Majumdar, R. et al. (Eds.) (2025). Proceedings of the 1st International Conference on Learning Evidence and Analytics. Asia-Pacific Society for Computers in Education

Picture-Cued Writing Using AI-Generated Images for Language Acquisition

Ka-Lai WONG[a*], Yuko TOYOKAWA[b] , Yiling DAI[b] , Brendan FLANAGAN[b] & Hiroaki OGATA[b]

[a]*Graduate School of Informatics, Kyoto University, Japan*
[b]*Academic Center for Computing and Media Studies, Kyoto University, Japan*
*wong.lai.37t@st.kyoto-u.ac.jp

Abstract: The integration of AI-generated images into foreign language writing tasks has gained increasing attention in educational research. This study explores the impact of text-to-image generation models on students' perceptions, revision behavior, and writing performance. We conducted a two-week experiment in a Japanese high school, where 34 students participated in picture-cued writing tasks facilitated by an AI system. Upon submission, students received AI-generated images based on their writing and automated feedback on their text, allowing them to independently decide whether to revise or finalize their writing. Findings indicate that while a majority of students improved their writing proficiency, most did not actively revise their work, and the proportion of lower-level writing increased over time. Despite positive perceptions of enjoyment and usefulness, students expressed neutral intentions toward future use, citing slow computation speed as a limitation. This study provides preliminary evidence that AI-generated images can support foreign language writing. The findings highlight the potential of multimodal AI tools in fostering linguistic accuracy in second language writing.

Keywords: AI-generated images, picture-cued writing, foreign language learning, text-to-image models, automated feedback

1. Introduction

Visualization has long been a key strategy in language learning, with picture-cued writing shown to enhance learner engagement, creativity, and critical thinking. Advances in AI have enabled automated scoring of such tasks (Zhao et al., 2023), yet writing remains fundamentally about communication—a process not fully captured by numerical evaluation. Writers often lack feedback on how readers interpret their texts, limiting their ability to revise effectively. Concurrently, transformer-based text-to-image generation models have advanced significantly, enabling the creation of visually rich, AI-generated content. While aesthetically compelling, their pedagogical value remains underexplored. Specifically, how do learners respond to images generated from their own writing, and can such visuals enhance learning? To explore this, we developed a system that generates AI-based images and provides automated feedback based on student writing. This study investigates: 1. How do learners perceive the Usefulness and Enjoyment of the system? 2. What are their Attitudes and Intentions toward the writing task? 3. How do they revise their writing to better match the images?

We conducted a two-week study at a Japanese high school with 38 participants. Students engaged in picture-cued writing tasks, received AI-generated images and feedback, and revised their work accordingly. We analyzed changes in their writing, lexical use, and CEFR-J proficiency levels.

2. Related Works

Visual prompts enhance student engagement and performance in narrative writing. To enable objective evaluation, we adopt the six-component rubric from Zhao et al. (2023) while streamlining assessment through rule-based scoring and BLIP-2 image-text similarity metrics (Li et al., 2023). Chatbots, shown to support interactive learning and real-time feedback, are integrated into our experiment to provide hints that help students iteratively refine their writing. Recent advances in text-to-image generation have improved coherence and prompt relevance, offering new ways to link language to visual context. Unlike error-driven systems like L-VEIGe (Sugita et al., 2023), which generate images based on incorrect answers, our approach provides AI-generated visuals for open-ended tasks regardless of accuracy. We then assess semantic similarity between original and generated images to evaluate writing quality.

3. Experiment design

The study was conducted in two stages: material preparation and experiment implementation. During preparation, we used DALL·E 3 to generate images based on a high school English curriculum. These images were randomly grouped and scheduled for release, ensuring that students had access to a new writing prompt daily. Additionally, we designed four-frame comic strips as materials for the pretest and posttest. In the experiment stage, the picture-cued writing system was introduced as a warm-up activity in a Japanese first-year high school English class from January 22 to February 4, 2025. Students could complete tasks during class or voluntarily. Upon submission, the system used DALL·E 3 to generate an image and GPT-4o to provide an automated evaluation. Students could revise or begin a new task after reviewing their results. The teacher implemented the activity in four class sessions, with students working in small groups (2–3). Students also used the system independently. In total, 38 students submitted 120 writing samples across 10 image prompts. The pretest and posttest, each a five-minute writing task using the same comic strip, were completed via Google Forms. To ensure comparability, only data from students who completed both tests (n = 34) were analyzed, resulting in 34 pre/posttest samples and 109 in-task writings. Student perceptions were collected via a questionnaire with seven 5-point Likert items and two open-ended questions, following Kim et al. (2021). Writing quality was assessed using CEFR-J levels via the CVLA3 tool (Uchida & Neigishi, 2025).

