International Journal of Educational Technology in Higher Education

RESEARCH ARTICLE **Open Access**

AI-generated feedback on writing: insights into efficacy and ENL student preference

Juan Escalante[1*], Austin Pack[1] and Alex Barrett[2]

*Correspondence:
Juan.escalante@byuh.edu

[1] Faculty of Education and Social Work, Brigham Young University-Hawaii, 55-220 Kulanui Street, Laie, HI 96762, USA
[2] College of Education, Florida State University, Stone Building, 114 West Call Street, Tallahassee, FL 32306, USA

Abstract

The question of how generative AI tools, such as large language models and chatbots, can be leveraged ethically and effectively in education is ongoing. Given the critical role that writing plays in learning and assessment within educational institutions, it is of growing importance for educators to make thoughtful and informed decisions as to how and in what capacity generative AI tools should be leveraged to assist in the development of students' writing skills. This paper reports on two longitudinal studies. Study 1 examined learning outcomes of 48 university English as a new language (ENL) learners in a six-week long repeated measures quasi experimental design where the experimental group received writing feedback generated from ChatGPT (GPT-4) and the control group received feedback from their human tutor. Study 2 analyzed the perceptions of a different group of 43 ENLs who received feedback from both ChatGPT and their tutor. Results of study 1 showed no difference in learning outcomes between the two groups. Study 2 results revealed a near even split in preference for AI-generated or human-generated feedback, with clear advantages to both forms of feedback apparent from the data. The main implication of these studies is that the use of AI-generated feedback can likely be incorporated into ENL essay evaluation without affecting learning outcomes, although we recommend a blended approach that utilizes the strengths of both forms of feedback. The main contribution of this paper is in addressing generative AI as an automatic essay evaluator while incorporating learner perspectives.

Keywords: Automated writing evaluation, ChatGPT, Artificial intelligence, Language education

Introduction

Automated writing evaluation (AWE) systems such as *Grammarly* and *Pigai* assist learners and educators in the writing process by providing corrective feedback on learner writing. These systems, and older tools such as spelling and grammar checkers, rely on natural language processing to identify errors and infelicities in writing and suggest improvements. However, with the recent unleashing of highly sophisticated generative pretrained transformer (GPT) large language models (LLMs), such as GPT-4 by *OpenAI* and *PaLM 2* by *Google*, AWE may be entering a new era.

As Godwin-Jones (2022) pointed out in his treatise on AWE tools in second language writing, GPT-powered programs are capable of not only correcting errors in essays, but can also compose essays. Given a simple prompt, generative artificial intelligence (GenAI) LLMs and chatbots that allow users to interface with LLMs, such as *ChatGPT* and *Bard*, can produce complete essays that are passable at the university level (Abd-Elaal et al., 2022; Herbold et al., 2023). It is also possible for English as a new language (ENL) writers to use GPT-powered machine translation to turn their essays written in their first language (L1) into an English essay (Godwin-Jones, 2022) take problematic writing and correct any mistakes wholesale, change its tone from informal to academic, or add cohesive elements like discourse markers (Tate et al., 2023). Educators have begun to use AI-powered plagiarism detectors to identify student submissions that were generated by AI, yet AI paraphrasing programs like *Quillbot* have been found to render AI-generated text indetectable by such tools (Krishna et al., 2023). With millions of users engaging with *ChatGPT* and other GenAI tools since *ChatGPT*'s debut in November of 2022, public discourse has speculated on the disruptive and problematic nature of these tools for the field of education (Lampropoulos et al., 2023).

The public reaction to GenAI in education has been diverse. In Fütterer et al.'s (2023) systematic review of popular publications across Australia, New Zealand, The U.K., and the U.S., general sentiment appeared evenly split between positive and negative, but concerns about academic integrity have been raised (Sullivan, 2023), with some educational institutions deciding to ban *ChatGPT* than to allow its use (Yang, 2023). The disruption GenAI represents for language education has been likened to the pocket calculator's impact on math education (Urlaub & Dessein, 2022), when institutions debated between prohibiting the technology or incorporating it by rethinking the educational objectives of math education. The prevailing sentiment on GenAI seems to be that reforms are needed to adapt educational practices in accommodation of the technology (Fütterer et al., 2023; Tseng & Warschauer, 2023). However, research is urgently needed so that teachers, students, and instructional designers can appropriately apply GenAI in education (Chiu et al., 2023).

This article represents a step in the direction of better understanding how GenAI might be used in language learning classrooms by examining how language teachers and learners employ it in the writing process. Specifically, we will attempt to investigate the efficacy of using GPT-4 as an AWE tool for generating corrective feedback on student writing and whether students will prefer this feedback over that of a human tutor.

Overview of relevant literature

ChatGPT is a public-facing GenAI chatbot that allows users to interface with LLMs. GenAI chatbots have been trained on a large corpus of language from the Internet to statistically predict the next most probable word in response to a user prompt; these responses are then put through an algorithm of reinforcement learning (OpenAI, 2023a). From this relatively simple premise these tools can generate, synthesize, or modify natural language to a high degree of sophistication (Elkins & Chun, 2020), and are rapidly becoming more sophisticated (Baktash & Dawodi, 2023). GenAI has proven capable at a variety of tasks including writing essays or creative texts such as poems or stories, writing or correcting computer programming code, answering questions, summarizing and paraphrasing provided text, and synthesizing disparate tones and styles to generate new and creative text. The vast capabilities and ease of use of GenAI chatbots have led to widespread concerns of the misuse of these tools by students (Yeo, 2023).

Educational systems currently rely on student formative and summative writing in assessment and instruction to develop and assess critical thinking, argumentation, synthesis of information, knowledge and competence, and language proficiency (Behizadeh & Engelhard, 2011); but the benefits of writing extend in other ways, such as learning about oneself, participating in a community, or simply to occupy free time (Florio & Clark, 1982). With writing being a beneficial and critical component of many educational systems, the task of reforming these systems to accommodate GenAI authoring apps seems both daunting and unappealing. Yet the historical lesson of pocket calculators shows that it is equally unappealing to prohibit the technology, or even ignore it (Urlaub & Dessein, 2022).

Godwin-Jones (2022) called for the "thoughtful, informed differentiation in the use and the advocacy of AI-enabled tools, based on situated practice, established goals, and desired outcomes" (p. 13). To address this involved agenda researchers and practitioners need to re-examine educational objectives which, for ENL writing instruction, includes identifying the purpose of writing in the curricula. Writing for placement or other summative writing will have a different objective than process-oriented writing, for instance. How AI-enabled tools can be integrated with these objectives remains unclear.

From a foundation of second language acquisition principles, [Ingley, 2023] proposed several practical ways in which GenAI might be used to improve academic writing in ENL contexts. For example, they propose questioning AI-enabled chatbots, and reflecting on output as a way of generating ideas or better understanding a topic versus simply asking the AI to brainstorm a topic for you. They also suggest that AI can help by serving in specific roles (e.g., a conference proposal reviewer, a writing teacher in a writing conference) and organize writing by drafting outlines or by providing feedback on a draft's organization. Similarly, they propose that feedback on coherence, grammar, vocabulary, and tone can be asked of these AI-tools to help support formative essay writing. Through purposeful prompting, AI-enabled chatbots can act as a more knowledgeable other (John-Steiner & Mahn, 1996) that can provide comprehensible input (Krashen, 1982) along different stages of the writing process.

