IIAI Open Conference Publication Series
IIAI Letters on Informatics and Interdisciplinary Research
Vol.006, LIIR427
DOI: https://doi.org/10.52731/liir.v006.427

# Assessing Reflective Learning through Human Revision of AI-Generated Essays: A Multi-Phase Study

Satoshi Kimura [*], Taoku Yasunaga [*]

## Abstract

This study examined how university students utilize generative AI in the context of writing admissions essays and how the depth of their reflective thinking affects the quality of AI-assisted writing. One hundred twenty-six students participated in five types of writing tasks modeled on university application prompts, with varying levels of AI involvement. Each submission was blind-reviewed using a four-level rubric designed to capture finer distinctions in structure, logic, and expression. The results showed that, while the influence of initial writing ability was limited to the early stages of AI engagement, the depth of reflection—measured as the Reflection Depth Score (RDS)—was significantly associated with the quality of outputs across all tasks. Participants with high RDS demonstrated greater score improvement in later tasks, while those with low RDS sometimes experienced declines in performance. These findings suggest that the educational effectiveness of generative AI depends not only on its available skills but also on the learner's metacognitive abilities, underscoring the importance of reflective and dialogic processes in AI-integrated writing instruction.

*Keywords:* generative AI, AI literacy, admissions essays, metacognition

## 1 Introduction

In recent years, the rapid advancement and widespread adoption of generative AI technologies have drawn increasing attention to their implications for education. In particular, ChatGPT, released in late 2022, saw rapid global adoption owing to its advanced capabilities in natural language processing, ease of use, and broad applicability [1]. The Ministry of Education, Culture, Sports, Science and Technology (MEXT) of Japan has emphasized the need to strengthen students' information literacy in light of the emergence of generative AI. In its official guidelines, MEXT states that learners are expected to "develop an attitude of active participation in an information-oriented society by appropriately and effectively utilizing generative AI for problem identification and problem-solving" [2]. Accordingly, generative AI is being explored in a variety of educational contexts, including learning support, instructional material development, and writing assistance [3, 4]. However, the Expert Panel convened by MEXT has reported that the utility of AI-generated output may vary depending on learners' language comprehension, reflective thinking, and metacognitive abilities [5].

One notable consequence of this reform has been the increasing use of comprehensive admission methods, including Admission Office (AO) entrance examinations and Designated School Recommendation admissions[6,7]. As of the 2023 academic year, more than 50% of

[*] Kyushu Institute of Technology, Fukuoka, Japan

admitted students in Japanese universities entered through such pathways [8]. These types of admissions place significant emphasis not only on test-based academic competencies but also on self-expressive documents such as statements of purpose and activity reports. These changes have elevated the importance of self-expressive documents, such as statements of purpose and activity reports, within the admissions process [7]. Consequently, many high schools in Japan now incorporate statement writing into educational activities such as the Period for Inquiry-Based Cross-Disciplinary Study, Homeroom Activities within Tokkatsu (Student-Led Activities), and career guidance programs. These opportunities encourage students to reflect on their future aspirations and articulate their academic and personal goals [9]. Reflecting these trends, a web-based survey conducted by the authors between October and December 2024 (N = 581) found that approximately 78% of respondents aged 15–26 believed it was acceptable to use generative AI in university admissions or job applications, while 56% reported a willingness to do so. These findings suggest that the psychological barriers to AI usage in self-expressive tasks are diminishing among younger populations. Nonetheless, the use of generative AI does not automatically translate into improved writing quality or favorable evaluations. A 2022 study by Kimura et al. [10] revealed that human raters could identify application essays written by AI with approximately 83% accuracy, and those perceived as AI-generated tended to receive lower scores. These results suggest that evaluative outcomes are influenced not only by textual quality but also by subjective impressions and expectations of the readers.

Given these findings, the focus must shift from whether generative AI is used to how it is used. Learners who revise and personalize AI-generated texts to reflect their own intentions and experiences may derive more benefit from such technologies. These skills involve more than technical proficiency; they require higher-order cognitive capacities such as linguistic metacognition and reflective judgment.

While prior studies have examined the effectiveness of generative AI in writing tasks, few have addressed how learners' cognitive and metacognitive attributes mediate that effectiveness. The present study investigates the relationship between generative AI usage strategies and the quality of written outcomes. A group of university students were asked to produce application essays under multiple conditions using generative AI. These essays were then blindly evaluated by human raters using a standardized rubric. The study further examines how students' AI usage skills and their depth of reflection—quantified using a Reflection Depth Score (RDS)—are associated with the quality of their final written products. Specifically, we analyze how scores vary across conditions where AI outputs were used as-is, interactively revised, or refined following peer review and self-reflection.

