targeted guidance aligned with established frameworks such as Toulmin's argument pattern (TAP) and the claim–evidence–reasoning (CER) model [8], [9]. Empirical evidence from studies in the United States and Europe indicates that hybrid human–AI feedback approaches can enhance students' revision practices, promote deeper reasoning, and improve the actionability of feedback. At the same time, these studies highlight the importance of pedagogical scaffolding and ethical safeguards to ensure responsible implementation [5]. While AI tools offer scalability, consistency, and adaptivity, their outputs may occasionally include inaccuracies, necessitating human oversight and explicit instructional guidance to ensure reliability and pedagogical soundness.

Despite growing interest in AI-supported learning, empirical evidence on its role in supporting scientific argumentation—particularly in secondary and postsecondary contexts—remains limited. Furthermore, relatively few studies adopt mixed-methods approaches that examine both learning outcomes and students' lived experiences with AI-generated feedback. A comprehensive understanding of the effectiveness, mechanisms, and limitations of AI-supported feedback is therefore essential for its responsible integration into science education [10].

This study advances the theoretical landscape of science education by explicitly integrating and extending three complementary frameworks: TAP, the CER model, and formative assessment theory [11]. Rather than treating these frameworks as discrete, the study conceptualizes AI-mediated feedback as a mediational scaffold that operationalizes formative assessment principles within argumentation-based instruction. In this integrated model, AI functions not merely as a feedback delivery tool but as an adaptive intermediary that supports the iterative construction, evaluation, and refinement of scientific arguments [12].

Specifically, the study refines existing theory by proposing that AI-generated feedback can enhance the alignment between students' argumentative structures (TAP/CER) and formative feedback processes (e.g., feedback loops, self-regulation, and revision cycles). It further extends formative assessment theory by introducing AI as a scalable mechanism for delivering contingent, component-specific feedback (i.e., claims, evidence, and reasoning), thereby addressing long-standing constraints related to feedback timeliness and granularity. At the same time, the study critically engages with the limitations of AI, incorporating the notion of human–AI co-regulation to account for the need for teacher oversight in mitigating potential inaccuracies. In doing so, the study contributes to the emerging discourse on human–AI collaboration in education by positioning AI as a complementary, rather than substitutive, agent in supporting students' development of evidence-based reasoning [13].

Therefore, this study aimed to examine the effects of AI-supported feedback on students' scientific argumentation using a quasi-experimental, explanatory sequential mixed-methods design. Specifically, it compared students receiving AI-supported feedback with those receiving traditional teacher feedback and explored how students experienced and utilized AI-generated guidance during the argumentation process. The findings are situated within broader discussions on actionable, ethical, and responsible AI use in education, with explicit consideration of the implications of potential AI-generated inaccuracies. Thus, the research questions of the study are:

- Is there a significant difference in students' scientific argumentation performance between those who receive AI-supported feedback and those who receive conventional teacher feedback?
- Which components of scientific argumentation (CER) are most influenced by AI-supported feedback?
- How do students perceive and experience AI-supported feedback in developing their scientific arguments?

2. METHOD

2.1. Research design

This study employed a quasi-experimental, explanatory sequential mixed-methods design [6]. Quantitative data were collected first to examine differences in students' scientific argumentation associated with AI-supported feedback, followed by qualitative data to contextualize and explain the observed patterns. A two-group pretest–posttest design was implemented using two intact classes assigned via convenience sampling to experimental and control conditions, as random assignment was not feasible in the school context.

Both groups received identical instructional content, argumentation tasks, and instructional time and were taught by the same teacher to minimize instructional variability. The primary distinction between groups was the source of formative feedback (AI-supported vs. teacher-provided). To reduce confounding due to feedback dosage, the number of feedback cycles, timing, and exposure were monitored and documented across conditions.