Suggestions such as these illustrate how instructors might frame acceptable and unacceptable use of AI-enabled writing tools by learners. Working in concert with AI in a creative and iterative process positions the learner as the driver in the writing process, as opposed to the learner prompting the AI to do all the thinking for them. Identifying and communicating the ethical and appropriate use of AI is an urgent task for practitioners. Since learners have increasingly relied on forms of AI in the writing process for decades, from the red or blue squiggly lines under text in word processors to recommendations on usage and style from *Grammarly*, they may not question using more enveloping forms of writing assistance.

From the perspective of learners, the use of AI by teachers and institutions may also need to be negotiated in terms of what is appropriate and ethical. Major exams such as the GRE and TOEFL often rely on AI-enabled AWE programs to score large numbers of essays (Elliot & Klobucar, 2013), as algorithmic assessment of writing reduces bias and noise and is likely more consistently accurate than the judgments of human experts

(Grove et al., 2000). But with easily accessed AWE tools like *Grammarly*, and GenAI tools like *ChatGPT*, it is simple for any teacher to offload the responsibility of essay evaluation to automated processes (Kumar, 2023). Personalized learning through evaluating and giving feedback on essay writing has been identified as a potential strength of GenAI (Chiu et al., 2023; Farrokhnia et al., 2023; Zhu et al., 2023), which can, in turn, help decrease teacher workload (Farrokhnia et al., 2023) and prevent teacher burnout. However, teachers will need to make informed decisions regarding if and when to incorporate AWE by consulting learner perceptions and considering the benefit to learning.

Although *Grammarly* has been shown to be useful as an AWE tool (Fitria, 2021), it is not yet known whether *ChatGPT* and similar GenAI tools can effectively or reliably be used for this purpose, nor whether learners will accept feedback from these tools. Programs like *Grammarly* and *Pigai* are specifically designed for essay evaluation and scoring using latent semantic analysis, a modeling approach that relies on large corpora of essays to determine whether a student's writing is statistically similar to writing in that corpora in terms of both mechanics and semantics (Shermis et al., 2013). The LLMs that *ChatGPT* interfaces with, on the other hand, are not trained with a corpora from a specific domain, such as essays, but with text scraped from the Internet. The domain-general nature of the LLMs behind *ChatGPT* means its efficacy as an AWE tool needs to be researched before being used as such.

In a recent feasibility study, Dai et al. (2023) used *ChatGPT* to provide corrective feedback in undergraduate writing. They found the GenAI feedback to be more readable and detailed than instructor feedback, but still maintained high agreement levels with instructor feedback on certain (but not all) aspects of student writing. Another study that examined ChatGPT for essay evaluation and feedback by Mizumoto and Eguchi (2023) fed a corpus of 12,100 essays by non-native English writers to *ChatGPT* and compared rubric-grounded feedback and scores to benchmark levels. Their results showed that *ChatGPT* was reasonably reliable and accurate. These studies suggest the feasibility and reliability of using GenAI tools like *ChatGPT* for the purpose of AWE, however the efficacy and student perceptions of *ChatGPT* AWE use needs to be better understood.

GenAI has many known and unknown limitations which need to be considered before using it as an AWE tool. One of the limitations identified by OpenAI itself is the tendency for *ChatGPT* to produce text that is untruthful and even malicious (OpenAI, 2023a). *ChatGPT* does not function as an information retrieving program in the way that internet search engines do, for example, and only produces text that is tailored to the user prompt using the statistically best-fitting combination of words. Considering students' tendency to accept information from AWE tools without verifying it (Koltovskaia, 2020), this suggests a need to teach learners to approach GenAI-produced output critically. Other relevant concerns about the use of GenAI are bias in output and privacy of user data (Derner & Batistič, 2023). Although OpenAI is working on solutions to these issues (OpenAI, 2023a), safeguards are still vulnerable to certain prompting practices (Derner & Batistič, 2023).

The accuracy and efficacy of GenAI chatbots relies to some extent on prompt engineering (Strobelt et al., 2023), as well as which LLM is used (e.g., BERT, GPT-4). According to Zhou et al. (2023), prompt engineering is the practice of optimizing the language of a prompt with the intention of eliciting the best possible performance from LLMs.

With prompt engineering, users can guide *ChatGPT* to desired behaviors by specifying things like task, context, outcome, length, format, and style.

Prompting for optimal AWE application is not yet fully explored. Mizumoto and Eguchi (2023) used a zero-shot prompting method where scoring samples were not included. Their scoring rubric was inputted in plain text format and they inserted all of their essays (n = 12,100) using a for loop in Python. Dai et al. (2023) used multi-turn prompting and pasted each essay (n = 103) at the end of each prompt. It may be within the capabilities of *ChatGPT* to therefore act as an AWE tool and, provided a scoring rubric or other criteria, furnish corrective feedback on student writing without fine-tuning. However, prompt engineering is a nascent science and it is not yet known whether such a practice would produce corrective feedback reliably from LLMs.

Ultimately, the success of any learning technology depends on whether users adopt it. According to Davis's (1989) seminal paper, the primary influences on user adoption of technology are in its perceived ease of use and perceived usefulness. Huawei and Aryadoust's (2023) review of AWE literature revealed that several studies reported that students and teachers viewed AWE scores negatively compared to scores provided by human raters. However, studies with ENL populations have noted that students often find human feedback to be confusing (Weigle, 2013). One advantage of LLMs is the ability to tailor the output by, for example, asking the chatbot to reiterate feedback in easier to understand terms or to explain things further. Roscoe et al.'s (2017) study of student perceptions of AWE described students' attitudes toward the system as "cautiously positive" (p. 212) which was influenced by presenting the AWE system as helpful, student initial expectations of the system, and student direct experience with the system feedback. Given the prominence that *ChatGPT* has garnered, its reputation will likely precede itself in the classroom and students and teachers may not initially trust it as an AWE tool, but whether students and teachers perceive the feedback from *ChatGPT* as being useful has not yet been studied.

The present study

ChatGPT represents a new technology with a vast array of capabilities in natural language processing. Educators are rushing to understand how this technology can appropriately be incorporated into classrooms, which has inspired new research agendas investigating its limitations and affordances. One avenue of research that is needed is understanding how GenAI can be included in the writing process in a way that is acceptable to both students and teachers. This study intends to examine the efficacy of *ChatGPT* and *GPT-4* as an AWE tool in terms of language improvement and student perceptions in the ENL population. Specifically, this study will be guided by the following research questions.

1. Does the application of AI-generated feedback result in superior linguistic progress among ENL students compared to those who receive feedback from a human tutor?
2. Does the preference for AI-generated feedback surpass that for human tutor-generated feedback among ENL students?

Method

Study 1 explored the first research question by means of a six-week longitudinal mixed repeated measures quasi experimental design. Study 2 investigated the second research question through a weekly survey administered over six weeks.

Participants

Both studies were conducted at a small liberal arts university in the Asia–Pacific region during the shortened Spring 2023 semester. A non-probability self-selection method was used to recruit 91 participants who were ENL students enrolled in an academic reading and writing language course. Based on the institution's English Language Admission Test, students were assessed as having at least a Common European Framework of Reference (CEFR) B1 English proficiency level. Standard ethical procedures were followed, with participants voluntarily consenting to participate. As part of the communication with the participants, it was explicitly stated that their involvement in the study would not be rewarded with additional academic credit.

