ABISATYA: Journal of Community Engagement
Vol.4 No.1 2026
ISSN: 3047-8596

Supporting Primary School Teachers in Developing English-Medium Science Materials Using Generative AI

Retno Wulan Dari[1]*, Eva Rahmawati[2], Suvi Akhiriyah[3], Wiwiet Eva Savitri[4], Him'mawan Adi Nugroho[5], Rizki Ramadhan[6]
[1]Universitas Negeri Surabaya, Indonesia (retnowulan@unesa.ac.id)
[2]Universitas Negeri Surabaya, Indonesia (evarahmawati@unesa.ac.id)
[3]Universitas Negeri Surabaya, Indonesia (suviakhiriyah@unesa.ac.id)
[4]Universitas Negeri Surabaya, Indonesia (wiwieteva@unesa.ac.id)
[5]Universitas Negeri Surabaya, Indonesia (himawannugroho@unesa.ac.id)
[6]Universitas Negeri Surabaya, Indonesia (rizkiramadhan@unesa.ac.id)

***Keywords*:**
Generative AI; Teacher Professional Development; English-Medium Instruction; Teaching Material Development

ABSTRACT

Developing English-medium Science materials for primary school students requires careful attention to language accessibility and subject accuracy. At primary school Labschool UNESA 2 Surabaya, initial coordination, teacher surveys, interviews, and reviews of existing materials highlighted the need to adapt Grade 3 Science texts to students' language proficiency. A community engagement program was therefore conducted to support teachers in this process. Twelve teachers attended two face-to-face sessions on 21 November and 2 December 2025. Working in three groups, they used ChatGPT and MagicSchool.ai to revise three Science chapters, followed by readability checks and facilitator feedback. Changes were assessed using Flesch Reading Ease, Flesch-Kincaid Grade Level, and estimated CEFR levels. A post-program questionnaire gathered participants' feedback. The revised texts showed improved readability. Average Flesch Reading Ease scores increased from 60.7 to 74.3, while Flesch-Kincaid Grade Level scores decreased from 8.7 to 5.5. Estimated CEFR levels shifted from B1+–B2 toward A2–B1 or B1. Teachers also produced revised reading texts, vocabulary sections, worksheets, and reusable prompts. Survey results showed that 83% rated their understanding of prompting as good, while 92% evaluated the program positively. However, neither student comprehension nor improvement in individual teacher competence was directly measured. Classroom testing is needed to assess the materials' effectiveness.

INTRODUCTION

English-medium programs at the primary level require teachers to manage two demands at the same time: subject concepts must remain accurate, while the language used to explain those concepts must be accessible to young learners. This is especially important when Science, Mathematics, or Social Studies are taught through English by teachers whose academic background is not English language education. Language difficulty can become an additional barrier when the wording of a text is more demanding than the learner's proficiency. The Common European Framework of Reference for Languages describes proficiency through reference levels from A1 to C2 and is widely used to support language curriculum, learning-material design, and assessment decisions (Council of Europe, 2020).

Artificial intelligence (AI) refers here to computer-based systems that can generate, transform, or analyse content in response to user instructions. Generative AI tools can support teachers by drafting explanations, creating questions, adjusting text, and producing alternative learning materials. Recent reviews show that these tools offer practical support for content creation and personalization, but they can also produce inaccurate, biased, or unsuitable output; therefore, teacher oversight remains essential

(Crompton & Burke, 2024; Lo, 2023; Montenegro-Rueda et al., 2023; Tlili et al., 2023). For school material development, the useful question is therefore not simply whether AI can generate a text, but whether teachers can guide, check, and revise that text for a specific learner group and subject purpose.

Effective technology integration also depends on teacher knowledge rather than access to a tool alone. The Technological Pedagogical Content Knowledge (TPACK) framework describes the interaction among technological knowledge, pedagogical knowledge, and content knowledge that teachers need when technology is integrated into instruction (Mishra & Koehler, 2006). Studies of technology-mediated teaching similarly emphasize the importance of faculty development, contextual adaptation, and practical support so that educators can use digital tools with confidence and pedagogical purpose (Rahayu et al., 2022; Salas-Pilco et al., 2022). This perspective is relevant to AI-assisted material development because teachers must preserve Science accuracy while simultaneously controlling language, task design, and learner suitability.