Because intact classes were used, causal inference is limited and findings are interpreted as associative estimates. To strengthen internal validity, pretest measures, statistical adjustment, and robustness checks were implemented. Potential threats—including selection bias, clustering by classroom, teacher expectancy effects, and implementation fidelity—are acknowledged. Given the clustered structure (students nested within classes), primary analyses were conducted using pretest-adjusted ANCOVA, supplemented by sensitivity analyses using: i) mixed-effects models with a random intercept for class and ii) cluster-robust standard errors. Intraclass correlation coefficients (ICC) for argumentation scores are reported to justify the unit of analysis, and permutation-based tests appropriate for two clusters were conducted to assess robustness.

2.2. Participants and context

Participants were 69 high school students enrolled in a science course where scientific argumentation formed part of the curriculum. Two intact classes were assigned to an experimental group (AI-supported feedback) and a control group (teacher-provided feedback). Baseline equivalence between groups was assessed using standardized effect sizes with 95% confidence intervals (CI). Posttest differences were estimated using analysis of covariance (ANCOVA) models adjusting for pretest scores. Given the modest sample size and clustered design, additional analyses included cluster-adjusted models and nonparametric approaches (e.g., rank-based ANCOVA). The ICC for the primary outcome was computed to quantify classroom-level dependence. While the sample reflects authentic classroom constraints, limitations include potential selection bias, modest statistical power, and clustering effects; these are addressed through transparent reporting and sensitivity analyses.

2.3. Instructional intervention

The six-week intervention consisted of inquiry-based science lessons integrating argumentation tasks. Students constructed written scientific arguments in response to phenomena-based questions and revised their work based on feedback.

2.3.1. Experimental group

Students received AI-supported formative feedback from a custom-built, rule-based natural language processing (NLP) system aligned with the CER framework. The system used deterministic rules (keyword matching, structural parsing, and template-driven logic) to generate targeted prompts. Rule development was calibrated against expert-coded responses (Cohen's κ>0.80). To enhance transparency, a descriptive evaluation of the feedback engine was conducted on held-out student responses, reporting precision and recall for CER prompts. Common failure modes included misclassification of implicit reasoning and over-triggering of evidence prompts in ambiguous cases. In such instances, teacher oversight and iterative refinement mitigated suboptimal feedback.

2.3.2. Control group

Students received conventional teacher feedback consisting of written comments and rubric-based scores aligned with CER criteria. Both groups engaged in comparable numbers of feedback cycles; however, latency differed, with AI feedback delivered immediately and teacher feedback provided after a delay. The number of cycles, average feedback time, and total exposure per student are reported to clarify potential dosage effects. TAP provided the theoretical basis for argument structure, while CER guided instruction and assessment [14]–[20].

2.4. Instrument

Scientific argumentation rubric (CER-based): students' written arguments were assessed using a CER-aligned rubric validated by three experts. Inter-rater reliability was acceptable (Fleiss' κ=0.82) [21], with component-level reliability also computed for CER scores. To enhance reproducibility, a detailed scoring guide with annotated exemplars for each level is provided. Raters were blinded to group assignment and assessment time (pre/post). Disagreements were resolved through consensus discussion, and periodic recalibration sessions were conducted to minimize rater drift, as shown in Table 1. Interview protocol: a semi-structured protocol explored students' perceptions of feedback, including clarity, usefulness, and impact on revision and learning processes [22]–[24].

Table 1. Scientific argumentation rubric

Component	Level 0	Level 1	Level 2	Level 3
Claim	No claim or incorrect claim	Vague or partially correct claim	Clear and correct claim	Clear, correct, and well-focused claim
Evidence	No evidence	Inappropriate or insufficient evidence	Relevant but limited evidence	Multiple, relevant, and sufficient evidence
Reasoning	No reasoning	Weak or incorrect reasoning	Partially connects evidence to claim	Clearly and logically connects evidence to claim using scientific principles

Note: the maximum possible score was 9 points.