The first research question was explored in study 1 with a total of 48 ENL students, 21 males and 27 females, ranging in age from 20 to 30 years old. They were divided into two groups: a control group (CG) that received feedback on their assignments from a human tutor, and an experimental group (EG) that was given feedback generated by AI (GPT-4).

To address the second research question, in study 2 a separate group of 43 ENL students, composed of 13 males, 30 females, whose ages ranged from 19 to 36 years, received written feedback on their weekly assignments from both AI and human tutors. Complete questionnaire responses varied among participants across the six weeks, from 32 to 41 with an average of 37.7.

Instruments

In study 1, to gauge the linguistic progress among students, the study implemented a pre-and post-test design. A diagnostic writing test administered on the first day of class served as the pretest and a final writing exam served as the posttest. For these assessments, and for a recurring weekly writing task, participants were required to write a 300-word paragraph centered around diverse academic topics discussed in class and integrate sources from readings.

In study 2, to assess student preferences between human and AI-generated feedback, a questionnaire was developed to gather quantitative and qualitative data (see appendix A). Eight five-point Likert scale items, arranged into four pairs, captured various dimensions of feedback preference, including, satisfaction, comprehensibility and clarity, helpfulness, and overall preference (e.g.,: "I am satisfied with the feedback I received from the ENL tutor this week compared with "I am satisfied with the feedback I received from the AI program this week"). Participants were asked to respond using a scale from "1—Strongly Disagree" to "5—Strongly Agree." In addition, participants were asked "If you were to only get one kind of feedback next week, which kind of feedback would you prefer?". A follow-up open-ended question in which participants were asked to provide an explanation for their choice was also included.

OpenAI's GPT-4 was utilized to generate feedback on student writing for both studies. GPT-4 is a multimodal LLM that can process image and text inputs and produce text outputs. GPT-4 was selected as we found it to provide the most suitable and accurate feedback out of the LLMs we tested. Furthermore, GPT-4 outperformed other LLMs on academic benchmarks, at least at the time of its release (OpenAI, 2023b) and when this study was conducted.

The prompt sent to ChatGPT to generate feedback on students' weekly writing consisted of several parts. Two experienced language educators familiar with the course and its assignments and assessments developed the prompt iteratively in line with the prompt-engineering framework presented by [Ingley, 2023]. First, GPT-4 was given the role of a professional language teacher who is an expert on providing feedback on the writing of English language learners. Second, the weekly writing prompt that students were given was included. Third, the LLM was instructed to, using simple language, comment on six areas of the students' writing: the topic sentence, the development of ideas, language that lowers the academic quality of the writing, the use of transitional phrases, the use of sources and evidence, and the grammatical accuracy of the language of the writing. The AI was instructed to put the feedback on grammatical accuracy into a table that organized the following elements: the sentence where the error in the writing is found, the error type, a description of what this kind of error is, and suggestions as to how to address the error. Lastly, the student's paragraph was copied into the prompt. An example prompt and resultant feedback are given in appendices B and C. A teaching assistant utilized GPT-4 and the prompt described above to generate feedback for each student. Feedback was emailed to students within two working days to ensure students had ample time to read the feedback and incorporate it into their revisions, should they desire to.

Data collection procedures

In study 1, students completed the pretest in the first week of the semester. For the next six weeks these students completed a weekly writing assignment. For each week students wrote a 300 word paragraph on a provided topic related to the material in class, received feedback either from a human tutor (CG) or AI (EG), revised, and submitted a final draft. Human tutors were paid trained English language tutors from the university's ENL Tutor Program and certified by the College Reading & Learning Association. Participants in the CG held one 30-min one-on-one tutoring session per week with the same tutor for the duration of the study. The EG was required to submit an initial draft of their writing assignment each week. These submissions were reviewed to remove any identifying information. We then utilized GPT-4 to generate individualized feedback for each student, which was subsequently sent to them via email. Upon receiving and reviewing the AI-generated feedback, students made revisions to their writing and submitted their final drafts. In week 8, the final week of the semester, these students completed the posttest. Both pre- and posttests were independently rated by two experienced academic English language instructors. An analytic rubric assessing four key writing areas, namely content, coherence, language use, and sources and evidence, was used to assess students' writing in the pre- and post test. Each category was scored individually on a scale of four, with descriptors outlining the performance at each level. Scores for each criterion varied from 1 to 4, with 1 representing an "Initial" level of performance and 4 representing a "Highly Developed" level of performance. The maximum achievable score on the rubric was 40. Inter-rater reliability was assessed by calculating intraclass correlation coefficients (ICC). These values indicated excellent inter-rater reliability for the CG and EG pretests (0.932, $p < 0.001$; 0.919, $p < 0.001$), as well as good reliability for the CG and EG posttests (0.877, $p < 0.001$; 816, $p < 0.001$).

In study 2, participants received feedback from a human tutor *and* from AI (GPT-4) on their weekly writing assignments. These human tutors were unaware that the participants were also receiving AI-generated feedback. The preference survey was distributed to participants via Qualtrics, once a week after students had completed their weekly writing assignment.

Data analysis procedures

To explore the first research question in study 1, a repeated-measure analysis of variance (RM-ANOVA) was conducted using SPSS 28. Using a general linear model, the time of the tests (pretest [T1] and posttest [T2]) served as a within-subjects variable, with the group serving as a between-subjects variable (experimental group or EG, and control group or CG). Missing data, outliers, and possible statistical assumption violations were examined. The scores of one student who only completed the posttest were not included in the analysis. Shapiro–Wilk test results suggested the data was normally distributed, except for the posttest control group (0.833, $p < 0.001$). Levene's test of equality of error variances indicated homogeneity of variance for both pre- ($p = 0.888$) and posttest ($p = 0.938$). RM-ANOVA results reported below follow a Greenhouse-Giesser correction. Lastly, two independent samples *t*-tests were conducted to explore potential differences in means at T1 and T2 to see if there were any significant differences in proficiency levels between the two groups before and after the treatment.

To explore the second research question in study 2, descriptive statistics were calculated for questions 1–9 for each week. Three researchers independently performed a thematic analysis of the qualitative data and then consulted on salient themes found in the data.