Readability and proficiency checks were used in this program as practical controls on AI-assisted revision. Flesch Reading Ease (FRE) is a readability index based on sentence length and word complexity; higher scores generally indicate easier text (Flesch, 1948). Flesch-Kincaid Grade Level (FKG), the abbreviation used in the program report, estimates the approximate United States school grade level required to understand a text; lower values indicate less demanding text (Kincaid et al., 1975). The Common European Framework of Reference for Languages (CEFR) provides proficiency bands from A1 to C2; in this program, A2-B1 was used as the target range for the Grade 3 materials (Council of Europe, 2020). These measures were treated as screening indicators for material revision, not as direct tests of students' comprehension. Thus, the intervention integrated three complementary components. TPACK informed the combination of technological tools, pedagogical decisions, and Science content knowledge. CEFR provided the program's target language range, while readability indices offered measurable indicators for evaluating text revision. Together, these components supported a structured process in which teachers generated, checked, and refined learning materials.

Primary school Labschool UNESA 2 Surabaya was selected as the partner school because English is used to teach subjects such as Science and Mathematics in its English-medium program. The initial needs assessment, conducted through school coordination, teacher surveys, and interviews, identified three challenges: adapting existing Science materials into accessible English, limited familiarity with AI-assisted material development, and the preparation demands associated with material revision. An initial examination of three Grade 3 Science chapters confirmed that their language difficulty exceeded the program's intended target. The program therefore used a hands-on revision workflow in which teachers diagnosed their existing Grade 3 Science texts, set a target language range,

wrote more specific prompts, generated or revised material, checked the new text with readability and CEFR indicators, and then edited it again after facilitator review.

Among these challenges, adapting existing Science materials to the learners' language level was prioritized because it directly affected the accessibility of instructional content. The program therefore concentrated on revising authentic teaching materials, while AI prompting and readability assessment were introduced as supporting strategies. Accordingly, this community engagement program aimed to strengthen the practical capacity of primary school Labschool UNESA 2 teachers to develop English-medium learning materials for international primary classes through a controlled AI-assisted process. Specifically, the program sought to help teachers (1) identify the language difficulty of existing materials, (2) write more specific prompts for revising reading texts and exercises, (3) move text difficulty toward the A2-B1 target used in the program, (4) maintain Science concept accuracy during simplification, and (5) produce reusable materials and prompt models for later units.

COMMUNITY ENGAGEMENT METHOD

Target Community and Participants

The target community consisted of teachers at primary school Labschool UNESA 2 Surabaya who were involved in the international-class program and in the preparation of English-medium learning materials. Twelve teachers participated in the training. The partner school contributed existing teaching-material drafts, joined the needs-analysis process, provided the venue and schedule, and participated in the revision and evaluation activities. The program team consisted of lecturers in English education from Universitas Negeri Surabaya.

Program Design and Implementation

The needs analysis involved coordination with school representatives, teacher surveys, interviews, and examination of existing Grade 3 Science materials. These activities were used to identify practical difficulties in preparing English-medium resources and to determine the focus of the training. The needs analysis identified difficulties in preparing English materials for non-language subjects, limited time for preparation, and limited familiarity with technology that could support material development. Based on these findings, the intervention was organized as two face-to-face sessions combining short conceptual input, direct practice, collaborative revision, presentation, review, and a mini clinic.

Each identified need was linked to an intervention component and an evaluation indicator. Language difficulty was addressed through guided text revision and assessed using FRE, FKG, and estimated CEFR levels. Limited AI familiarity was addressed through prompting practice and evaluated through post-program participant feedback. Preparation demands were addressed through reusable prompts and teaching-material templates, although actual time savings were not measured.

Table 1. Program implementation and immediate outputs

Session	Date and duration	Main activities	Immediate outputs
1	21 November 2025; 13.00–15.30	Draft-book review and book mapping; unit structure and learning outcomes; readability and CEFR introduction; prompting for reading texts and exercises; group revision of three Grade 3 Science chapters.	Initial readability/CEFR profile; first revised drafts; before–after files; initial reusable prompts.
2	2 December 2025; 13.00–15.00	Group presentations; review and validation by the team; mini clinic; final revision; reflection and program evaluation.	Revised reading texts, vocabulary-in-focus sections, worksheets/exercises, and standardized prompt examples.