This study addresses the below question: How do initial writing ability and reflection depth shape the outcomes of AI-supported writing, particularly across different modes of AI involvement?

## 2 Methods and Results

This study modeled short-form application essays commonly required in actual university admissions processes in Japan. These essays are typically submitted via online application systems and require applicants to express their motivations concisely and logically within a character limit. Rather than employing complex classification models, the study focused on understanding overall trends through descriptive statistics, non-parametric group comparisons, and rank-based correlation analyses to examine how different AI usage strategies and reflective thinking influenced essay quality. All procedures involving human participants were approved by the Human Research Ethics Committee of Kyushu Institute of Technology.

## 2.1 Methods

This study was conducted with 126 third-year undergraduate students majoring in Computer Science and Systems Engineering at Kyushu Institute of Technology. Participants were asked to write multiple versions of a mock application essay in response to the following prompt, which simulates an actual university entrance examination:

*"Describe a social or personal issue you wish to address through the study of computer science and systems engineering. Explain what you want to learn and achieve at Kyushu Institute of Technology to realize that goal. (Within 400 Japanese characters)"*

Each student completed five writing tasks in the following order:

- Method A: Written without using AI
- Method B: Generated by AI only
- Method C: Revised using AI dialogue based on the text from Method B
- Method D: Revised based on peer feedback on Method C
- Method E: Generated again using AI, informed by the experiences of Methods A–D

This process yielded a total of 655 responses. After completing all tasks, participants were also asked to submit a written reflection. All essays were anonymized and randomly shuffled. A group of five evaluators (two university faculty members and three undergraduate or graduate students) independently scored the essays using a 4-point rubric designed to capture three dimensions: Suitability (I), Literacy (II), and Proactivity (III). This rubric was newly developed to provide higher resolution in evaluating the quality of reflective and expressive writing in university admissions contexts. Each dimension was rated from 0 to 3 for a total possible score of 0 to 9 (Table 1).

Table 1: Evaluation Rubric for Admissions Essay Assessment

| Criterion | 0 | 1 | 2 | 3 |
|---|---|---|---|---|
| **I.**<br>**Relevance**<br>*Alignment with engineering fields* | No issue or topic is stated, or the topic presented is inappropriate for an engineer to address. | A topic that is reasonably appropriate for an engineer is presented. | A topic appropriate for an engineer is presented, with a vague image of how to approach the solution. | A topic suitable for an engineer is presented, along with a clear idea of how it might be addressed, such as through learning or practical methods. |
| **II.**<br>**Literacy**<br>*Logical structure and clarity in limited-length writing* | The response does not address the assigned prompt or shows no awareness of the reader, resulting in a self-centered narrative. | The response addresses the required points. While there may be some lack of fluency, the logic is generally coherent. | The response is logically structured, easy to read, and clearly addresses the assigned task. | The writing demonstrates awareness of the reader and uses expressions appropriate for conveying ideas within a limited word count. |
| **III. Self-Direction** | No clear reason for applying is pro- | A general statement of motiva- | The motivation is expressed based | The motivation is clearly and con- |

| | | | | |
|---|---|---|---|---|
| *Personal motivation and clarity of goals* | vided (i.e., what the applicant wants to learn, achieve, or strive for), or the stated motivation lacks coherence with personal experiences. | tion (e.g., learning goals, aspirations, or efforts) is provided, even if in abstract terms. | on personal experiences or reflections, even if somewhat abstract or loosely connected. | cretely articulated based on specific personal experiences or ideas. Use of insightful abstract or metaphorical expressions is encouraged. |

To assess metacognitive engagement, we also evaluated each participant's reflection using the Reflection Depth Score (RDS). The RDS was derived from ratings based on Table 2's rubric. Three sources rated each reflection: the first author, ChatGPT-4o, and -o1 (both versions released after April 3, 2024). The RDS rubric was adapted from the four-level reflection model proposed by Kember et al. [11] to fit the context of generative AI use. It evaluates learners' cognitive engagement with AI-generated content, such as prompt design, structural evaluation, and awareness of AI limitations. Each model received the same reflection text and rubric. The final RDS for each participant was calculated as the median of the three scores. Data were analyzed using Jamovi (version 2.4) [12]. Specifically, we conducted non-parametric tests such as Kruskal–Wallis tests, Dunn's post hoc comparisons, and Spearman's rank correlation analyses to accommodate non-normal distributions and ordinal-scale rubric data.