Results

Relating to RQ1 in Study 1

Descriptive statistics for the EG and CG scores for the pre- (T1) and posttest (T2) along with results of the Shapiro–Wilk tests are reported in Table 1. Both groups made similar

Table 1 Descriptive statistics of writing scores across groups and time

Test	Group	Descriptive statistics					Shapiro–Wilk		
		N	Mean	SD	Skewness	Kurtosis	Statistic	df	Sig.
T1	EG	23	27.522	1.301	0.287	− 0.505	0.939	23	0.167
	CG	25	26.820	1.391	− 0.882	0.331	0.926	25	0.070
T2	EG	23	33.370	1.908	− 0.361	− 1.001	0.934	23	0.137
	CG	25	33.680	2.066	1.570	2.698	0.833	25	0.000*

[*] $p < 0.001$

Table 2 Results of RM-ANOVA

	Type III sum of squares	df	Mean square	F	Sig.	Partial Eta squared (η_p^2)
Between-subjects						
Group	0.917	1	0.917	0.241	0.626	0.005
Error	174.989	46	3.804			
Within-subjects						
Time	967.251	1	967.251	487.661	0.000*	0.914
Time * Group	6.136	1	6.136	3.094	0.085	0.063
Error (Time)	91.239	46	1.983			

Table 3 Results of independent samples *t*-tests of between-subject variable (group)

Test	Group	N	Mean	SD	*t*	*df*	Sig. (2-tailed)	Cohen's d
1	EG	23	27.522	1.301	1.801	46	0.078	0.520
	CG	25	26.820	1.391				
2	EG	23	33.370	1.908	− 0.539	46	0.591	− 0.156
	CG	25	33.680	2.066				

progress in their academic writing over the six week long treatment, as indicated by the increase in mean scores for the EG (5.848) and CG (6.86).

The results of the mixed 2 × 2 RM-ANOVA analysis (Table 2) revealed there was no significant interaction effect between group and time ($F = 3.094$, $p = 0.085$, $\eta_p^2 = 0.063$) The effect size of the difference (η_p^2) signifies that this two-way interaction accounts for 6.3% of the variance in scores. The test of between-subjects effects revealed no significant difference between the EG and CG ($F = 0.241$, $p = 0.626$, $\eta_p^2 = 0.005$), indicating that the method of providing feedback (human tutor or AI) did not have a significant effect on students' posttest scores; the between-subjects variable (group) only accounted for 0.5% of the variance.

Independent sample *t*-tests of the between-subjects variable at T1 and T2 are reported in Table 3. No significant difference was found between EG and CG means in the pre- and posttest, suggesting no significant difference in writing proficiency between the groups at either times of measurement.

Relating to RQ2 in Study 2

The descriptive statistics for question items 1–9 of the weekly preference survey are shown in Appendix A. The data suggest a near even split in preferences. Question 9, for example, asked "If you were to only get one kind of feedback next week, which kind of feedback would you prefer?". Figure 1 shows the slight fluctuations in mean responses to this question, with a value of 1.5 indicating an even split in preference. The six-week average of the number of students that preferred feedback from human tutors was 18, which was slightly lower than the average of those that preferred AI-generated feedback (19.667), although this may be due to fewer students completing the survey in the final week. The means for human tutor feedback (H) for the entire six weeks, were slightly higher in each pair of items: satisfaction (Q1&2), H = 4.277 AI = 4.262; clarity (Q3&4),

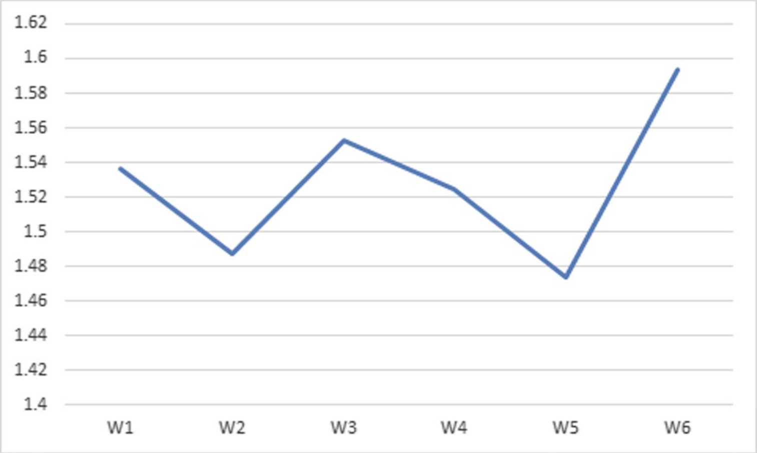

Fig. 1 Mean responses to item 9, where 1 equals preference for human feedback and 2 equals preference for AI feedback

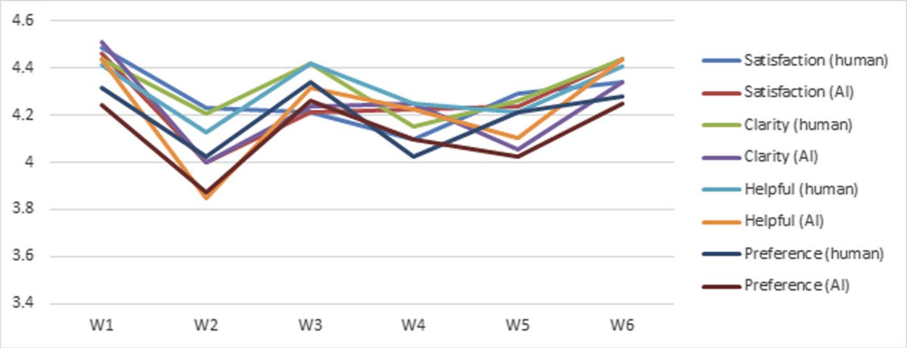

Fig. 2 Means of items 1–8

H = 4.319 AI = 4.236; helpfulness (Q5&6), H = 4.305 AI = 4.228; and preference (7–8), H = 4.2 AI = 4.126. Given how close each pair of means were, and to further explore the data, students were grouped according to how they answered Q9 in week six, and a *t*-test for comparing the mean response of these two groups for each item was conducted. All means were nonsignificant between these groups, further suggesting that student preference was equally split.

Line graphs that included items 1–8 (Fig. 2), and each pair of items (Figs. 3, 4, 5, 6) show two trends. First, preference for human tutor feedback is generally slightly higher. Second, means were highest in weeks one, three, and six, with slight dips in weeks two, four, and five. Important to note, however, is that the means for items Q1-8 in all weeks, except for Q6&8 in week two, were above 4, suggesting that students generally perceived both forms of feedback as being of value.

An analysis of the qualitative data did not yield any insights into why the means of items 1–8 dipped in weeks two, four, and five. It may be that the writing topics proved more difficult in these weeks, or that students did not get as much out of the human or AI-generated feedback because of other pressures in their academic and/or personal lives, such as having unit tests during these weeks.

