A Cynical Pilot Study on AI-Assisted University Student Writing Development

James T. Thomas
Central Washington University
Melissa Huntley
University of Shimane

Reference Data:

Thomas, J. T. & Huntley, M. (2026). A cynical pilot study on AI-assisted university student writing development. In B. Lacy, M. Swanson, & H. Yoder (Eds.), *LanguageS: Learning, Teaching, Assessing – JALT 50 Years – Challenges and Perspectives*. JALT. https://doi.org/10.37546/JALTPCP2025-05

AI tools are frequently presented as a means to improve EFL student writing; however, the extent to which AI genuinely enhances writing skills—rather than merely composing polished final products—remains an open question. This longitudinal experiment involved 93 first-year Japanese university students completing English writing tasks over one semester. Students were divided into two groups: the self-editing group (control), and the ChatGPT-assisted group (variable). Writing samples were assessed on length, relevance, linguistic diversity, and accuracy. Analysis revealed that ChatGPT-assisted editing does little to improve students' actual writing skills. While the final drafts of AI-assisted essays were relatively error-free, students' pre-edited texts showed no statistically significant differences in overall writing quality when compared to the control group. This suggests that AI-assisted editing polishes the final composition rather than developing skills, calling into question the prevailing narrative that AI tools enhance writing ability.

AIツールは英語を外国語として学ぶ学生のライティング能力向上手段として頻繁に紹介されるが、AIが単なる完成度の高い最終成果物の作成ではなく、真にライティングスキルを向上させる程度については明確ではない。本研究では、日本の大学1年生93名を対象に、1学期にわたり英作文の課題を実施した。学生は自己編集グループ（対照群）とChatGPTによる支援グループ（実験群）の2群に分けられた。作文は、長さ、内容の関連性、語彙の多様性、正確性に基づいて評価された。分析の結果、ChatGPTによる編集支援は学生の書く力の向上にはほとんど寄与しないことが明らかになった。AI支援による最終稿は比較的誤りが少なかったものの、編集前の文章は対照群とは統計的に有意な差が認められなかった。このことから、AIは作文スキルを育成するよりも、最終的な文章を磨き上げるだけのものであり、AIツールがライティング能力を高めるという一般的な見解に疑問を投げかける結果となった。

There has been a recent inundation of discussion about generative AI, particularly OpenAI's large language model (LLM) ChatGPT. Some hail it as a miracle technology that will lead humanity into a new "golden age," while others condemn it as a harbinger of the end times. This divided rhetoric is similarly reflected in the EFL community. Even before the release of ChatGPT, researchers like Birdsell (2022) were looking at the effects of advanced computing and neural networks on EFL writing. This attention has increased substantially with generative AI becoming widely available to the public. Furthermore, assertions about the importance or even necessity of generative AI in EFL have been increasing, with some researchers advocating for a cautious and awareness-driven approach (Dilekli & Boyraz, 2024; Hougham & Higginbotham, 2024; Matsuno, 2025) while others strongly advocate for broad adoption of generative AI as a tool in EFL education (Al-Khresheh, 2024; Solak, 2024; Tseng & Lin, 2024).

However, despite numerous academic studies (e.g., Rahmi et al., 2024, Tseng & Lin, 2024) and a plentitude of presentations at academic conferences emphasizing the benefits of generative AI incorporation into EFL writing courses, few longitudinal studies have been conducted involving large numbers of participants with balanced control-experimental groupings, which led this study's researchers to cynically ask, are the effects of generative AI being overexaggerated? This gap between perception and data is what the current study seeks to address, specifically seeking to answer the following question: Do students who use ChatGPT as a revision tool actually become better writers?

To explore this question, 93 Japanese university students in two first-year English courses were divided into control and experimental groups. Both groups composed five essays over the course of one semester. The experimental group was tasked with using ChatGPT as a revision tool while the control group was tasked with self-revising without AI assistance. Successive rough drafts were then assessed on metrics of length, relevance, linguistic diversity, and accuracy to measure ChatGPT's influence.

Literature Review

Although generative AI is an emerging technology, research concerning potential applications is plentiful, particularly in the field of education. Indeed, a plethora of recent papers on generative AI in EFL have been published, an indication of the impact this technology is having on the field. However, as Toncelli & Kostka (2024) acknowledged, "studies that explore how educators incorporate GenAI into their practice are scarce" (p. 78). Moreover, they noted that many teachers perceive strong pressures to incorporate generative AI into courses despite not fully understanding it themselves or without adequate training, leading to a broad range of implementation strategies.