Table 2: Rubric for Reflection Depth Score (RDS)

| **Score** | **Reflection Level** | **Descriptors** |
|---|---|---|
| **0 (Shallow)** | Surface-level or procedural reflection only | - Feedback consists of vague impressions such as "It was convenient" or "Amazing"<br>- No clear evaluation of AI outputs or feedback is extremely ambiguous<br>- Lacks reference to concrete personal experiences or contains overly general content |
| **1 (Somewhat Shallow)** | Basic observations or impressions about AI | - Mentions differences between AI and humans, but in abstract terms with little connection to practice<br>- Describes experience but lacks reference to improvement or application<br>- Refers to writing quality but with insufficient rationale |
| **2 (Moderate)** | Concrete evaluation of interaction with AI and applied adjustments | - Describes specific operations, e.g., editing, testing, or modifying prompts<br>- Identifies challenges or difficulties in use and discusses how these led to adjustments or changes in perspective<br>- Acknowledges division of roles or complementary functions between humans and AI |
| **3 (Deep)** | Reflection involving metacognitive awareness and future application | - Demonstrates consideration of structural limitations, errors, or biases in AI output<br>- Shows strategic understanding of prompt design and AI-human dialogue<br>- Expands perspective to compare with human thinking and consider societal implications<br>- Applies feedback to writing structure and discusses future use in practice |

## 2.2 Comparison of Writing Methods Based on Rubric Scores

Figures 1 display the distribution of rubric scores for the three evaluation criteria—Aptitude (I), Literacy (II), and Autonomy (III)—across the five writing methods. The rubric used assigns

0 to 3 points per criterion, with a maximum total of 9 points.

As shown in Figure 1, Kruskal–Wallis tests revealed that Method B (AI-only generation) produced significantly higher scores than Method A (non-AI writing) across all evaluation criteria (I: p = .006; II: p = .003; III: p = .021). This suggests that even AI-generated outputs, when used directly, may exhibit a certain level of structural and expressive quality.

Furthermore, there was a general upward trend in scores from Method B to Method C (dialogue-based human revision of AI-generated draft) and Method D (peer-reviewed and self-revised version of Method C). Particularly in the dimensions of Literacy and Self-Direction, Methods C, D, and E demonstrated higher median scores than both Methods A and B. However, these differences were not statistically significant in all cases. Post-hoc comparisons using the Dwass–Steel–Critchlow–Fligner test indicated marginally significant improvements between Methods B and C in Literacy (p = .072) and Self-Direction (p = .081). These findings suggest that learners who engaged in interactive editing and self-reflection processes were more successful in enhancing the structural and expressive quality of their texts.

The median total score increased from Method A (3.20) to Method E (4.20), with Methods C, D, and E all showing higher medians than Method A and B. In Criterion I, the median scores rose from 1.00 in Method A to 1.40 in Method E; in Criterion II, from 1.00 to 1.40; and in Criterion III, from 1.20 to 1.40. This pattern suggests progressive refinement through reflective engagement and peer or AI-based revision.  This pattern suggests that the different forms of human-AI interaction may be associated with varying levels of writing quality across methods.

In particular, Method E—AI-only generation informed by prior task experiences—yielded significantly higher median scores than Method B (p = .002), indicating a potential relationship between prior engagement and improved outputs.