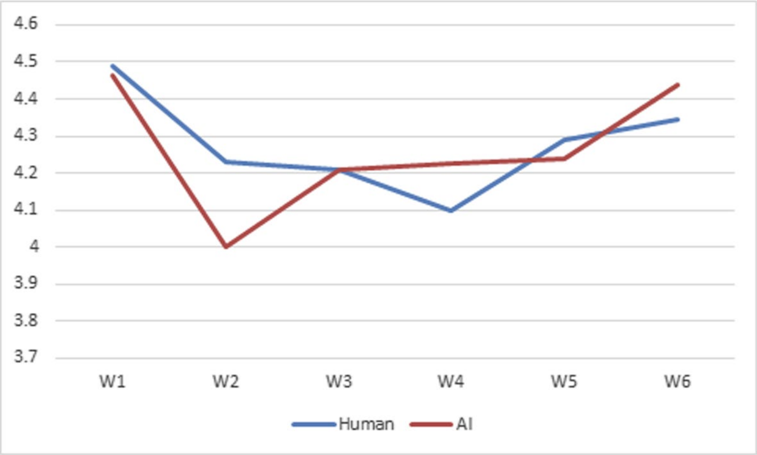

Fig. 3 Means of items 1 and 2 relating to satisfaction with feedback

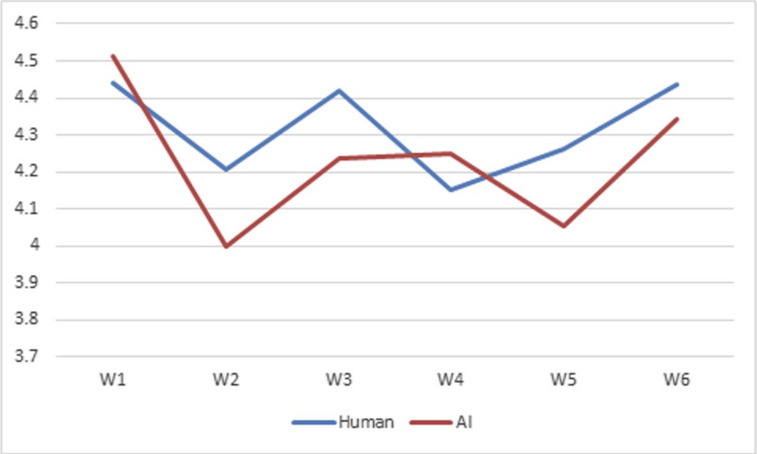

Fig. 4 Means of items 3 and 4 relating to clarity

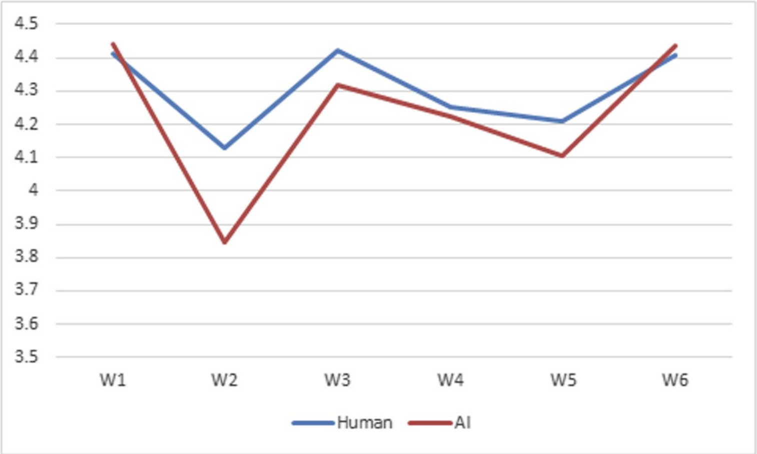

Fig. 5 Means of items 5 and 6 relating to helpfulness

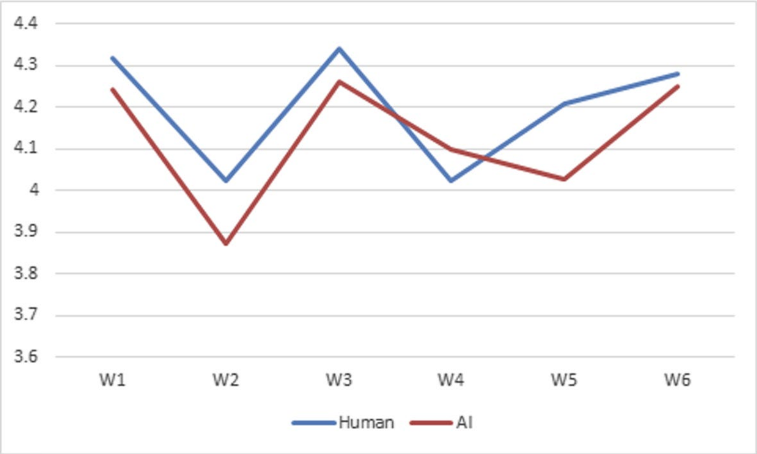

Fig. 6 Means of items 7 and 8 relating to preference

There were several recurring themes in the qualitative data as to why some students preferred receiving feedback from a human tutor. Perhaps the most prominent was the affordances of sitting down face to face and interacting with a human. For some this was viewed as beneficial because they found personal interaction to be "more engaging" and "a fun way to learn" when compared to just reading through AI-generated feedback. Some cited the ability to ask follow-up questions and get immediate feedback as being instrumental. Others noted that interacting with a human tutor allowed them to develop their writing and speaking skills at the same time. One student observed that "AI comment is super helpful, but personal characteristic [sic] could be missing after AI revision", suggesting that reliance on AI might unintentionally result in the erasure of the personal voice of the writer.

As for those students that preferred receiving AI-generated feedback, clarity, understandability, consistency, and specificity of feedback, especially in regards to academic writing style and vocabulary, were common themes found in the data. Several students commented on how the AI-generated detailed feedback on errors in academic writing. For example, one student wrote "the AI program provides me with concrete feedback and easy-to-understand documentation of where the errors were." Another noted "The AI feedback was so accurate and it also suggested academic words that I could use in place of the words that I used which was not very academic." Several students commented how AI does not have constraints of time or availability, and students can review the feedback whenever they want: "AI gives me correct and accurate feedback on every sentence, no matter the time. However, [ENL] tutors have limited time and can only get limited feedback."

Several students highlighted the advantages of both forms of feedback. For example, one student wrote "I prefer to have both. The reason why I like my [ENL] tutor, is because I think interaction helps my brain to learn and focus, so I could improve and progress. On the other hand, the AI also was very helpful, it pointed out the problems precisely and clear to understand. I would say they are a good combination for students." Another student wrote "If there was a question whether to receive feedback from both [ENL] tutors and AI program I would say yes because I would be able to learn more from both".

Discussion and conclusion

Study 1 compared human tutor and AI-generated feedback to see if one would influence linguistic gains more than the other. The results indicate that AI-generated feedback did not result in superior linguistic progress among ENL students compared to those who received feedback from a human tutor. The between-subject variable of group did not have a significant effect on writing scores, suggesting that one method of feedback was not better than another in terms of scores.

While the study found that AI-generated feedback did not lead to superior linguistic progress among ENL students compared to human tutor feedback, it is important to consider the potential time-saving benefits offered by AI-generated feedback for educators. Utilizing AI for providing feedback can potentially significantly reduce the time teachers spend on reviewing and responding to each student's assignment, thereby freeing up valuable time for other tasks. Furthermore, the time efficiency of AI-generated feedback can be particularly advantageous in large classes where providing individualized feedback by the instructor is logistically challenging and time-consuming.

Study 2 investigated which form of feedback ENL students preferred and why. We found about half the students preferred receiving feedback from a human tutor, and half preferred AI-generated feedback. Those that preferred sitting down and discussing their feedback with a tutor cited the face-to-face interaction as having affective benefits, such as increasing engagement, as well as benefits for developing their speaking abilities. Those that preferred AI-generated feedback primarily cited the clarity and specificity of the feedback as being useful for improving their writing. This echoes the findings of Dai et al. (2023), namely that AI-generated feedback was found to be more readable and detailed than feedback from an instructor.

We offer several suggestions as to how the inclusion of AI-generated feedback might be better incorporated into practice. In this study, as students were emailed the feedback generated by the AI, the students had no opportunity to ask follow-up questions to the AI. This is because we wanted to limit students' access to the AI and prevent potential misuses where students asked the AI to write their assignments for them. However, providing opportunities for students to ask follow-up questions to the AI may lead to a greater preference in AI-generated feedback.