2.5. Ethical considerations

This study adhered to established ethical standards for research involving human participants, particularly minors, and was conducted in accordance with internationally recognized guidelines, including the principles of the Declaration of Helsinki and institutional standards for educational research ethics. Ethical clearance was obtained through an equivalent review process, including evaluation by the dissertation panel and formal authorization from the Schools Division Office where the study was conducted, covering all procedures related to data collection, intervention implementation, and AI-supported processing of student responses [25].

Informed consent and assent were obtained following standardized ethical protocols for research with minors. Written informed consent was secured from school authorities and parents/guardians, while informed assent was obtained from student participants. All participants were explicitly informed about the study's purpose, procedures, the role of AI in generating feedback, the nature of the data collected, and their right to withdraw from the study at any point before data anonymization without penalty.

To ensure confidentiality and data protection, all data were de-identified using coded identifiers prior to analysis. The AI system processed only anonymized textual data and did not access personally identifiable information at any stage. Data were stored in secure, password-protected systems accessible only to the research team, and no identifiable data were shared with external parties [26]. Furthermore, the study was not preregistered; this is acknowledged as a limitation. To mitigate potential bias, all methodological and analytic decisions are reported transparently to support rigor and reproducibility.

2.6. Data collection and analysis procedures

All participants completed pretest and posttest assessments evaluated using a validated CER rubric. The study employed a quasi-experimental design with intact classes and no random assignment; thus, causal inferences are made cautiously, and statistical controls were applied to mitigate baseline differences. To complement the quantitative data, ten students from the experimental group, representing high, medium, and low performers, were purposively selected for semi-structured interviews [27].

Quantitative analyses included descriptive statistics and ANCOVA with pretest scores as covariates, with adjusted means, standard errors, and 95% CI reported. Model assumptions (normality, homogeneity of variance, and linearity) were examined using diagnostic plots and statistical tests. To strengthen robustness, multiple sensitivity analyses were conducted, including mixed-effects models to account for class-level clustering, cluster-robust standard errors, rank-based ANCOVA for nonparametric confirmation, and permutation tests. Full statistical outputs are reported to ensure transparency and reproducibility [28].

Given the absence of a priori power analysis, a post hoc sensitivity analysis was conducted to determine the minimum detectable effect sizes based on the sample size, design, and significance level, thereby contextualizing the statistical power and interpretation of findings. Moreover, to address the validity of the AI-mediated feedback system, a dedicated methodological component details the architecture and calibration of the NLP engine. The system was designed using rule-based and pattern-matching algorithms aligned with CER and TAP, enabling identification of CER structures. Calibration involved iterative refinement using expert-coded training data, with inter-rater agreement metrics used to benchmark system performance against human evaluators. These procedures were implemented to ensure the reliability and validity of automated feedback before deployment. Qualitative data were analyzed using thematic analysis, following systematic coding, categorization, and theme development. Findings were integrated with quantitative results through an explanatory sequential mixed-methods approach to elucidate patterns in students' argumentation development and their interactions with AI-generated feedback [29].

3. RESULTS AND DISCUSSION

3.1. Overall effects of AI-supported feedback on scientific argumentation

To address research question 1, overall scientific argumentation scores (maximum=9) were analyzed using a pretest–posttest design. Because intact, nonrandomized groups were used, posttest differences were examined using ANCOVA with pretest scores as a covariate, providing adjusted estimates of group effects, as seen in Table 2. Pretest differences between groups were small (d≈0.03), suggesting minimal baseline imbalance; however, equivalence was not formally tested. After adjusting for pretest scores, the ANCOVA revealed a statistically significant group effect favoring the experimental condition. The adjusted mean difference was 1.31 points (95% CI [0.42, 2.20]), corresponding to a large effect size (d=1.18). Meanwhile, a sensitivity analysis treating total rubric scores as ordinal (nonparametric comparison of gain distributions) yielded a consistent directional effect, supporting the robustness of the findings. Thus, these results suggest that AI-supported feedback was associated with greater improvements in overall scientific argumentation compared to conventional teacher feedback. However, due to the quasi-experimental design, conclusions should be interpreted as associative rather than strictly causal.