Crompton & Burke (2024) conducted a detailed survey of existing literature on the use of ChatGPT in education, focusing on three facets: how teachers can utilize ChatGPT, how students can utilize ChatGPT, and what problems and potential for misuse may exist. They specifically addressed the topic of writing support, discussing multiple studies that demonstrated or proposed utilizing ChatGPT for a variety of writing-related tasks, ranging from error correction to guidance through the entire writing process. While potential benefits were discussed, they also addressed the increased ease of plagiarism. Other authors have also sought to educate educators about potential abuses of generative AI and how to guide students in its proper use. Raine (2023) demonstrated that prompts could be written to produce text mimicking typical EFL writing mistakes but speculated that such prompt writing was likely to be beyond the language capability of the students most likely to use it. Overall, he optimistically concluded that "[w]hile some students may be tempted to use these shortcuts in homework and coursework, others may be able to use them in ways that genuinely help them to improve their language proficiency" (p. 41). This cautious optimism is also reflected in Hougham & Higginbotham (2024), who suggested guiding students' use of ChatGPT as a peer reviewer, seeking to provide legitimatized, structured usage of a technology that students will inevitably begin using as well as proposing a set of "golden rules" for ethical use of AI generated content. Further examples of productively channeling generative AI include Dilekli & Boyraz (2024) and Matsuno (2025), who conducted awareness-raising activities with their students by having them actively reflect on their own output compared to that of generative AI.

Other authors have more emphatically advocated for the incorporation of generative AI into EFL classrooms, sometimes venturing into uses that many educators would consider problematic. Emblematic of the enthusiastic embrasure of generative AI in the classroom are studies such as Al-Khresheh (2024), Rahmi et al. (2024), and Tseng & Lin (2024).

Al-Khresheh (2024) addressed teacher perceptions regarding ChatGPT utilization in an EFL educational context, arguing that a lack of research into the perceptions of the broader international EFL educator community "poses a serious problem, as understanding these perspectives is essential for the meaningful and widespread adoption of ChatGPT in ELT" (p. 1). While Al-Khresheh (2024) enthusiastically advocated for AI integration into multiple aspects of EFL teaching, others endorsed it specifically for EFL writing contexts. In Rahmi et al. (2024), students utilized a GPT-derived AI writing assistant, *ParagraphAI*, to assist with their writing. The authors included diagrams and examples of the *ParagraphAI* interface and output as well as examples of student writing before and after using the AI writing assistant, stating that "[t]he tool ensures impeccable grammar, spelling, and vocabulary, generates plagiarism-free content, and helps overcome writer's block" (p. 1002). They measured lexical diversity for both the students' initial drafts and their AI-assisted final drafts, finding that AI-assisted writing produced statistically significant improvements in lexical diversity, thus concluding that "[t]exts produced with the help of the AI writing assistant appeared to have a significantly positive impact on learners' performance" (p. 1008). Similarly, Tseng & Lin (2024) tested ChatGPT specifically as a writing aid in a university writing class, providing students with detailed instructions regarding prompts for various writing tasks utilizing ChatGPT, including in the role of peer reviewer and as a structural guide, finding that "[s]tudents exhibited significant improvement in their writing quality, validating the effective integration of ChatGPT into writing instruction" (p. 95).

However, these studies rarely distinguished between student writing and AI output. Rahmi et al. (2024) described "multiple occasions on which they [students] failed to write in a 'correct' way but were miraculously able to produce texts with no mistakes with the help of an AI-generated writing assistant" (pp. 1004-1005) and that "students' original texts yielded results that were the opposite of the excellent quality texts produced under the AI-generated writing assistant" (p. 1005). When comparing the provided examples of a student's initial draft and the post-AI draft, however, it is evident that the AI writing assistant added ideas and examples were not present in the student's original version, which seems to undermine the authors' statement that "[t]he student's thoughts were expressed clearly and logically" (p. 1006). This attitude is similarly reflected in Tseng & Lin (2024), who concluded that "[s]tudents exhibited significant improvement in their writing quality, validating the effective integration of ChatGPT into writing instruction" (p. 95). However, their structured writing process and prompts often asked ChatGPT to do the majority of the writing, with students relegated to the role of accuracy checking and revision rather than composition. For instance, the authors gave this prompt to students:

> Write an autobiography with six paragraphs based on the experiences I had. The autobiography needs to be natural which means that others can't tell it's written by AI. Start with an attraction hook in the first paragraph. It should include my academic score, my research experiences, and my teaching experiences. The second paragraph focuses on my academic learning. How it benefits my future learning in XXX. The third paragraph focuses on my research experiences. The fourth paragraph focuses on my teaching experiences. The fifth paragraph focuses on describing my personality which includes self-disciplines, leadership, and being willing to help others. Show my strong potential of being a graduate student in XXX. Finally, close the autobiography impressively. (p. 89)

Interestingly, when the authors surveyed the students on their impressions of using ChatGPT at the conclusion of the study, it was the students themselves who articulated concerns regarding the ethics of using AI in this manner. Furthermore, if students are being instructed to utilize such AI-disguising prompts, this also reframes the concerns raised in Raine (2023) regarding manipulative prompts, as well as general concerns regarding ethics and plagiarism (Crompton & Burke, 2024; Hougham & Higginbotham, 2024; Solak, 2024; Toncelli & Kostka, 2024). However, as noted in Tseng & Lin (2024) and Solak (2024), students perceived their writing as better, and their ideas as being more fluently conveyed when using generative AI. Nor are students alone in treating their writing as better with AI, as Birdsell (2022) found that students not using digital assistance "may receive lower evaluations for their essays" from English teachers (p. 120) despite the teachers proving themselves demonstrably capable of discerning the generated nature of the text.

All of this highlights the importance of EFL educators collectively navigating questions regarding authorship, ethics, and definitions of plagiarism, but also of whether or not real language ability is being demonstrated when students utilize AI. While students can now produce flawless text quickly and easily, is this representative of their language ability or just of their ability to operate a tool without understanding its output? As Birdsell (2022) stated after having students experiment with neural machine translation (NMT) software (in this context a comparable form of AI), "if students simply cut and paste their writings into the NMT and then use the generated output, as is, very little learning occurs" (p. 121). It follows that using generative AI to produce writing would have similar results. Furthermore, although there are now a substantial number of published studies on the topic of generative AI use in EFL writing contexts, many are perception-based rather than analyses of student output (e.g., Al-Khresheh, 2024; Solak, 2024; Toncelli & Kostka, 2024). Additionally, few studies have been conducted with large sample sizes, both control and experimental groups, or multiple writing samples over an extended period. For instance, Birdsell (2022) analyzed 17 students divided evenly into experimental and control groups but who wrote a single essay; Dilekli & Boyraz (2024) focused on just five students; Gayed et al. (2022) analyzed ten EFL learners in a counter-balanced design; Matsuno (2025) had a sample size of 34 students but only one writing task; Rahmi et al. (2024) had a sample size of only four undergraduates; and Tseng & Lin (2024) had just 15 participants in their study. It is likely that the major contributing factor to the dearth of participants and samples within these studies is simply the newness of the technology. However, it is also true that with such small sample sizes, external validity becomes a concern when applying these findings to a larger student population or duration of study. Indeed, it is against this rising tide of optimism without substantive data that this study was conceived.

Whether it be checking grammar and spelling (Crompton & Burke, 2024), offering feedback while drafting or revising (Gayed et al., 2022; Matsuno, 2025; Rahmi et al., 2024), guiding students through the writing process step by step (Crompton & Burke, 2024), translating directly from L1 to L2 (Birdsell 2022), or producing entire texts based on prompts and personal detail input (Tseng & Lin 2024), three points are evident: 1) instructors are and will continue to be encouraging EFL students to use generative AI for writing; 2) the delineation between what constitutes genuine student work versus plagiarism is at issue; and 3) claims regarding generative AI improving students' writing ability are being propagated without strong data support. It is this third point that will be addressed in this study.

Methodology

The experimental design for this study was a four-month-long repeated measures design between a control group and an experimental group of Japanese university students in a first-year required English communication class. Participation in the research activities was part of class activities, but access to data for research and presentation purposes was given explicitly via written waiver provided in both English and Japanese. The waiver and each step of the experimental design were approved by the university's institutional review board and adhered to Japanese research compliance standards.