In light of the major findings highlighted above, we believe a mixed approach to providing feedback may be most beneficial for both language educators and students. By utilizing GenAI, language educators may be able to produce more detailed feedback in a shorter amount of time for each individual learner. Providing opportunities for students to discuss AI-generated feedback with a human tutor and ask follow up questions affords students with the benefits of each modality, namely the clarity and specificity of the AI-generated feedback, and the benefits of interacting with another human, such as engagement and the ability to practice speaking.

With AI industry leaders predicting artificial capable intelligence, able to perform day to day tasks, being available in two years (Suleyman, 2023), and super artificial intelligence arriving potentially this decade (Leike & Sutskever, 2023), it is of growing importance that practitioners in the field of language education become more familiar with this rapidly changing technology, its potential uses, and how it may drastically influence and personalize language education in the future.

Currently GPT-4 is not optimized for AWE purposes and therefore features like text annotation, common in existing AWE programs, are cumbersome to reproduce. However, some established AWE programs have begun to integrate GPT technology (e.g., *GrammarlyGO*). It is likely only a matter of time before large language models exist that have been fine tuned and optimized for language learning and teaching purposes, including assessing writing. Furthermore, as the public gets increased access to these models (e.g., through application program interfaces, or APIs), GenAI will likely be woven into the learning management systems commonly used in educational institutions, thereby becoming a more integral part of the practices of educators and students.

A potentially fruitful avenue of research would be to examine how the proficiency levels of students affect their ability to understand and learn from AI-generated feedback. Furthermore, investigating GenAI's ability to reliably assess and score writing would be insightful. Lastly, in a similar vein to this study, further research could focus on investigating the efficacy of AI-generated feedback on native English-speaking students.

In addition, we encourage language educators to consider the following questions:

- What aspects of the language learning process are best performed by GenAI, now and in the future?
- What aspects of the language learning process are best performed by humans, now and in the future?
- As GenAI becomes more capable and prevalent, what skills will become more important for language educators to cultivate?

To conclude, while admittedly there are a number of vehicles for personalized learning, the potential of GenAI in this area merits further attention. As GenAI continues to be developed and permeate the sphere of language education, it becomes imperative to ensure a balanced approach, one that capitalizes on its strengths while duly recognizing the indispensable contributions of human pedagogy. The endeavor of comprehending and assessing the capabilities of GenAI, along with its potential influence on language learning and teaching, is arguably now of paramount importance.

Appendix A

See Table 4.

Table 4 Mean and SD for survey questions across six weeks

Item	W1 (n = 41)	W2 (n = 38)	W3 (n = 37)	W4 (n = 40)	W5 (n = 38)	W6 (n = 32)	Grand Mean
Q1 I am satisfied with the feedback I received from the ENL tutor this week	4.49 (0.98)	4.23 (1.35)	4.21 (1.02)	4.1 (1.28)	4.29 (1.16)	4.34 (0.97)	4.28 (1.13)

Table 4 (continued)

Item	W1 (n = 41)	W2 (n = 38)	W3 (n = 37)	W4 (n = 40)	W5 (n = 38)	W6 (n = 32)	Grand Mean
Q2 I am satisfied with the feedback I received from the AI program this week	4.46 (0.9)	4 (1.38)	4.21 (1.04)	4.23 (1.25)	4.24 (1.1)	4.44 (0.95)	4.26 (1.1)
Q3 The feedback from the ENL tutor was clear and easy to understand	4.44 (1.12)	4.21 (1.36)	4.42 (0.98)	4.15 (1.31)	4.26 (1.16)	4.44 (0.91)	4.32 (1.14)
Q4 The feedback from the AI program was clear and easy to understand	4.51 (0.95)	4 (1.12)	4.24 (1.1)	4.25 (1.21)	4.05 (1.27)	4.34 (0.83)	4.24 (1.11)
Q5 The feedback from the ENL tutor helped me improve my writing	4.42 (0.97)	4.13 (1.32)	4.42 (0.95)	4.25 (1.17)	4.21 (1.28)	4.41 (0.91)	4.31 (1.1)
Q6 The feedback from the AI program helped me improve my writing	4.44 (1.0)	3.85 (1.42)	4.32 (0.99)	4.23 (1.23)	4.11 (1.23)	4.44 (0.95)	4.23 (1.14)
Q7 I prefer receiving feedback from the ENL tutor	4.32 (0.96)	4.03 (1.33)	4.34 (0.91)	4.03 (1.33)	4.21 (1.26)	4.28 (0.99)	4.2 (1.13)
Q8 I prefer receiving feedback from the AI program	4.24 (1.04)	3.87 (1.38)	4.26 (0.98)	4.1 (1.19)	4.03 (1.33)	4.25 (0.92)	4.13 (1.14)
Q9 If you were to only get one kind of feedback next week, which kind of feedback would you prefer?	1.54 (0.51)	1.49 (0.51)	1.55 (0.5)	1.53 (0.51)	1.47 (0.51)	1.59 (0.5)	1.53 (0.5)

Appendix B

Sample prompt sent to GPT-4 to generate feedback

You will be a professional language teacher who is an expert on providing feedback on the writing of English language learners. Here is the writing prompt that students are given: [the weekly writing prompt was inserted here].

Below I will share with you a student's writing. Based on their writing, comment on the following:

1. Using simple language, comment on the quality of the topic sentence and if it addresses the writing prompt. Provide suggestions for improvement but don't write a new topic sentence for the student. Provide an example of an improved topic sentence that is about a different topic than the student's writing. Start your feedback with the header in bold "Feedback on the quality of the topic sentence:".

4. Using simple language, comment on the development of ideas throughout the paragraph. Specifically comment on the development of the main idea through supporting ideas and elaborating details such as examples and evidence. Start your feedback with the header in bold "Feedback on the development of ideas throughout the paragraph:".
5. Using simple language, identify language in the paragraph that lowers the paragraph's academic quality. Provide examples of academic words and phrasing that would improve the writing. Start your feedback with the header in bold "Feedback on the academic quality of the language:".
6. Using simple language, examine the use of transitional phrases in the paragraph. Are they appropriate for academic writing? How could they be improved? Start your feedback with the header in bold "Feedback on the transitional phrases:".
7. Using simple language, evaluate the paragraph according to its use of sources and evidence. Start your feedback with the header in bold "Feedback on the use of sources and evidence:".
8. Using simple language, comment on the grammatical accuracy of the language of the writing, such as spelling, capitalization, punctuation, singular and plural nouns, verb tense, subject verb agreement, word form, awkward phrasing, prepositions, articles, and sentence fragments and run-ons. Put the feedback on grammatical accuracy in a table with four columns. The first column includes the sentence where the error in the student's writing is found. The second column is the error type. The third column is a description of what this kind of error is. The fourth column is a suggestion on how to address the error.

Here is the student's paragraph. Using simple language, please provide feedback in the manner described above.

[The student's paragraph was inserted here.]