Table 2. Pretest and posttest scientific argumentation scores by group

Group	Pretest mean (SD)	Posttest mean (SD)	Mean gain	Adjusted posttest mean	Adjusted mean difference	95% CI	Effect size (d)
Experimental (AI-supported teacher feedback)	5.42 (1.21)	8.85 (1.34)	3.09	8.72	1.31	[0.42, 2.20]	1.18
Control (conventional teacher feedback)	5.38 (1.18)	7.21 (1.47)	1.83	7.41	-	-	-

Note: Max. score=9. Adjusted means from ANCOVA to control pretest scores.

3.2. Component-level effects of AI-supported feedback

To address research question 2, separate ANCOVA models were conducted for each CER component (claim, evidence, and reasoning), controlling for corresponding pretest scores; the results are presented in Table 3. After controlling for baseline performance, the experimental group outperformed the control group across all CER components, with the largest effects observed in reasoning (*d*=0.71) and evidence (*d*=0.65), and a moderate effect for claim development (*d*=0.42), with CI excluding zero. These results indicate that AI-supported feedback was associated with stronger posttest performance, particularly in components requiring deeper conceptual justification such as evidence and reasoning. The structured AI prompts that asked students to clarify claims, provide data, and explain logical connections likely supported these improvements, and sensitivity analyses using ordinal models showed comparable directional effects, confirming the robustness of the findings, as presented in Table 4. The qualitative examples illustrate that students in the experimental group demonstrated more substantive structural revisions, including clearer scientific claims, more specific evidence, and explicit causal reasoning. Improvements in the control group were comparatively modest and often limited to elaboration rather than conceptual strengthening.

Table 3. Adjusted posttest scores by CER component

Component	Experimental adjusted mean (SE)	Control adjusted mean (SE)	Adjusted mean difference	95% CI	Effect size (d)
Claim	2.74 (0.08)	2.51 (0.09)	0.23	[0.04, 0.42]	0.42
Evidence	2.81 (0.07)	2.45 (0.08)	0.36	[0.15, 0.57]	0.66
Reasoning	2.84 (0.07)	2.45 (0.09)	0.39	[0.18, 0.60]	0.71

Note: Max. per component=3. Adjusted means from ANCOVA to control each component pretest scores.

Table 4. Exemplar student responses

Group	Test	Student response
Experimental	Pre-test	Claim: "*Plants need water*"; Evidence: "*I watered my plant*"; Reasoning: "*Plants drink water*"
	Posttest	Claim: "*Plants require water to perform photosynthesis*"; Evidence: "*Watered plants grew taller and greener*"; Reasoning: "*Water supports photosynthesis, explaining growth differences*"
Control	Pre-test	Claim: "*Water helps plants*"; Evidence: "*I gave water.*"; Reasoning: "*Plants need it.*"
	Posttest	Claim: "*Water helps plants grow*"; Evidence: "*Plant grew taller*"; Reasoning: "*Because water helps growth*"

3.3. Qualitative insights into student learning processes

Qualitative findings provide explanatory depth to the quantitative results and address research question 3. Three major themes emerged from students' reflective journals and interviews: i) iterative revision and immediate guidance; ii) improved understanding of evidence–reasoning links; and iii) enhanced metacognitive awareness.

3.3.1. Theme 1: iterative revision and immediate guidance

Students consistently reported that AI-supported feedback enabled them to revise their arguments multiple times before submission. One student noted:

> "*Every time I revised, the feedback pointed out what was missing. It helped me improve step by step instead of guessing what the teacher wanted.*" (Student 31)

This theme helps explain the significantly higher gains observed in the experimental group. The opportunity for iterative revision aligns with formative assessment principles, allowing students to treat argument construction as a process rather than a one-time task.