Participants were recruited from two sections of a first-year required English communication course, and writing tasks were conducted directly by the instructors of each section. This specific English course contained two cohorts of students: an A

group that participated in communication class with an instructor for 45 minutes, then attended sustained silent reading (SSR) for 45 minutes, and a B group that attended SSR, then took class with the instructor. The A group was treated as the control group (self-editing), and the B group was treated as the experimental group (AI-assisted editing).

Participants were tasked with five writing assignments throughout the first semester of the class (April to July) that involved three steps: rough draft, editing process, and resubmission. All participants composed rough drafts in the classroom under instructor supervision via the institutional LMS for Step 1, which were collected for quality analysis. The task was a 15-minute essay on a given topic (Task 1: self-introduction, Task 2: discuss a hobby, Task 3: describe a place you know well, Task 4: summer vacation plans, Task 5: reflect on your first semester of university). In Step 2, participants in both groups were given an editing checklist (e.g., "Did I make any spelling errors?" "Do my ideas connect clearly and logically from one to the next?" "Can I replace any simple words with stronger or more specific words?"). In instructions for how to use the check list, the control group were told to revise their rough drafts using the checklist. The experimental group were instructed to use the checklist as a guide for asking Open AI's large language model (LLM) ChatGPT to provide feedback on their essays and to revise based on feedback received. Step 3 consisted of participants submitting their revised drafts as proof of successful completion of the research task.

Writing quality was assessed for only the first drafts of each writing assignment. By comparing unassisted and unedited writing samples between the control and experimental groups throughout the course of the research, participant writing ability could be isolated and measured without the confounding variable of AI-produced writing obscuring authentic participant-produced compositions. A researcher-designed formula combining qualitative and quantitative measures of writing quality was created by coding composition length, lexical diversity, sentence-level relevance, and language accuracy to create a simple 0 – 4 scale of assessed quality (see Table 1).

Table 1

Components of assessment scale with example

Scale components	1. Word Count Ratio	2. Type / Token Ratio	3. Relevancy Rate	4. Accuracy Rate	5. Total Quality Score
	1 – (1/2 average baseline word count ÷ word count)	Unique words ÷ total word count	Relevant sentences ÷ total sentence count	1 – (number of errors ÷ total word count)	Sum of scores 1 - 4
Example TA003.1	1 - (1/2 *87.95 [baseline average] ÷ 151 words)	96 unique words ÷ 151 words	24 relevant sentences ÷ 24 total sentences	1 - (25 errors ÷ 151 words)	
	= **.708**	= **.636**	= **1**	= **.8344**	**3.179**

Length was quantified by first establishing a baseline by averaging the word count for the first writing task (87.95 words), then dividing the word count total of a submission by the baseline. For example, a submission of 151 words would result in a length score of .708. Lexical diversity was measured via type/token ratio where the number of unique words was divided by the total word count of a submission. These two quantitative assessment items were conducted via the *TextInspector* software (Alshehri, 2022; Gayed et al., 2022; Raufi et al., 2024). Instructors coded the qualitative relevance and accuracy measures. Relevancy scores were determined by dividing the number of sentences relevant to the given topic by the total number of sentences. Conversational phrases such as "Hello" and "Thank you for reading" were determined irrelevant by this measure. Accuracy was measured by reverse coding the error rate of five categories of errors (subject-verb agreement, determiner/plural usage, verb tense/form, preposition usage, and word choice). Each component score was combined to create a total quality score for the submitted composition. Total quality scores and individual component scores were compared between each of the five writing sessions to measure longitudinal growth in writing ability and between groups to identify discrepancies in ability growth.

Alteration to Experimental Design

Open AI's LLM was chosen for its accessibility to a large number of participants as a free version of the model could be accessed via browser or phone application without requiring an account. For the timeframe of this research, participants in the experimental group used the model GPT-4o. However, in the first research task, it was found that participants attempting to access ChatGPT on a mass scale on the same network were unable to engage with the model due to limitations of the free version. As a result, the design was adjusted so that participants were independently engaged in the editing and resubmission process outside of the classroom.