During the first session, participants were divided into three groups of four teachers. Each group worked on one chapter of a Grade 3 Science book. The initial texts were measured to identify their readability and estimated Common European Framework of Reference level. The results indicated that the texts were generally above the program target for Grade 3 learners. Participants then revised the texts using artificial intelligence, but early outputs were reviewed because some were still unsuitable. During the training, teachers were guided to construct prompts specifying five elements: the learner group, target CEFR level, expected text length, required vocabulary, and exercise format. Revised prompts were collected as reusable examples for subsequent material development.

The second session focused on product refinement. Each group presented its revised reading text and exercises together with the before-and-after readability and language-level information. The program team reviewed the products in terms of science concept accuracy, target language level, clarity of exercise instructions, and consistency of format. A mini clinic allowed the groups to revise their materials immediately after receiving feedback. The generated materials were not accepted as final products without review. Facilitators examined Science concept accuracy, language difficulty, clarity of instructions, and consistency of presentation. Teachers subsequently revised the texts and exercises during the mini clinic in response to the feedback.

Tools, Data Collection, and Analysis

ChatGPT and MagicSchool.ai were used to support the generation and revision of Grade 3 Science texts and accompanying exercises. Their outputs were subjected to readability assessment and facilitator review before further revision. The learning products included reading texts, vocabulary-in-focus sections, worksheets, exercises, and a reusable prompt bank. Program evaluation used two forms of evidence: product-based before-and-after measurements and a post-program participant survey.

For the product analysis, the three group texts were measured before and after revision using Flesch Reading Ease (FRE), Flesch-Kincaid Grade Level (FKG), and an estimated Common European Framework of Reference for Languages (CEFR) level. FRE was used as an indicator of general reading ease, FKG as an approximate grade-level indicator, and CEFR as a proficiency-reference indicator. The analysis was descriptive and examined whether the revised texts moved toward easier readability and the program target of A2-B1. Because the unit of analysis was the three collaboratively revised chapter texts, no inferential statistical test was applied.

After the second session, all 12 participants completed a simple program-evaluation questionnaire. Ten items asked about understanding of readability, language-level checking, prompting, vocabulary selection, worksheet development, relevance of the training, clarity of explanation, usefulness of mentoring, time effectiveness, and overall satisfaction. Responses were summarized in three categories: Good, Fair, and Poor. The final report did not provide psychometric validation information for this questionnaire; therefore, the survey is treated as descriptive feedback on the community engagement program rather than as a validated research scale.

RESULTS AND DISCUSSION

Results

The needs analysis confirmed that the main difficulty was not the absence of teaching materials, but the process of adapting them for English-medium Grade 3 instruction. Teachers already had draft materials, yet the language was often too difficult for the intended learners. The training therefore began from the teachers' own products rather than from generic examples. This choice increased the practical relevance of the activity because every group worked on a real chapter that could later be used in the school's material-development process.

The first measurement showed that all three Science texts were above the intended target. Group 1 recorded a Flesch Reading Ease (FRE) score of 61 and a Flesch-Kincaid Grade Level (FKG) score of 8.9 with an estimated Common European Framework of Reference for Languages (CEFR) level of B2. Group 2 recorded FRE 63 and FKG 8.2 with an estimated B1+-B2 level, while Group 3 recorded FRE 58 and FKG 9.1 with an estimated B2 level. After revision, all groups moved toward easier text profiles. The average FRE increased from 60.7 to 74.3, while the average FKG decreased from 8.7 to 5.5. The estimated language levels moved to B1 or A2-B1, closer to the target used in the program.

Table 2. Readability and language-proficiency profile before and after revision

Group / chapter	Before revision (FRE / FKG / CEFR)	After revision (FRE / FKG / CEFR)	Main change
Group 1 / Chapter 1	FRE 61 FKG 8.9 CEFR B2	FRE 74 FKG 5.6 CEFR B1	Shorter text, simpler sentence structure, and more controlled vocabulary.