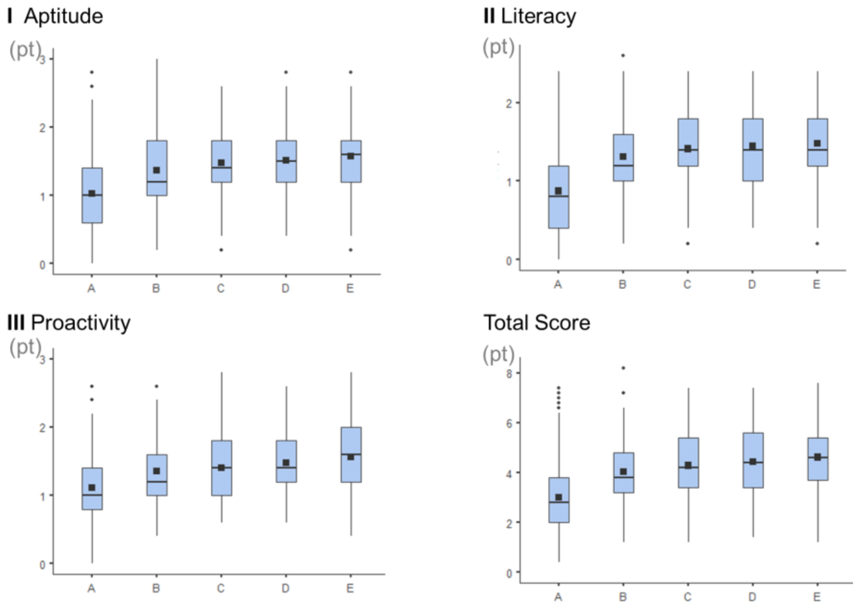

Figure 1:  Distribution of Scores by Evaluation Criteria

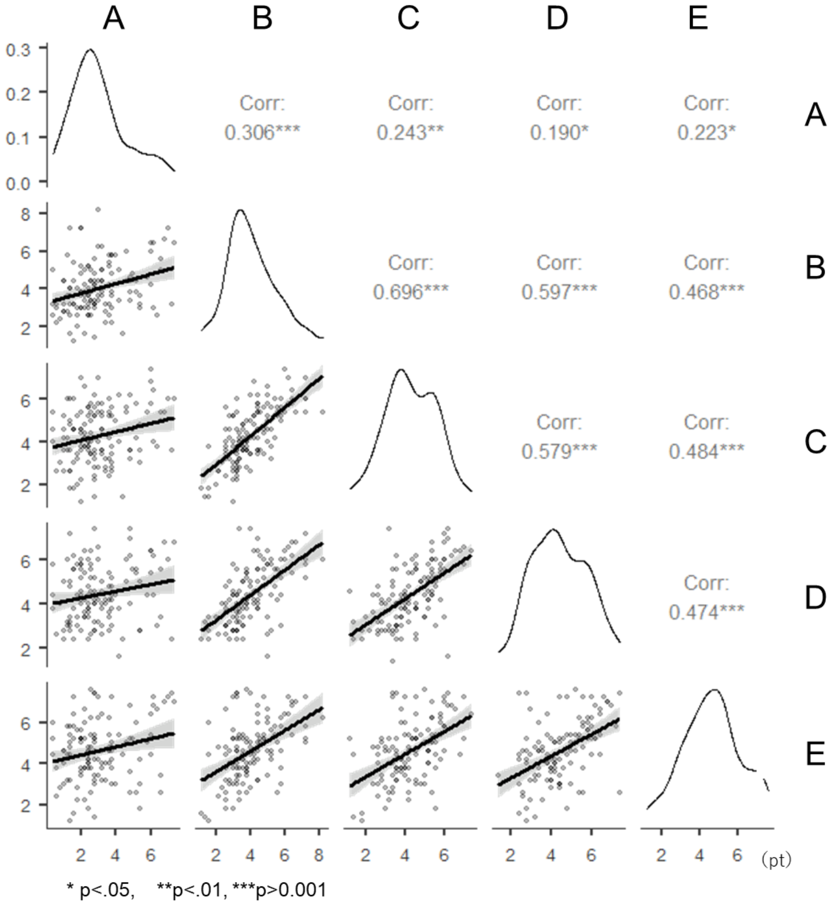

Figure 2: Rank Correlation of Total Scores among Different Writing Methods

## 2.3 Correlation Between Writing Methods

To address the ordinal nature and non-normal distribution of the rubric-based scores, Spearman's rank correlation coefficients (ρ) were calculated across Methods A through E (Figure 2). The analysis revealed a statistically significant monotonic correlation between Method A and Method B ($\rho = .300$, $p < .001$) and a marginally significant correlation with Method C ($\rho = .187$, $p < .1$). No significant associations were observed between Method A and Methods D or E.

These results suggest that foundational writing ability influences AI-supported writing only in the early phases, with its impact diminishing in later tasks involving peer feedback or AI regeneration.

The significant correlation between Methods A and B further indicates that learners' ability to write and structure content may affect the quality of the prompts provided to the AI, highlighting the importance of linguistic and metacognitive engagement in prompt formulation.