Results

Descriptive Statistics

Of the 93 students who participated in this research, 91 sets of writing data were collected, with 49 sets of data in the control group (unassisted self-editing) and 42 sets of data in the experimental group (ChatGPT-assisted editing). Initial analysis on the average total quality score and individual quality components revealed that participants showed a steady improvement from the first writing task through the fourth writing task, with a decrease in the fifth writing task (see Figure 1 and Tables 2 & 3). Researchers concluded that the topic of the final writing task was a more difficult topic than previous assessments and removed the Task 5 data set from inferential analysis.

Figure 1

Average Total Quality Scores from Task 1 to 5

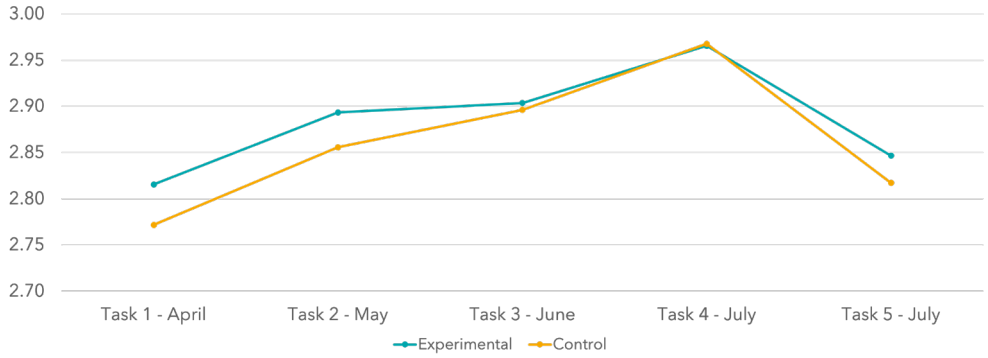

Table 2

Average Quality Component and Total Scores from Tasks 1 to 5: Control Group

	Task 1	Task 2	Task 3	Task 4	Task 5
Word count ratio	.464	.435	.463	.526	.462
Type token ratio	.677	.644	.652	.596	.653
Relevance ratio	.736	.920	.934	.989	.880
Accuracy ratio	.895	.857	.848	.8568	.822
Total Score	2.772	2.856	2.896	2.968	2.817

Table 3

Average Quality Component and Total Scores from Tasks 1 to 5: Experimental Group

	Task 1	Task 2	Task 3	Task 4	Task 5
Word count ratio	.460	.469	.462	.547	.512
Type token ratio	.684	.631	.642	.590	.628
Relevance ratio	.787	.927	.930	.941	.908
Accuracy ratio	.884	.865	.869	.900	.799
Total Score	2.816	2.894	2.904	2.966	2.847

Inferential Statistics

A paired sample two-tailed *t*-test was conducted to compare total quality scores and component scores between the control and the experimental groups from the Task 1 data sets and Task 4 data sets (see Tables 4 & 5). Total quality scores and all assessment components were statistically significant from Task 1 to Task 4 with the exception of the accuracy rate of the experimental group.

Table 4

Means for Total and Component Scores: Control Group

	μ	*σ*	*df*	*t*	*p*
Word count ratio	.063	.014	92	2.41	.018
Type token ratio	-.078	.004	92	-5.68	<.001
Relevance ratio	.260	.003	92	6.589	<.001
Accuracy ratio	-.037	.003	92	-3.197	.002
Total Score	.206	.015	90	4.449	<.001

Table 5

Means for Total and Component Scores: Experimental Group

	μ	*σ*	*df*	*t*	*p*
Word count ratio	.08	.019	74	2.540	.013
Type token ratio	-.098	.006	75	-5.616	<.001
Relevance ratio	.176	.005	75	5.024	<.001
Accuracy ratio	.002	.003	74	.188	.425
Total Score	.176	.018	74	4.06	<.001

There were statistically significant differences in the relevancy and accuracy scores between the control and experimental groups (see Table 6), with the control group showing a higher rate of growth in relevancy ratios and the experimental group displaying a higher rate of growth in accuracy ratios, but no statistically significant differences in total quality score were found between the control group (μ =.21 =, σ = .069) and the experimental group (μ = .17, σ = .06); t (84) = 0.713, p = .239.