3.3.2. Theme 2: improved understanding of evidence–reasoning links

Students also reported clearer understanding of how evidence must be explicitly connected to claims through reasoning. As one participant explained:

> "*Before, I just put data and thought it was enough. The AI kept asking me to explain why the data mattered, so I learned how to connect them properly.*" (Student 23)

This qualitative evidence supports the quantitative finding that evidence and reasoning components showed the largest improvement. The AI feedback functioned as a cognitive prompt that made implicit expectations explicit.

3.3.3. Theme 3: enhanced metacognitive awareness

A third theme highlights students' growing awareness of argument quality criteria. Several students indicated that they began to internalize the rubric:

> "*Eventually, I didn't need the feedback as much because I already knew to check my claim, evidence, and reasoning.*" (Student 5)

Table 5 presents a synthesis of quantitative and qualitative findings, illustrating how observed improvements in students' scientific argumentation correspond with reported learning processes and engagement with AI-supported feedback.

Table 5. Meta-inferences: integration of quantitative and qualitative findings

Quantitative results	Supporting qualitative evidence	Meta-inference
Experimental group showed greater overall gains in CER scores (Cohen's d=1.35)	Students reported iterative revision and stepwise guidance from AI feedback	Iterative, structured AI feedback likely facilitated substantial overall improvements in scientific argumentation
Largest posttest gains observed in Evidence and Reasoning components	Students described improved understanding of how evidence supports claims and explicit reasoning	AI prompts helped students make connections between evidence and claims, supporting higher-order argumentation skills
Control group showed smaller, surface-level improvements	Students reported minimal strategy use and limited engagement with reasoning	Conventional feedback may be insufficient for deep, reflective argumentation development
Enhanced metacognitive awareness noted in qualitative data	Students internalized the rubric and monitored their own revisions	AI-supported feedback scaffolds self-regulated learning and encourages reflection on argument quality

This study shows that AI-supported feedback enhances students' scientific argumentation and fosters metacognitive skills essential for self-regulated learning. Quantitative results indicated that only the experimental group receiving AI-generated feedback achieved significant gains in overall CER scores, with the largest improvements in evidence and reasoning. Qualitative data corroborated these findings, revealing that students used feedback iteratively, developed a clearer understanding of evidence–reasoning connections, and engaged in reflective revision, suggesting that structured AI prompts effectively guided learning and supported deeper argumentation skills.

Consistent with prior research on AI-assisted formative assessment [14], [16], [20], [24], the findings indicate that AI feedback provides immediate, specific, and systematic guidance that may be difficult to deliver consistently through traditional teacher feedback due to time constraints. Beyond replicating established benefits, this study extends the literature by demonstrating that AI feedback enhances both overall argumentation performance and the quality of individual components of scientific arguments. The structured feedback helped students refine claims, incorporate relevant evidence, and strengthen reasoning, suggesting that targeted prompts can support complex cognitive processes involved in constructing evidence-based explanations.

Importantly, the design of the AI-supported feedback system relied on deterministic rule-based prompts aligned with the CER framework, which shaped how students engaged cognitively with the revision process. Rather than generating open-ended or unpredictable suggestions, the system provided structured prompts that explicitly highlighted missing or weak argument components (e.g., insufficient evidence or unclear reasoning). From a cognitive perspective, these rule-based prompts may reduce extraneous cognitive load by directing students' attention to specific elements of their arguments while maintaining focus on the conceptual task. By making evaluation criteria explicit and segmenting complex argumentative tasks into manageable components, the feedback likely supported students' cognitive strategies during iterative revisions. Students' reflections suggest that these prompts functioned as scaffolds that guided them to reconsider how evidence supports claims and how reasoning explains this relationship, thereby facilitating metacognitive monitoring and strategic revision. In this way, the deterministic structure of the AI feedback may have supported both cognitive processing (argument construction) and metacognitive regulation (reflection and revision) during the learning process.