Group / chapter	Before revision (FRE / FKG / CEFR)	After revision (FRE / FKG / CEFR)	Main change
Group 2 / Chapter 2	FRE 63 FKG 8.2 CEFR B1+–B2	FRE 76 FKG 5.2 CEFR A2–B1	Clearer exercise instructions, more accessible reading, and more contextual examples.
Group 3 / Chapter 3	FRE 58 FKG 9.1 CEFR B2	FRE 73 FKG 5.8 CEFR B1	Fewer complex clauses and passive structures; simpler explanations of terms.
Average	FRE 60.7 FKG 8.7	FRE 74.3 FKG 5.5	Overall readability moved toward an easier level and the language level moved closer to the target.

The changes were also visible in the teachers' working process. At the beginning, prompts tended to be general, and the generated text did not always match the desired level or format. During mentoring, teachers learned to state a target language level, specify the learner group, limit text length, identify required vocabulary, and define the expected exercise format. They then checked the output and edited it instead of using it automatically. The resulting products included revised reading texts, vocabulary-in-focus sections, worksheets or exercises, and prompt examples that could be reused for later chapters. The program outputs consisted of facilitator-reviewed teaching-material drafts, including revised reading texts, vocabulary sections, worksheets, and reusable prompts. Although these products underwent internal review and revision, they had not been validated through classroom implementation or direct student comprehension testing.



Figure 1. The facilitator providing feedback on participants' teaching materials

Participant feedback supported the product-based findings. The highest positive response for understanding was related to prompting: 83% of participants selected Good for their understanding of how to write prompts that simplify reading texts without changing meaning. Seventy-five percent selected Good for understanding readability

and for preparing vocabulary-in-focus sections, while 67% selected Good for language-level checking and artificial-intelligence-assisted worksheet development. The relevance of the training, clarity of the team's explanations, and usefulness of the review and mini clinic each received 83% Good responses. Overall satisfaction received the strongest rating, with 92% Good.

Table 3. Participant understanding and satisfaction after the program (n = 12)

Evaluation item	Good	Fair	Poor
Understanding Flesch Reading Ease and Flesch-Kincaid readability results	9 (75%)	3 (25%)	0 (0%)
Checking Common European Framework of Reference for Languages level for Grade 3	8 (67%)	4 (33%)	0 (0%)
Prompting to simplify reading without changing meaning	10 (83%)	2 (17%)	0 (0%)
Preparing vocabulary in focus for topic and learner level	9 (75%)	3 (25%)	0 (0%)
Developing AI-assisted worksheets with clear instructions	8 (67%)	4 (33%)	0 (0%)
Training materials matched international-class needs	10 (83%)	2 (17%)	0 (0%)
Program team explained materials clearly and systematically	10 (83%)	2 (17%)	0 (0%)
Review and mini clinic helped improve the products	10 (83%)	2 (17%)	0 (0%)
Two-session schedule and flow were effective	8 (67%)	3 (25%)	1 (8%)
Overall satisfaction with the program	11 (92%)	1 (8%)	0 (0%)

Discussion

The results show that the program produced an immediate improvement in the linguistic accessibility of the three Grade 3 Science texts and in the teachers' revision process. This finding supports the practical use of generative artificial intelligence as a teaching-support tool rather than as an autonomous material writer. Reviews of ChatGPT in education report that teachers can use generative artificial intelligence for content creation, task automation, and personalized materials, while also warning that generated content can be inaccurate or unsuitable and therefore requires checking and professional judgment (Crompton & Burke, 2024; Lo, 2023). In the present program, the value of the technology came from combining generation with explicit language targets, readability checks, subject-content review, and teacher editing.