## 2.4 Relationship Between Initial Writing Ability and AI-Assisted Writing Outcomes

To examine how initial writing ability influenced the quality of AI-assisted outputs, participants were grouped into three levels based on their scores on Method A (non-AI writing): A-L (A < 2.5), A-M (2.5 ≤ A < 4.0), and A-H (A ≥ 4.0). The scores for Methods B through E were then compared across these groups using the Kruskal–Wallis test and subsequent post-hoc comparisons via the Dwass–Steel–Critchlow–Fligner test (Figure 3).In Method B, the A-H group scored significantly higher than the A-L group (p = .010), while the A-M group did not differ significantly from either (p = .091 vs. A-L). In Method C, both A-M and A-H groups showed significantly higher scores than A-L (p = .039 and .030, respectively). No significant differences were observed in Methods D or E. These results suggest that learners with stronger initial skills may be more effective at revising AI-generated content, particularly in the earlier stages of human-AI interaction. The median total score in Method B was 3.6 for the A-L group and 4.4 for the A-H group, while in Method E, the corresponding medians were 4.4 and 5.2. This narrowing of the gap (1.80 → 1.00) suggests that repeated engagement with AI, including prompting, evaluation, and reflection, can help reduce disparities  stemming from differences in initial writing ability.

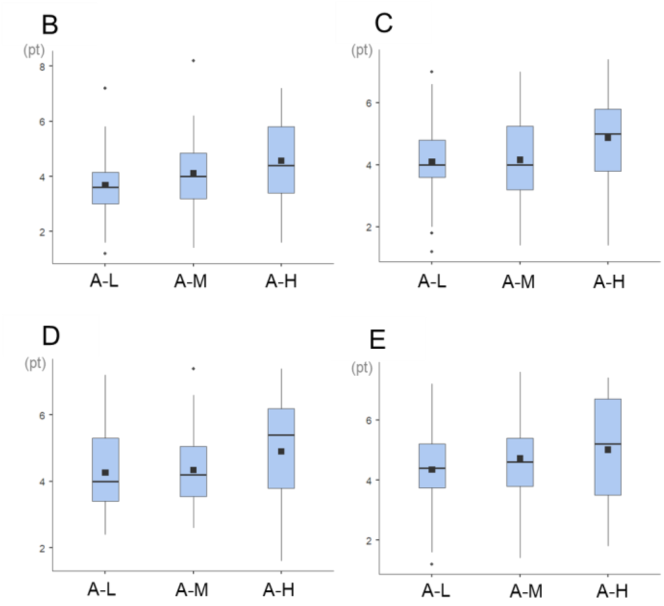

Figure 3: Score Distribution by Initial Writing Ability Groups (A-L/M/H) across Methods B to E

As shown in Figure 4, among participants in the A-L group, those with high RDS scores showed significantly higher performance in Methods C (p = .003), D (p = .001), and E (p = .007) than those with low RDS. In Method C, the median total scores were 3.0 for RDS-L and 4.0 for RDS-H, with mean scores of 3.22 and 3.99, respectively. In Method D, the medians were equal at 3.8, but mean scores differed (RDS-L = 3.67, RDS-H = 4.04). In Method E, the median was 3.8 for RDS-L and 4.4 for RDS-H, with mean scores of 3.74 and 4.53.

These findings underscore the importance of reflective engagement in improving writing quality, especially for learners who initially struggled with AI-assisted writing tasks.

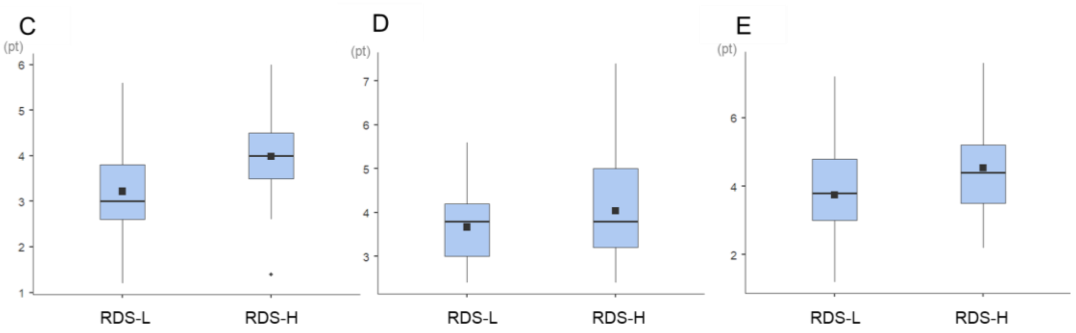

Figure 4: Score Improvements in Methods C–E by RDS, Limited to the B-L Group (Low Scorers in Method B)

## 2.5 Reflection Depth and Its Impact on Revision and AI Utilization Skills

To examine the relationship between revision performance and the depth of reflection (RDS), analyses were conducted focusing on participants whose initial scores on Method B (AI-only generation) were below the overall average This restriction was applied because participants with already high scores (B-H group) had limited room for improvement, making it difficult to assess the true effects of reflection. Based on their RDS scores, participants were classified into two groups: RDS-L (RDS < 2) and RDS-H (RDS ≥ 2). Mann–Whitney U tests were used to compare scores on Methods C, D, and E, as well as their differences relative to Method B (i.e., C−B, D−B, and E−B).