Several limitations must be acknowledged. The study involved a relatively small sample, a short intervention period, and the use of a single AI feedback system within one instructional context, which may constrain generalizability. Additionally, because the intervention relied on intact classes within a quasi-experimental design, causal interpretations should be made cautiously. Nonetheless, the findings have both theoretical and practical implications. Theoretically, they contribute to constructivist and self-regulated learning perspectives by illustrating how structured AI feedback can function as a scaffold that supports reflection, criterion internalization, and iterative knowledge construction. Practically, AI-supported feedback offers a scalable mechanism for providing individualized guidance without substantially increasing instructional workload. Future research should examine longer interventions, cross-disciplinary applications, and hybrid feedback models that integrate AI-generated guidance with teacher facilitation, as well as adaptive systems that dynamically respond to evolving student skill levels. Overall, the results highlight the potential of AI-supported feedback to enhance both scientific argumentation and metacognitive development, supporting learners in developing transferable reasoning skills.

4. CONCLUSION

AI-supported feedback enhances students' scientific argumentation, particularly improving the quality of claims, evidence, and reasoning, by providing timely, structured, and iterative guidance that promotes metacognitive reflection. While promising, these findings are limited by the study's design and should be interpreted cautiously. Therefore, AI feedback is best used as a complementary tool alongside teacher guidance. Future research should employ randomized and longitudinal designs to examine causal effects, long-term impacts, and the most effective strategies for integrating AI feedback with classroom instruction.

FUNDING INFORMATION

This research received no external funding.

AUTHOR CONTRIBUTIONS STATEMENT

This journal uses the Contributor Roles Taxonomy (CRediT) to recognize individual author contributions, reduce authorship disputes, and facilitate collaboration.

Name of Author	C	M	So	Va	Fo	I	R	D	O	E	Vi	Su	P	Fu
Joelash R. Honra	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
John Lorence A. Villamin	✓		✓	✓			✓			✓	✓		✓	✓

C : **C**onceptualization
M : **M**ethodology
So : **So**ftware
Va : **Va**lidation
Fo : **Fo**rmal analysis
I : **I**nvestigation
R : **R**esources
D : **D**ata Curation
O : Writing - **O**riginal Draft
E : Writing - Review & **E**diting
Vi : **Vi**sualization
Su : **Su**pervision
P : **P**roject administration
Fu : **Fu**nding acquisition

CONFLICT OF INTEREST STATEMENT

The authors declare no conflicts of interest.

DATA AVAILABILITY

Deidentified quantitative data (pretest and posttest CER scores), qualitative interview transcripts, and analysis code are available from the corresponding author upon reasonable request, subject to institutional and ethical restrictions to protect student privacy.

REFERENCES

[1] N. Ma, J. H. Guo, and Y. F. Sun, "Evaluating scientific argumentation ability: a framework based on scientific argumentation maps," *International Journal of Science Education*, pp. 1–20, 2025, doi: 10.1080/09500693.2025.2462796.

[2] L. L. Heng, J. Surif, and C. H. Seng, "Malaysian students' scientific argumentation: do groups perform better than individuals?" *International Journal of Science Education*, vol. 37, no. 3, pp. 505–528, 2015, doi: 10.1080/09500693.2014.995147.

[3] D. S. Clayton and C. Gautier, "Scientific argumentation in earth system science education," *Journal of Geoscience Education*, vol. 54, no. 3, pp. 374–382, 2006, doi: 10.5408/1089-9995-54.3.374.

[4] X. Li, W. Wang, and Y. Li, "Systematically reviewing the potential of scientific argumentation to promote multidimensional conceptual change in science education," *International Journal of Science Education*, vol. 44, no. 7, pp. 1165–1185, 2022, doi: 10.1080/09500693.2022.2070787.

[5] M. Zhu, H. S. Lee, T. Wang, O. L. Liu, V. Belur, and A. Pallant, "Investigating the impact of automated feedback on students' scientific argumentation," *International Journal of Science Education*, vol. 39, no. 12, pp. 1648–1668, 2017, doi: 10.1080/09500693.2017.1347303.