Figure 2. Participants presenting their revised teaching materials

The program generated evidence at two levels: changes in the readability profiles of three collaboratively revised texts and participants' post-program perceptions. These findings should not be interpreted as direct evidence of individual teacher learning gains or sustained institutional adoption, neither of which was systematically assessed. The readability change provides the clearest product-level evidence of the intervention. Flesch Reading Ease increased from an average of 60.7 to 74.3, while Flesch-Kincaid Grade Level decreased from 8.7 to 5.5. Because higher FRE values represent easier text and lower FKG values represent a lower estimated grade demand, both indicators moved in the intended direction (Flesch, 1948; Kincaid et al., 1975). The estimated CEFR profile also moved from B1+-B2 toward B1 or A2-B1, which was closer to the program target (Council of Europe, 2020). These measures do not prove student comprehension, but they show that teachers were able to make language demands more deliberate. This is pedagogically relevant because unnecessary linguistic complexity can add cognitive load that is not directly related to learning the Science concept itself (Sweller, 1988).

Another important outcome was the change in prompting behavior. Teachers moved from general instructions to prompts that specified learner profile, target CEFR level, text length, required vocabulary, and exercise format. This shift is important because educational uses of generative AI depend heavily on the quality of human instructions and on subsequent evaluation of the output. Reviews have repeatedly emphasized the need for teacher training, critical checking, and responsible human oversight when generative AI is used for educational content (Montenegro-Rueda et al., 2023; Tlili et al., 2023). The facilitators' content review was particularly important in this program because simplifying language can unintentionally distort a science concept even when the resulting text is grammatically fluent.

The program design also benefited from using authentic school materials and practice-based mentoring. The Technological Pedagogical Content Knowledge (TPACK) perspective suggests that meaningful technology integration emerges from the interaction of technology, pedagogy, and disciplinary content rather than from technical skill alone (Mishra & Koehler, 2006). The training reflected this interaction: teachers used

AI as the technology, readability and proficiency targets as part of pedagogical adaptation, and Science accuracy as the content constraint. The mini clinic, presentation, and immediate revision cycle also aligns with evidence that educators need contextual professional development and opportunities to adapt technology to their actual teaching tasks (Rahayu et al., 2022).

Community participation was visible throughout the program. The school contributed to problem identification and scheduling, teachers supplied their own book drafts, and the 12 participants revised the materials collaboratively in three groups. Starting from existing Grade 3 chapters made the intervention directly relevant to the partner's workflow. Contextual adaptation is important because digital learning practices are shaped by local resources, teacher experience, and institutional conditions rather than by the technology alone (Ngoasong, 2022). Likewise, personalized or adaptive learning resources require explicit criteria for how materials should be adjusted to users' needs (Thalmann, 2014). The program's reusable prompt bank and common material components provide an initial mechanism for continued adoption.

Several limitations should be considered. First, the before-and-after analysis was based on three collaboratively revised chapter texts rather than individual teacher pre-tests and post-tests. Second, the post-program survey was a simple descriptive evaluation instrument, and the final report did not provide evidence of its validity or reliability. Third, the planned reduction in teachers' preparation time was not measured, so the article does not claim a verified time-saving effect. Fourth, readability formulas and CEFR estimates are screening tools and cannot substitute for direct evidence that Grade 3 students understand the revised texts. Fifth, the materials had not yet been piloted with students when the final report was prepared. Finally, only 67% of participants rated the two-session schedule as good, suggesting that more time is needed for product finalization.

For sustainability, the school can continue the process through a short internal material clinic every two or three weeks using the same sequence of target setting, prompting, output review, readability and CEFR checking, and final editing. Continued teacher support is consistent with evidence that successful technology-mediated learning depends on professional development, engagement, and adaptation to local conditions (Salas-Pilco et al., 2022). The next stage should include content validation, a common unit template, and a limited classroom pilot so that teacher observations and student responses can be used to refine the materials before wider implementation.

CONCLUSION

The community engagement program supported twelve teachers at primary school Labschool UNESA 2 Surabaya in revising English-medium Science materials for Grade 3 students. Through two training sessions, teachers worked collaboratively to revise three Science chapters using generative AI, readability assessment, and facilitator feedback. The revised texts showed measurable changes in language difficulty. The average Flesch Reading Ease score increased from 60.7 to 74.3, while the Flesch-Kincaid Grade Level decreased from 8.7 to 5.5. Estimated CEFR levels also shifted from B1+–B2 towards B1 or A2–B1, bringing the materials closer to the program's intended language target. In

addition to the revised reading texts, teachers produced vocabulary sections, worksheets, and reusable prompts. Participant feedback was generally positive, with 92% rating the program as good overall.