A statistically significant difference was observed in the initial scores of Method B between RDS groups (p = .006, rank-biserial correlation r = 0.426). In Method C, the RDS-H group outperformed the RDS-L group (p = .009, r = 0.396). In Method D, the difference remained significant (p = .020, r = 0.361). In contrast, no significant difference was observed in E–B (p = .322, r = 0.079) or D–B (p = .615, r = 0.048). C–B (p = .436, r = 0.029) and D–C (p = .892, r = 0.280) were also not significant. To supplement these findings, descriptive statistics are provided. In Method C, the median scores were 3.0 for RDS-L and 4.0 for RDS-H, with means of 3.22 and 3.99, respectively. In Method D, the medians were equal at 3.8, but the mean scores were 3.67 (RDS-L) and 4.04 (RDS-H). In Method E, the median was 3.8 for RDS-L and 4.4 for RDS-H, and the mean scores were 3.74 and 4.53.

Further analysis focused on learners in the A-L group to examine how reflection depth affected D–B scores. A Kruskal–Wallis test revealed a statistically significant difference among the four subgroups based on Method B performance and RDS level ($\chi^2$ = 8.75, df = 3, p = .033, $\varepsilon^2$ = 0.186). Although pairwise comparisons using the Dwass–Steel–Critchlow–Fligner test did not identify significant differences between all groups, the RDS-H subgroup that had high Method B scores showed the most stable outcomes (−0.05), while the RDS-L subgroup in the same category exhibited a marked decline (−1.85) (figure 5).

These results suggest that deeper reflection, as measured by RDS, may buffer against decreases in writing quality during peer-reviewed revision. In contrast, learners with low RDS may be more vulnerable to performance decline, even if their initial AI-generated drafts were strong.

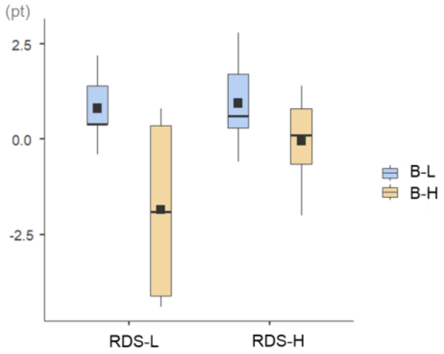

Figure 5: Score Change (D-B) by RDS  (Low Scorers in Method A)

# 3 Discussion and Conclusion

## 3.1 Advantages of a Four-Level Rubric for Capturing Learning Progression

The implementation of a four-level rubric enabled a more nuanced evaluation of students' writing across multiple dimensions. Although scoring with this rubric required slightly more time, the additional effort is justified by its ability to visualize learner growth and provide targeted feedback. The quantitative distinctions also enabled subsequent analysis of reflection and writing quality, suggesting that this rubric is valuable not only for assessment but also for learning process modeling.

Interestingly, the current results showed that AI-generated texts (Method B)  tended to score higher than human-only texts (Method A), particularly in structural and linguistic aspects. This contrasts with our previous study [10], in which AI-generated essays were rated lower than human-written ones. However, that study focused on perceived AI-generatedness, which may have influenced scoring bias. While this shift may partly be attributed to differences in rubric design, it is also likely that the improved capabilities of generative AI systems over the past year played a substantial role.

## 3.2 Consistency of Writing Quality Across Methods

Spearman's correlation analysis revealed a statistically significant monotonic association between Method A and Method B ($\rho = .300$, $p < .001$) and a marginally significant association with Method C ($\rho = .187$, $p < .1$). In contrast, no significant correlations were found with Methods D or E (Figure 2).

These findings suggest that foundational writing ability plays a role in early AI-supported writing but exerts minimal influence on later-stage outputs shaped by peer or iterative feedback. The observed correlation between Methods A and B further indicates that linguistic competence affects not only how learners refine AI outputs but also how they construct prompts, underscoring the importance of supporting learners' metacognitive and communicative engagement when integrating generative AI into writing tasks.