Table 6

Means for Total and Component Scores: Between Groups

	μ (control - exp)	*df*	*t*	*p*
Word count ratio	-.025	83	-.887	.189
Type token ratio	.018	83	1.12	.133
Relevance ratio	.554	83	10.27	<.001
Accuracy ratio	-.039	83	-2.851	.006
Total Score	.039	83	.7134	.238

Discussion

The overall results showed little difference between the control and experimental groups. Instead, both groups showed a general increase in quality regardless of utilizing AI-assisted revision which is consistent with their receiving otherwise equivalent English instruction over the course of the semester.

However, there were two minor but statistically significant variations between the groups that merit discussion. The first significant difference was the higher increase in relevancy rate in the control group. It may be that the experimental group expected ChatGPT to account for problems and were therefore less proactive about actively applying instructor feedback while the control group was more inclined to study feedback from previous writing tasks.

The second significant difference was the experimental group showing greater improvement in accuracy than the control group. There are two theories that may account for this improvement. The first is that students utilizing ChatGPT improved their accuracy because using ChatGPT did teach them, consciously or unconsciously, to write slightly more accurately. The second theory has to do with priming. The control group wrote their drafts in their first English instruction environment of the day. Conversely, the experimental group spent 45 minutes in English SSR before writing their English essays. It is therefore possible that they were primed before writing. Overall, however, the growth shown by both groups was mostly consistent.

Finally, there was one unexpected consequence of introducing the experimental group to generative AI, which was that the experimental group submitted an increased number of plagiarized works outside of the parameters of the experiment, both in other class

assignments and in the rough draft process of this research (those drafts were removed from the data pool).

Limitations

Due to the pilot nature of this study, there were several limiting factors that could be addressed in future research. First, although students in the control group were asked not to use AI in their revisions, because the limitations in the free version of ChatGPT resulted in students revising outside of a controlled environment, it was impossible to guarantee that they did so unaided. Furthermore, it was impossible to guarantee that students in either group were allotting adequate time for methodical revisions. Second, the initial datasets were limited to five writing samples over a single semester (continued research is ongoing). Third, as demonstrated with writing task 5, prompts could have been more tightly scaffolded to avoid variable complexity issues. Fourth, students were not provided with detailed instructions on how to utilize ChatGPT as a learning aid. This was a conscious decision by the researchers, as it reflects the real-world circumstances in the broader education field regarding the range of AI familiarity amongst teachers and students (Al-Khresheh, 2024; Dilekli & Boyraz, 2024; Hougham & Higginbotham, 2024; Toncelli & Kostka, 2024). However, longitudinal research addressing both substantive ChatGPT instruction and general writing instruction together could yield different results. Fifth, the metrics used to assess student writing were limited due to researcher time constraints: Many types of grammatical, semantic, and structural errors were not included in the counts, and nor were all possible positive writing metrics measured. Finally, as this research is ongoing, the priming theory discussed above will be tested, for the two groups will alternate timeslots in the second semester, meaning that the control group will be attending SSR first and thus subject to the same theorized priming effect.

Conclusion

The goal of this study was to answer the question of whether students who use ChatGPT as a revision tool actually become better writers. Overall, this study found that while using ChatGPT as a revision tool helped students produce better final drafts, it did not improve their actual L2 writing ability. Students in both the experimental and control groups saw similar writing skill growth over time. However, apart from a minor increase in accuracy by the experimental group (which may be attributable to an outside priming factor), the experimental group did not demonstrate a statistically significant increase in writing ability when compared to the control group. This fits with the results (if not always the conclusions) of previous research on AI in EFL writing courses (Birdsell, 2022; Gayed et al., 2022; Rahmi et al., 2024; Tseng & Lin, 2024).

It is obvious that ChatGPT can generate text exceeding the writing quality of an EFL learner. But caution should be exercised to avoid falsely equating the output of generative AI with student language level. Furthermore, institutions, instructors, and learners may perceive the integration of generative AI as inevitable, but that does not mean that it should be used as a substitute for proper instruction. It may be true that the EFL community is at a crossroads. The question students and educators need to ask is this: Is the goal of EFL education genuine skill development, or does the illusion of it suffice? While it may not be possible to "put the genie back in the bottle," that does not mean that generative AI should be enthusiastically incorporated into every aspect of learning. Instead, a mindful, scaffolded integration of generative AI should be used supplementarily if at all. Nor does it mean that the most optimistic views on generative AI should be embraced unquestioningly. A healthy degree of cynicism is warranted.