The program also highlighted the importance of teacher involvement when using AI to develop learning materials. Generating simpler texts was only one part of the process. Teachers still needed to check language difficulty, review Science content, and revise the materials to suit their students. However, the findings are limited to changes in three collaboratively revised texts and participants' post-training perceptions. Improvements in individual teacher competence and student comprehension were not directly measured. The materials had also not been tested in actual classroom teaching. Further mentoring, classroom trials, and feedback from teachers and students are therefore needed to assess their practical suitability and support continued material development at the partner school.

ACKNOWLEDGMENTS

The authors thank the leadership and teachers of primary school Labschool UNESA 2 Surabaya for their collaboration and active participation. The program was supported through the 2025 Faculty Policy Community Engagement scheme of the Faculty of Languages and Arts, Universitas Negeri Surabaya. The final program report did not identify any conflict of interest.

REFERENCES

Crompton, H., & Burke, D. (2024). The educational affordances and challenges of ChatGPT: State of the field. *TechTrends, 68*, 380-392. https://doi.org/10.1007/s11528-024-00939-0

Council of Europe. (2020). *Common European Framework of Reference for Languages: Learning, teaching, assessment - Companion volume*. Council of Europe Publishing. https://www.coe.int/en/web/common-european-framework-reference-languages/cefr-companion-volume-and-its-language-versions

Flesch, R. (1948). A new readability yardstick. *Journal of Applied Psychology, 32*(3), 221-233. https://doi.org/10.1037/h0057532

Kincaid, J. P., Fishburne, R. P., Jr., Rogers, R. L., & Chissom, B. S. (1975). *Derivation of new readability formulas (Automated Readability Index, Fog Count and Flesch Reading Ease Formula) for Navy enlisted personnel* (Research Branch Report 8-75). Chief of Naval Technical Training. https://stars.library.ucf.edu/istlibrary/56/

Lo, C. K. (2023). What is the impact of ChatGPT on education? A rapid review of the literature. *Education Sciences, 13*(4), 410. https://doi.org/10.3390/educsci13040410

Mishra, P., & Koehler, M. J. (2006). Technological pedagogical content knowledge: A framework for teacher knowledge. *Teachers College Record, 108*(6), 1017-1054. https://doi.org/10.1111/j.1467-9620.2006.00684.x

Montenegro-Rueda, M., Fernández-Cerero, J., Fernández-Batanero, J. M., & López-Meneses, E. (2023). Impact of the implementation of ChatGPT in education: A

systematic review. *Computers, 12*(8), 153. https://doi.org/10.3390/computers12080153

Ngoasong, M. Z. (2022). Curriculum adaptation for blended learning in resource-scarce contexts. *Journal of Management Education, 46*(4), 622-655. https://doi.org/10.1177/10525629211047168

Rahayu, G. R., Utomo, P. S., Riskiyana, R., & Hidayah, R. N. (2022). Opportunity amid crisis in medical education: Teaching during the pandemic of COVID-19. *Journal of Multidisciplinary Healthcare, 15*, 2493-2502. https://doi.org/10.2147/JMDH.S379140

Salas-Pilco, S. Z., Yang, Y., & Zhang, Z. (2022). Student engagement in online learning in Latin American higher education during the COVID-19 pandemic: A systematic review. *British Journal of Educational Technology, 53*(3), 593-619. https://doi.org/10.1111/bjet.13190

Sweller, J. (1988). Cognitive load during problem solving: Effects on learning. *Cognitive Science, 12*(2), 257-285. https://doi.org/10.1207/s15516709cog1202_4

Thalmann, S. (2014). Adaptation criteria for the personalised delivery of learning materials: A multi-stage empirical investigation. *Australasian Journal of Educational Technology, 30*(1), 45-60. https://doi.org/10.14742/ajet.235

Tlili, A., Shehata, B., Adarkwah, M. A., Bozkurt, A., Hickey, D. T., Huang, R., & Agyemang, B. (2023). What if the devil is my guardian angel: ChatGPT as a case study of using chatbots in education. *Smart Learning Environments, 10*, Article 15. https://doi.org/10.1186/s40561-023-00237-x