Appendix C

Example AI-generated feedback

Original student writing

The most important aspect of culture is a culture of perception, which helps build better societies and relationships. I know this is a unique culture for other people and has some disadvantages however, this is a very Japanese thing and important for Japanese people. In my opinion, this is a virtue of the Japanese people and something to be proud of. There are two reasons why I think the culture of perception is the most important aspect of culture. First of all, this culture led us to be the people who think and act on their own. For example, when I was 13 years old, I had an opportunity to help my father's work. And I observed what my father doing and thought about how I can help him to work more efficiently. Thanks to that experience, I could move to actions beforehand my other coworkers tell me what want me to do. It had helped me many times. Moreover, even in other situations, such as assisting with activities and building good relationships. The second reason is that it leads us to be able to work more effectively and efficiently. In any kind of situation, communication will be an obstacle to working faster sometimes. I have one example of this. In the doctor industry, surgery is a race against time, and people in the surgery room should work efficiently and must be consistent in them. Furthermore, people who work for doctors should support them by observing their movement so that doctors can work without stress. I know that this culture also leads us in negative ways, such as exclusivism and less skill in conversation. I think that we have to understand and use this culture correctly so that we avoid negative results. In conclusion, the culture of perception is the most important aspect of culture and helps us to work more effectively and efficiently.

AI-generated feedback

Feedback on the quality of the topic sentence

The current topic sentence introduces the main idea of the paragraph—the importance of the culture of perception in building societies and relationships. However, the topic sentence could be improved by being more concise and clearly stating the main idea. For example, for a different topic, an improved topic sentence might be: "Implementing renewable energy sources is crucial for mitigating the effects of climate change and promoting sustainable development."

Feedback on the development of ideas throughout the paragraph

The paragraph does a decent job of developing the main idea of the culture of perception being the most important aspect of culture. The author provides two reasons for this belief and elaborates on each reason with examples and evidence. However, the paragraph could benefit from more specific details and evidence to support the claims. Additionally, connecting the examples back to the main idea more explicitly would strengthen the paragraph.

Feedback on the academic quality of the language

The paragraph contains some informal language that detracts from its academic quality. For example, "this is a very Japanese thing" could be replaced with "this cultural aspect is particularly significant in Japan." Instead of "I know this is a unique culture for other people," a more academic phrasing could be "It is evident that this cultural aspect may be distinct for individuals from other backgrounds."

Feedback on the transitional phrases

The paragraph includes some transitional phrases, such as "First of all," "Moreover," and "In conclusion." These phrases are generally appropriate for academic writing. However, the transitions could be improved by using more varied phrases, such as "Additionally" or "Furthermore," to enhance the flow of ideas and maintain a consistent academic tone.

Feedback on the use of sources and evidence

The paragraph includes personal anecdotes and examples to support the main idea. While these examples help to illustrate the points being made, incorporating more external sources, such as studies, expert opinions, or statistics, would strengthen the argument and lend credibility to the claims.

Feedback on the grammatical accuracy

Sentence	Error type	Description	Suggestion
The most important aspect of culture is a culture of perception, which helps build better societies and relationships	Word form	"a culture of perception" is a repetitive phrase	Replace with "perception as a cultural aspect"
I know this is a unique culture for other people and has some disadvantages however, this is a very Japanese thing and important for Japanese people	Run-on sentence	Two independent clauses are not separated properly	Use a period or a semicolon before "however"
Thanks to that experience, I could move to actions beforehand my other coworkers tell me what want me to do	Awkward phrasing	The sentence structure is unclear	Rewrite as "Thanks to that experience, I could take action before my coworkers told me what they wanted me to do."
It had helped me many times	Verb tense	The past perfect tense is not necessary here	Change "had helped" to "has helped"
The second reason is that it leads us to be able to work more effectively and efficiently	Wordiness	The sentence could be more concise	Rewrite as "The second reason is that it enables us to work more effectively and efficiently."
I know that this culture also leads us in negative ways, such as exclusivism and less skill in conversation	Preposition	Incorrect use of preposition	Change "leads us in" to "leads us to"

Abbreviations

AI	Artificial intelligence
ENL	English as a new language
GPT-4	Generative Pretrained Transformer 4 (from OpenAI)
ChatGPT	A specific chatbot interface for the GPT models
AWE	Automated writing evaluation
LLMs	Large language models
GenAI	Generative Artificial Intelligence
PaLM 2	A language model from Google
GRE	Graduate record examination
TOEFL	Test of English as a Foreign Language
BERT	Bidirectional Encoder Representations from Transformers (another type of language model)

Acknowledgements
We would like to extend our sincere gratitude to our Research Assistant.

Author contributions
JE and AP: Research design, instruments. JE: Data collection. AP and AB: Data analysis. JE, AP, and AB: Writing and revising.

Funding
This research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors.

Availability of data and materials
The datasets generated and/or analyzed during the current study are available from the corresponding author on reasonable request.

Declarations

Competing interests
The authors declared that they do not have any competing interests.

Received: 11 August 2023 Accepted: 18 October 2023
Published online: 27 October 2023

References

Abd-Elaal, E.-S., Gamage, S., & Mills, J. (2022). Assisting academics to identify computer generated writing. *European Journal of Engineering Education*. https://doi.org/10.1080/03043797.2022.2046709

Baktash, J. A. & Dawodi, M. (2023). Gpt-4: A review on advancements and opportunities in natural language processing. [preprint in arXiv]. https://doi.org/10.48550/arXiv.2305.03195

Behizadeh, N., & Engelhard, G., Jr. (2011). Historical view of the influences of measurement and writing theories on the practice of writing assessment in the United States. *Assessing Writing, 16*(3), 189–211. https://doi.org/10.1016/j.asw.2011.03.001

Chiu, T. K. F., Xia, Q., Zhou, X., Chai, C. S., & Cheng, M. (2023). Systematic literature review on opportunities, challenges, and future research recommendations of artificial intelligence in education. *Computers and Education Artificial Intelligence*. https://doi.org/10.1016/j.caeai.2022.100118

Dai, W., Lin, J., Jin, F., Li, T., Tsai, Y.-S., Gašević, D. & Chen, G. (2023). Can large language models provide feedback to student? A case study on ChatGPT. [Preprint from EdArXiv]. https://doi.org/10.35542/osf.io/hcgzj

Davis, F. D. (1989). Perceived usefulness, perceived ease of use, and user acceptance of information technology. *MIS Quarterly*, 319–340.

Derner, E. & Batistič, K. (2023). Beyond the safeguards: Exploring the security risks of ChatGPT. [preprint in arXiv], abs/2305.08005. https://doi.org/10.48550/arXiv.2305.08005

Elkins, K., & Chun, J. (2020). Can GPT-3 pass a writer's Turing Test. *Journal of Cultural Analytics*. https://doi.org/10.22148/001c.17212

Elliot, N. & Klobucar, A. (2013). Automated essay evaluation and the teaching of writing. In M. D. Shermis & J. Burstein (Eds.), *The Handbook of automated essay evaluation: Current applications and new directions*. Routledge.

Farrokhnia, M., Banihashem, S. K., Noroozi, O., & Wals, A. (2023). A SWOT analysis of ChatGPT: Implications for educational practice and research. *Innovations in Education and Teaching International*. https://doi.org/10.1080/14703297.2023.2195846

Fitria, T. N. (2021). Grammarly as AI-powered English writing asssistant: Students' alternative for writing English. *Metathesis, 5*(1), 65–78. https://doi.org/10.31002/metathesis.v5i1.3519

Florio, S., & Clark, C. M. (1982). The functions of writing in an elementary classroom. *Research in the Teaching of English, 16*(2), 115–130.