Bio Data

Melissa Huntley Omuro has worked at the University of Shimane since 2016 and specializes in cross-cultural communication and psychology. In addition to examining AI use in the classroom, she presents and publishes regularly on communicative competence in intercultural exchanges, Japanese culture and religions, and virtual exchange as part of Kakenhi Grant 22K00795.
<m-huntley@u-shimane.ac.jp>

James Thomas holds an M.A. in TESOL and is the current Math & Writing Tutoring Manager at Central Washington University. He previously served as an English lecturer at the University of Shimane for five years and taught in the Preparatory English Program at King Fahd University of Petroleum and Minerals in Saudi Arabia for two years. His research interests include rhetoric, pedagogy, and language learning.
<jtylerthomas@gmail.com>

References

Al-Khresheh, M. H. (2024). Bridging technology and pedagogy from a global lens: Teachers' perspectives on integrating ChatGPT in English language teaching. *Computers and Education: Artificial Intelligence, 6*, 100218. https://doi.org/10.1016/j.caeai.2024.100218.

Alshehri, A. (2022). Writing proficiency of Saudi EFL learners: Examining the impact of lexical diversity. *Journal of Language and Linguistic Studies, 18* (1), 519–529.

Birdsell, B. J. (2022). Student writings with DeepL: Teacher evaluations and implications for teaching. In P. Ferguson, & R. Derrah (Eds.), *Reflections and New Perspectives*. JALT. https://doi.org/10.37546/JALTPCP2021-14

Crompton, H., & Burke, D. (2024). The educational affordances and challenges of ChatGPT: State of the field. *TechTrends 68*(2), 380–392. https://doi.org/10.1007/s11528-024-00939-0

Dilekli, Y., & Boyraz, S. (2024). From "Can AI think?" to "Can AI help thinking deeper?": Is use of Chat GPT in higher education a tool of transformation or fraud?. *International Journal of Modern Education Studies, 8*(1), 49–71. https://doi.org/10.51383/ijonmes.2024.316

Gayed, J. M., Carlon, M. K. J., Oriola, A. M., & Cross, J. S. (2022). Exploring an AI-based writing assistant's impact on English language learners. *Computers and Education: Artificial Intelligence*, *3*, 100055. https://doi.org/10.1016/j.caeai.2022.100055

Hougham, D., & Higginbotham, G. (2024). Generative AI in writing classes: Seven golden rules. *The Language Teacher, 48*(2), 14–17. https://doi.org/10.37546/JALTTLT48.2

Matsuno, S. (2025). How ChatGPT Improves Writing for EFL Students. *JACET* 授業学ジャーナル, *5*, 56–66.

Rahmi, R., Amalina, Z., Andriansyah, A., & Rodgers, A. (2024). Does it really help? Exploring the impact of Al-generated writing assistant on the students' English writing. *Studies in English Language and Education*, *11*(2), 998–1012. https://doi.org/10.24815/siele.v11i2.35875

Raine, P. (2023). ChatGPT: Initial implications for language teaching and learning. *The Language Teacher*, *47*(2), 38–41. https://doi.org/10.37546/JALTTLT47.2

Raufi, B. S., Mulyono, H., Ilyas, H. P., & Zulaiha, S. (2024). Exploring Indonesian EFL students' lexical diversity and its correlation with academic vocabulary use in an online academic writing environment. *Ampersand 13,* 100196. https://doi.org/10.1016/j.amper.2024.100196

Solak, E. (2024). Revolutionizing language learning: How ChatGPT and AI are changing the way we learn languages. *International Journal of Technology in Education (IJTE), 7*(2), 353–372. https://doi.org/10.46328/ijte.732

Toncelli, R., & Kostka, I. (2024). A love-hate relationship: Exploring faculty attitudes towards GenAI and its integration into teaching. *International Journal of TESOL Studies, 6*(3), 77–94. https://doi.org/10.58304/ijts.20240306

Tseng, Y, & Lin, Y. (2024). Enhancing English as a foreign language (EFL) learners' writing with ChatGPT: A university-level course design. *The Electronic Journal of e-Learning*, *22*(2), 78–97. https://doi.org/10.34190/ejel.21.5.3329