[6] U. Kulatunga, R. S. Moog, and J. E. Lewis, "Use of Toulmin's argumentation scheme for student discourse to gain insight about guided inquiry activities in college chemistry," *Journal of College Science Teaching*, vol. 43, no. 5, pp. 78–86, 2014, doi: 10.2505/4/jcst14_043_05_78.

[7] D. Lazarou and S. Erduran, "'Evaluate what I was taught, not what you expected me to know': evaluating students' arguments based on science teachers' adaptations to Toulmin's argument pattern," *Journal of Science Teacher Education*, vol. 32, no. 3, pp. 306–324, 2021, doi: 10.1080/1046560X.2020.1820663.

[8] R. Gray and N. H. Kang, "The structure of scientific arguments by secondary science teachers: comparison of experimental and historical science topics," *International Journal of Science Education*, vol. 36, no. 1, pp. 46–65, 2014, doi: 10.1080/09500693.2012.715779.

[9] M. Förster, H. R. Broder, M. C. Fahr, M. Klier, and L. Fink, "Tell me more, tell me more: the impact of explanations on learning from feedback provided by artificial intelligence," *European Journal of Information Systems*, vol. 34, no. 2, pp. 323–345, 2025, doi: 10.1080/0960085X.2024.2404028.

[10] H. Uğraş, M. Uğraş, S. Papadakis, and M. Kalogiannakis, "ChatGPT-supported education in primary schools: the potential of ChatGPT for sustainable practices," *Sustainability*, vol. 16, no. 22, p. 9855, 2024, doi: 10.3390/su16229855.

[11] J. Zhang, "School wireless network classroom teaching system based on artificial intelligence," *Applied Artificial Intelligence*, vol. 37, no. 1, p. e2219563, 2023, doi: 10.1080/08839514.2023.2219563.

[12] G. Lampropoulos and S. Papadakis, "The educational value of artificial intelligence and social robots," in *Social Robots in Education*, G. Lampropoulos and S. Papadakis, Eds., Cham: Springer, 2025, vol. 1194, pp. 3–13, doi: 10.1007/978-3-031-82915-4_1.

[13] S. Papadakis *et al.*, "Embracing digital innovation and cloud technologies for transformative learning experiences," in *Proceedings of the 11th Workshop on Cloud Technologies in Education (CTE 2023)*, 2024, pp. 1–21.

[14] B. G. Koffman, K. J. Kreutz, and K. Trenbath, "Integrating scientific argumentation to improve undergraduate writing and learning in a global environmental change course," *Journal of Geoscience Education*, vol. 65, no. 3, pp. 231–239, Aug. 2017, doi: 10.5408/16-232.1.

[15] E. M. Nussbaum and O. V. Edwards, "Critical questions and argument stratagems: a framework for enhancing and analyzing students' reasoning practices," *Journal of the Learning Sciences*, vol. 20, no. 3, pp. 443–488, Jul. 2011, doi: 10.1080/10508406.2011.564567.

[16] J. R. Honra, "Understanding the influence of biomimetic projects on transdisciplinary thinking in biology education," *Innovations in Education and Teaching International*, vol. 63, no. 3, pp. 766–780, 2025, doi: 10.1080/14703297.2025.2512386.

[17] J. R. Honra, "Development and validation of Filipino learners' academic resilience scale," *American Journal of Education and Technology*, vol. 1, no. 2, pp. 107–113, 2022, doi: 10.54536/ajet.v1i2.664.

[18] J. T. Shemwell, K. R. Gwarjanski, D. K. Capps, S. Avargil, and J. L. Meyer, "Supporting teachers to attend to generalisation in science classroom argumentation," *International Journal of Science Education*, vol. 37, no. 4, pp. 599–628, 2015, doi: 10.1080/09500693.2014.1000428.

[19] A. García-Carmona, G. Muñoz-Franco, A. M. Criado, and M. Cruz-Guzmán, "Validation of an instrument for assessing basic science process skills in initial elementary teacher education," *International Journal of Science Education*, vol. 46, no. 4, pp. 362–381, 2024, doi: 10.1080/09500693.2023.2232936.