Fütterer, T., Fischer, C., Alekseeva, A., Chen, X., Tate, T., Warschauer, M., & Gerjets, P. (2023). ChatGPT in education: Global reactions to AI innovations. *Research Square*. https://doi.org/10.21203/rs.3.rs-2840105/v1

Godwin-Jones, R. (2022). Partnering with AI: Intelligent writing assistance and instructed language learning. *Language Learning Technology, 26*(2), 5–24.

Grove, W. M., Zald, D. H., Lebow, B. S., Snitz, B. E., & Nelson, C. (2000). Clinical versus mechanical prediction: A meta-analysis. *Psychological Assessment*. https://doi.org/10.1037//1040-3590.12.1.19

Herbold, S., Hautli-Janisz, A., Heuer, U., Kikteva, Z. & Trautsch, A. (2023). AI, write an essay for me: A large-scale comparison of human-written versus ChatGPT-generated essays. [preprint in ArXiv], abs/2304.14276. https://doi.org/10.48550/arXiv.2304.14276

Huawei, S., & Aryadoust, V. (2023). A systematic review of automated writing evaluation systems. *Education and Information Technologies, 28*, 771–795. https://doi.org/10.1007/s10639-022-11200-7

Ingley, S. J., & Pack, A. (2023). Leveraging AI tools to develop the writer rather than the writing. *Trends in Ecology Evolution, 38*(9), 785–787. https://doi.org/10.1016/j.tree.2023.05.007

John-Steiner, V., & Mahn, H. (1996). Sociocultural approaches to learning and development: A Vygotskian framework. *Educational Psychologist, 31*(3–4), 191–206. https://doi.org/10.1080/00461520.1996.9653266

Koltovskaia, S. (2020). Student engagement with automated written corrective feedback (AWCF) provided by Grammarly: A multiple case study. *Assessing Writing*. https://doi.org/10.1016/j.asw.2020.100450

Krashen, S. D. (1982). *Principles and practice in second language acquisition*. Pergamon Press Inc.

Krishna, K., Song, Y., Karpinska, M., Wieting, J. & Iyyer, M. (2023). Paraphrasing evades detectors of AI-generated text, but retrieval is an effective defense. [preprint in ArXiv], abs/2303.13408. https://doi.org/10.48550/arXiv.2303.13408

Kumar, R. (2023). Faculty members' use of artificial intelligence to grade student papers: A case of implications. *International Journal for Educational Integrity*. https://doi.org/10.1007/s40979-023-00130-7

Lampropoulos, G., Ferdig, R. E., & Kaplan-Rakowski, R. (2023). A social media data analysis of general and educational use of ChatGPT: Understanding emotional educators. *SSRN*. https://doi.org/10.2139/ssrn.4468181

Leike, J. & Sutskever, I. (2023). *Introducing superalignment*. OpenAI. https://openai.com/blog/introducing-superalignment#fn-A

Mizumoto, A., & Eguchi, M. (2023). Exploring the potential of using an AI language model for automated essay scoring. *Research Methods in Applied Linguistics*. https://doi.org/10.1016/j.rmal.2023.100050

OpenAI. (2023a). GPT-4 System Card. https://cdn.openai.com/papers/gpt-4-system-card.pdf

OpenAI. (2023b). GPT-4 Technical Report. https://cdn.openai.com/papers/gpt-4.pdf

Roscoe, R. D., Wilson, J., Johnson, A. C., & Mayra, C. R. (2017). Presentation, expectations, and experience: Sources of student perceptions of automated writing evaluation. *Computers in Human Behavior, 70*, 207–221. https://doi.org/10.1016/j.chb.2016.12.076

Shermis, M. D., Burstein, J., & Bursky, S. A. (2013). Introduction to automated essay evaluation. In M. D. Shermis & J. Burstein (Eds.), *The handbook of automated essay evaluation: Current applications and new directions*. Routledge.

Strobelt, H., Webson, A., Sanh, V., Hoover, B., Beyer, J., Pfister, H., & Rush, A. M. (2023). Interactive and visual prompt engineering for ad-hoc task adaption with large language models. *IEEE Transactions on Visualization and Computer Graphics, 29*(1), 1146–1156. https://doi.org/10.1109/TVCG.2022.3209479

Suleyman, M. (2023). *My new Turing test would see if AI can make $1 million*. MIT Technology Review. https://www.technologyreview.com/2023/07/14/1076296/mustafa-suleyman-my-new-turing-test-would-see-if-ai-can-make-1-million/

Sullivan, M., Kelly, A., & McLaughlan, P. (2023). ChatGPT in higher education: Considerations for academic integrity and student learning. *Journal of Applied Learning & Teaching*. https://doi.org/10.37074/jalt.2023.6.1.17

Tate, T. P., Doroudi, S., Ritchie, D., Xu, Y., & Uci, M. W. (2023). Educational research and AI-generated writing: Confronting the coming Tsunami. [preprint in EdArXiv]. https://doi.org/10.35542/osf.io/4mec3

Tseng, W., & Warschauer, M. (2023). AI-writing tools in education: If you can't beat them, join them. *Journal of China Computer-Assisted Language Learning*. https://doi.org/10.1515/jccall-2023-0008

Urlaub, P., & Dessein, E. (2022). From disrupted classrooms to human-machine collaboration? The pocket calculator, Google Translate, and the future of language education. *L2 Journal, 14*(1), 45–59. https://doi.org/10.5070/L214151790

Weigle, S. C. (2013). English as a second language writing and automated essay evaluation. In M. D. Shermis & J. Burstein (Eds.), *The handbook of automated essay evaluation: Current applications and new directions*. Routledge.

Yang, M. (2023). *New York City schools ban AI chatbot that writes essays and answers prompts*. The Guardian. https://www.theguardian.com/us-news/2023/jan/06/new-york-city-schools-ban-ai-chatbot-chatgpt

Yeo, M. A. (2023). Academic integrity in the age of artificial intelligence (AI) authoring apps. *TESOL Journal*. https://doi.org/10.1002/tesj.716

Zhou, Y., Muresanu, A. I., Han, Z., Paster, K., Pitis, S., Chan, H. & Ba, J. (2023). Large language models are human-level prompt engineers. International Conference on Learning Representations 2023.

Zhu, C., Sun, M., Luo, J., Li, T. & Wang, M. (2023). How to harness the potential of ChatGPT in education? *Knowledge Management & E-Learning, 15*(2), 133–152. https://doi.org/10.34105/j.kmel.2023.15.008

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Juan Escalante is an Assistant Professor of English Language Teaching and Learning at Brigham Young University-Hawaii. His research interests include technology enhanced-language education, teacher training, and language assessment.

Austin Pack is Assistant Professor of English Language Teaching and Learning at Brigham Young University-Hawaii. His research interests include language learning motivation, technology-enhanced language education, and complex dynamic systems.

Alex Barrett is a PhD candidate in the department of Educational Psychology and Learning Systems at Florida State University. He designs and researches learning technologies and has research interests in the topics of human-computer interaction, immersive learning, and technology-enhanced language education.