## 3.3 Contribution of Initial Writing Ability to AI-Supported Output Quality

A comparative analysis was conducted to examine how initial writing ability influenced the quality of outputs generated through different AI-supported writing methods. Based on their scores in Method A (human-only composition), participants were grouped into three levels: A-L (low), A-M (medium), and A-H (high). As shown in Figure 3, learners in the A-H group tended to achieve higher scores, particularly in Methods B and C, suggesting that foundational writing skills provide a consistent advantage in AI-supported contexts. The significant score gap in Method C suggests that stronger writers revised AI-generated drafts more effectively. This implies that linguistic metacognitive skills—such as the ability to assess, interpret, and selectively integrate AI feedback—play a crucial role in enhancing the quality of AI-assisted writing.

However, in Methods D (peer-reviewed and self-revised version of Method C) and E (AI-only generation informed by A–D), these score differences diminished and were no longer statistically significant. This finding suggests that reflective and collaborative processes may help reduce disparities stemming from initial ability by offering all learners opportunities to iteratively refine their work and internalize effective revision strategies.

Furthermore, the analysis found a significant positive correlation between scores in Method A and Method B (AI-only generation), underscoring the influence of learners' linguistic abilities on the quality of their prompts. Since the effectiveness of AI-generated output is partially determined by prompt clarity and structure, this result highlights the important role of the human user in shaping AI performance beyond the system's inherent capabilities.

### 3.4 Reflection Depth as a Key Driver of Writing Improvement in AI-Assisted Tasks

To examine how the depth of reflection (RDS) impacts learners' ability to improve their writing through AI-assisted tasks, we focused on participants with below-average performance in Method B (B-L group). As shown in Figure 4, learners with higher RDS scores (RDS-H group) consistently achieved greater gains in Methods C through E. In Method E, the RDS-H group significantly outperformed the RDS-L group, while improvements in C−B and E−B did not reach statistical significance. These results suggest that metacognitive reflection enables learners—even those with lower initial skills—to interpret AI feedback more critically, revise more effectively, and enhance the quality of their outputs.

Further evidence was observed within the A-L group, representing learners with initially low writing ability. While RDS-H learners improved moderately, RDS-L learners sometimes declined in quality due to insufficient reflection, leading to disorganized revisions (Figure 5).

These findings further suggest that learners with higher RDS not only revised more effectively but also appeared to possess stronger skills in selecting and integrating feedback. In contrast, those with lower RDS sometimes failed to benefit from revision activities, potentially due to difficulty in processing feedback. This indicates that structured reflection practices may help students better evaluate AI and peer suggestions, reinforcing the need to cultivate metacognitive skills in AI-supported writing instruction.

These results align with self-regulated learning theory, which highlights metacognitive strategies like reflection and self-monitoring in improving outcomes and revision [13, 14]. Learners with stronger metacognitive skills can better revise based on feedback. Therefore, beyond teaching AI tools as technical instruments, educational design should incorporate strategies to foster reflective learning—positioning generative AI not merely as a writing assistant but as a partner in cognitive development.

### 3.5 Summary and Implications of Findings

The effectiveness of generative AI in writing was mediated by learners' initial writing ability and reflection depth (RDS) rather than guaranteeing improvement on its own. Participants with higher RDS were more capable of evaluating and revising AI-generated text, often leveraging dialogic processes and peer feedback to produce more coherent and purposeful compositions. Conversely, learners with lower RDS scores occasionally experienced diminished quality due to superficial feedback integration, leading to structural inconsistency or weakened argumentation.

These findings collectively clarify how initial writing ability and reflection depth influence AI-supported writing outcomes across varying modes of AI involvement. While initial ability impacted early-stage performance, its influence diminished over time. In contrast, reflection depth (RDS) remained consistently associated with improved outcomes, particularly in revision-based or dialogic tasks. This highlights the need to incorporate reflective skill development into AI-supported writing pedagogy—not only to improve technical output but also to foster deeper learner engagement.

### 3.6 Limitations and Future Research

This study has several limitations. First, participants were limited to university students in the engineering field, which may affect the generalizability of the findings. Second, the writing task focused on a specific genre (admissions essays), and future studies should examine diverse writing contexts. Third, the research was conducted within a single session, preventing the investigation of longitudinal learning effects. Finally, the study did not control for the exact AI tools or prompts used, limiting detailed analysis of the influence of specific generative systems.