4. Result and discussion

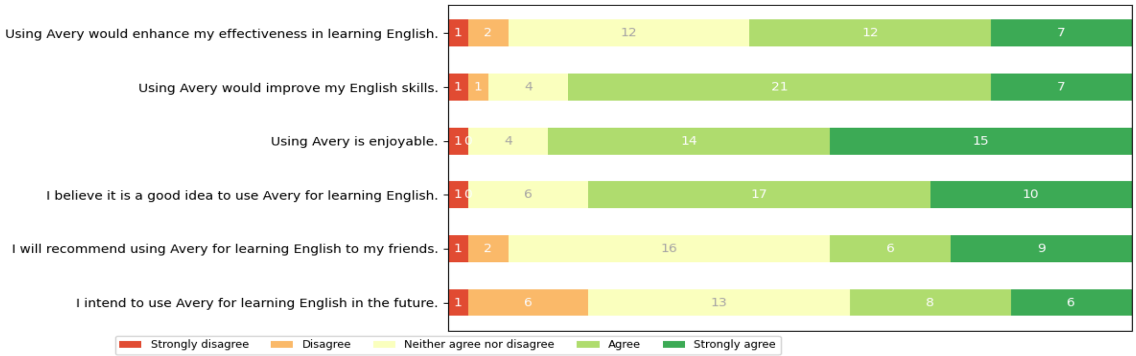

Figure 1. Questionnaire results

Student feedback was mostly positive (Figure 1), with high ratings for enjoyment and usefulness. However, intentions to continue using or recommend the system remained neutral, likely due to concerns about slow computation speed. In the open-ended responses, 25 out of 34 participants expressed satisfaction with the AI-generated images and automatic scoring system, while 20 out of 34 reported frustration with system responsiveness, emphasizing the need for optimization to improve user experience. Analysis of writing submissions revealed limited engagement with revision; most students did not revise their work after receiving AI feedback. CEFR-J analysis showed that 45.8% of texts were at pre-A1 and 19.2% at A1.1 levels. Interestingly, writings from voluntary sessions tended to score higher than those completed in class. However, the proportion of pre-A1 writing increased over time (Figure 2, right), possibly reflecting limited vocabulary or low engagement with the feedback process. An analysis of pretest and posttest writing using CVLA3 revealed that 55.9% of students achieved a higher CEFR score in the posttest than in the pretest, indicating an overall improvement in their writing proficiency. The mean word count in the pretest was 56 (SD = 20.20), with a mean CEFR score of 0.47 (SD = 0.73). In contrast, the posttest had a lower mean word count of 49 (SD = 14.80) but a higher mean CEFR score of 0.61 (SD = 0.74). On average, sentence length decreased by 6.82 words, while the CEFR score increased by 0.14. While students may have retained some memory of the pretest prompts, they tended to use fewer words with greater accuracy in the posttest, suggesting improved lexical precision and syntactic control.

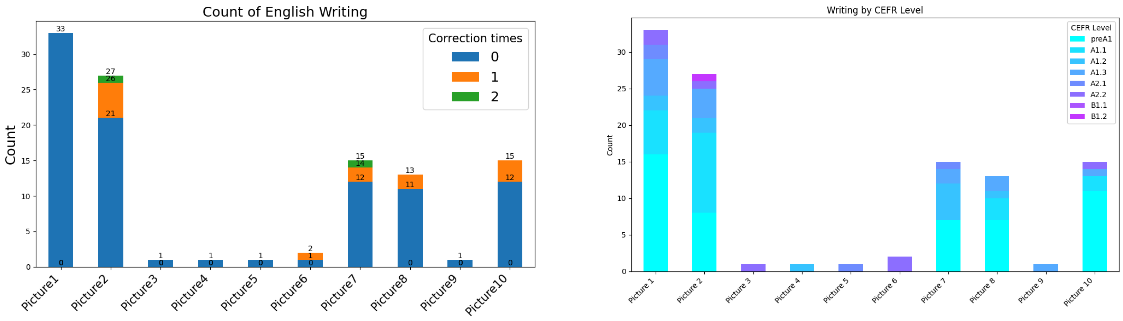

Figure 2. Writing samples Overview (left) by correction times; (right) by CEFR-J level

5. Conclusion

This study offers preliminary evidence that AI-generated images can support foreign language writing development. Improvements in students' writing performance suggest a potential learning effect. However, given the study's short duration and small sample size, future research should employ quasi-experimental designs with larger participant groups to more robustly assess the impact of AI-generated imagery on writing outcomes. Additionally, the integration of lightweight image generation models, such as Gemini 2.0 Flash, warrants exploration to improve system responsiveness.

Acknowledgements

This work was supported in part by the following grants: Japan Society for the Promotion of Science (JSPS) (B) JP20H01722, JP23H01001, (Exploratory) JP21K19824 and CSTI SIP Program JPJ012347.

References

Kim, H.-S., Cha, Y., & Kim, N. Y. (2021). Effects of AI chatbots on EFL students' communication skills. 영어학, 21, 712–734.

Li, J., Li, D., Savarese, S., & Hoi, S. (2023, July). Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In ICML (pp. 19730-19742). PMLR.

Sugita, K., Ota, K., Gu, W., & Hasegawa, S. (2023). L-VEIGe: Vocabulary Learning Support System Using Error Image Generation A study of criteria for image suggestibility. JSiSE Research Report, 38,n2. (2023-7)

Uchida S., & Neigishi, M. (2025) Estimating the CEFR-J level of English reading passages: Development and accuracy of CVLA3『英語コーパス研究』32.

Zhao, R., Zhuang, Y., Zou, D., Xie, Q., & Yu, P. L. (2023). AI-assisted automated scoring of picture-cued writing tasks for language assessment. Education and Information Technologies, 28(6), 7031-7063.