[20] J. Tripto, O. B.-Z. Assaraf, Z. Snapir, and M. Amit, "The 'What is a system' reflection interview as a knowledge integration activity for high school students' understanding of complex systems in human biology," *International Journal of Science Education*, vol. 38, no. 4, pp. 564–595, 2016, doi: 10.1080/09500693.2016.1150620.

[21] J. R. Honra, "Enhancing high school biology education students' transdisciplinary thinking through design thinking," *The Journal of Experimental Education*, pp. 1–17, 2025, doi: 10.1080/00220973.2025.2606241.

[22] G. Shank, "Six alternatives to mixed methods in qualitative research," *Qualitative Research in Psychology*, vol. 3, no. 4, pp. 346–356, 2006, doi: 10.1177/1478088706070842.

[23] L. Restivo and T. Apostolidis, "Triangulating qualitative approaches within mixed methods designs: A theory-driven proposal based on a French research in social health psychology," *Qualitative Research in Psychology*, vol. 16, no. 3, pp. 392–416, 2019, doi: 10.1080/14780887.2019.1605670.

[24] M. Heron, E. Medland, N. Winstone, and E. Pitt, "Developing the relational in teacher feedback literacy: exploring feedback talk," *Assessment & Evaluation in Higher Education*, vol. 48, no. 2, pp. 172–185, 2023, doi: 10.1080/02602938.2021.1932735.
[25] A. Mijatović, B. Kačer, D. Miletić, T. Vranković, N. Vuco, and A. Marušić, "Ethics tools for strengthening ethics review: a scoping review," *Accountability in Research*, p. 2642202, 2026, doi: 10.1080/08989621.2026.2642202.
[26] C. T. Lye, M. M. Caron, L. Walsh, B. E. Bierer, and M. Barnes, "Disparate data retention standards in biomedical research," *Accountability in Research*, vol. 33, no. 4, pp. 1–28, 2026, doi: 10.1080/08989621.2025.2543884.
[27] D. B. Resnik, "The role of intuition in risk/benefit decision-making in human subjects research," *Accountability in Research*, vol. 24, no. 1, pp. 1–29, 2017, doi: 10.1080/08989621.2016.1198978.
[28] Y. Tsong, W. J. Chen, and C. W. Chen, "ANCOVA approach for shelf life analysis of stability study of multiple factor designs," *Journal of Biopharmaceutical Statistics*, vol. 13, no. 3, pp. 375–393, 2003, doi: 10.1081/BIP-120022761.
[29] V. Clarke and V. Braun, "Thematic analysis," *The Journal of Positive Psychology*, vol. 12, no. 3, pp. 297–298, 2017, doi: 10.1080/17439760.2016.1262613.

BIOGRAPHIES OF AUTHORS

Joelash R. Honra is an associate professor in the School of Arts and Sciences at National University, Philippines. He holds a doctorate in Biology Education from the University of the Philippines Diliman. His research interests lie at the intersection of innovative pedagogy and science instruction, with a specific focus on career-focused teaching, problem-based learning (PBL), and design thinking frameworks. He can be contacted at email: jrhonra@nu-moa.edu.ph; jrhonra@alum.up.edu.ph.

John Lorence A. Villamin is an assistant professor at National University–MOA, Philippines, and currently serves as Cluster Head of the Math and Science Cluster under the School of Arts and Sciences. He earned his Bachelor of Secondary Education major in Biological Science from Occidental Mindoro State College and completed his Master of Science in Science Education at Ateneo de Manila University. His research interests include science education, instructional design, and technology-enhanced learning, with focus on mobile learning, flipped and hyflex instructional models, learning management systems, and artificial intelligence in education. His work examines instructional approaches that support learner engagement and teaching effectiveness in higher education. He can be contacted at email: jlavillamin@nu-moa.edu.ph.