Future work should explore broader populations and develop targeted interventions to foster metacognitive skills, especially in the context of generative AI use. As metacognition has been shown to mediate learning outcomes and self-efficacy [11], instructional strategies that emphasize structured reflection may enhance students' ability to interact effectively with AI. Additionally, collecting detailed data on the AI tools (including version information) and the specific prompts used by learners will enable deeper analysis of the interaction between user input and AI output quality. Combined with advances in automated scoring and learning analytics, such research may contribute to building sustainable and scalable writing support systems in education.

## Acknowledgement

This work was supported in part by JSPS KAKENHI Grant Number JP24K06084. The authors would like to express their sincere appreciation to Dr. Takashi MIYAURA (Kyushu Institute of Technology) and Dr. Izumi TANAKA (Fukuoka Institute of Technology, Junior College)) for their valuable contributions to the development of the assessment rubrics and their insightful advice throughout the study. The authors also thank the students and faculty members who participated in the study for their cooperation. The authors used ChatGPT (OpenAI) for language correction and translation assistance. The authors made all final decisions regarding content and wording.

## References

[1] Y. Kimura, S. Fukushima, T. Aoki, Y. Shimada, H. Takashima, A. Fukui, R. Maruyama, & R. Yoshiwaki, (2023). *Artificial Intelligence Research Trends 2: Impacts of Foundation Models and Generative AI* (in Japanese). Center for Research and Development Strategy, Japan Science and Technology Agency (JST)

[2] Ministry of Education, Culture, Sports, Science and Technology – Japan. (2024). Guidelines for the Utilization of Generative AI at the Elementary and Secondary Education Level (Version 2.0). (in Japanese)

[3] Ritsumeikan University. (2023). Trial introduction of machine translation and ChatGPT-supported English class support system. Retrieved April 7, 2025, from https://www.ritsumei.ac.jp/news/detail/?id=3103

[4] Maki, D., Mitsumune, Y., Takahashi, Y., Mesaki, M., Muko, H., & Ohnishi, Y. (2024). A study on the educational use of generative AI through classroom practices in middle schools. *Bulletin of the Center for Research in Science Education, Faculty of Education, Ehime University*, 3. (in Japanese)

[5] Imai, M. (2024). Distribution Material 3 for the 10th Meeting of the Expert Panel on the Future of Curriculum, Instruction, and Assessment. Ministry of Education, Culture, Sports, Science and Technology. (in Japanese)

[6] Japan Association of National Universities. (2018). *Basic Policy on the University Admission System after 2020* (in Japanese)

[7] Ministry of Education, Culture, Sports, Science and Technology. (2015). *Implementation Plan for High School–University Connection Reform*. Retrieved April 7, 2025, from https://www.mext.go.jp/b_menu/shingi/chukyo/chukyo12/sonota/__icsFiles/afieldfile/2015/01/23/1354545.pdf

[8] Ministry of Education, Culture, Sports, Science and Technology – Higher Education Bureau, University Admissions Division. (2023). *Overview of the Implementation Status of University and Junior College Entrance Examinations for F Y2023*. (in Japanese)

[9] Matsue Higashi Senior High School, Shimane Prefecture. (2025). *2nd-Year Period for Inquiry-Based Cross-Disciplinary Study (February 2025)*. Retrieved April 7, 2025, from https://www.matsuehigashi.ed.jp/news/news-1/5148 (in Japanese)

[10] S. Kimura, T. Yasunaga, T. Miyaura, I. Tanaka. (2024). The Impact of Generative AI on Document Review of Statements of Purpose. *Proceedings of the 1st Annual Conference of Japanese Association for Research on University Admissions*;52-55. (in Japanese)

[11] Kember, David, Jones, Alice, Loke, Alice, McKay, Jan, Sinclair, Kit, Tse, Harrison, Webb, Celia, Wong, Frances, Wong, Marian, and Yeung, Ella. (1999). *Determining the level of reflective thinking from students' written journals using a coding scheme based on the work of Mezirow. International Journal of Lifelong Education 18*;18-30.

[12] The jamovi project. (2023). *jamovi (Version 2.4)* [Computer software]. Retrieved from https://www.jamovi.org

[13] Zimmerman, B. J. (2002). Becoming a Self-Regulated Learner: An Overview. *Theory Into Practice 41(2)*:64-70

[14] D. Akamatsu, M. Nakaya & R. Koizumi (2019). Effects of Metacognitive Strategies on the Self-Regulated Learning Process: The Mediating Effects of Self-Efficacy. *Behavioral Sciences 9 (12)*: doi:10.3390/bs912012