Impact of ChatGPT on ESL students' academic writing skills: a mixed methods intervention study

Santosh Mahapatra[1*]

*Correspondence:
santosheflu@gmail.com

[1] K 127, BITS Pilani Hyderabad
Campus, Hyderabad 500078,
India

Abstract

This paper presents a study on the impact of ChatGPT as a formative feedback tool on the writing skills of undergraduate ESL students. Since artificial intelligence-driven automated writing evaluation tools positively impact students' writing, ChatGPT, a generative artificial intelligence-propelled tool, can be expected to have a more substantial positive impact. However, very little empirical evidence regarding the impact of ChatGPT on writing is available. The current mixed methods intervention study tried to address this gap. Data were collected from tertiary level ESL students through three tests and as many focus group discussions. The findings indicate a significant positive impact of ChatGPT on students' academic writing skills, and students' perceptions of the impact were also overwhelmingly positive. The study strengthens and advances theories of feedback as a dialogic tool and ChatGPT as a reliable writing tool, and has practical implications. With proper student training, ChatGPT can be a good feedback tool in large-size writing classes. Future researchers can investigate the impact of ChatGPT on various specific genres and micro aspects of writing.

Keywords: ChatGPT, ESL, Writing skills, Intervention, Feedback, AWE, AI

Introduction

Formative feedback positively impacts students' writing (Anderson & Ayaawan, 2023; Butterfuss et al., 2022Huisman et al., 2019). Usually associated with formative assessment strategies such as self-and peer assessment, formative feedback can be directed to address students' needs for explanation (Bozorgian & Yazdani, 2021; Zhang et al., 2023) concerning various aspects of writing such as content, organization, grammar, vocabulary, and style. Since formative feedback is a continuous process, it offers real-time support to students as they write (Zhu et al., 2020). However, it is time-consuming and challenging to implement in a large classroom (Golzar et al., 2022). Such crowded classrooms are parts of everyday realities in many educational institutions in most developing and underdeveloped countries across the Global South. Addressing students' individual feedback needs in such classrooms is a significant challenge. Computer-mediated feedback has evolved as a potential solution to this problem in writing classrooms (Taskiran & Goksel, 2022; Yamashita, 2021), especially in institutions of higher

education in countries where students have access to portable computing devices and the Internet (Mahapatra, 2021). Globally, artificial intelligence (AI)-driven automated feedback is fast becoming a norm (Rad et al., 2023). Based on a large language model, ChatGPT, a relatively new addition to the repertoire of AI tools, can provide meaningful writing samples (Barrot, 2023), adjust the difficulty level of texts matching leaners' proficiency level (Bonner et al., 2023), offer advice regarding various structural aspects of a text and translate it (Imran & Almusharraf, 2023), and facilitate guided writing (Kohnke et al., 2023). These affordances can aid self-reliance and fulfill instant feedback needs of students. It can provide human-like expert assistance to students in idea generation, organization, maintaining accuracy and choosing appropriate vocabulary (Tai et al., 2023). However, currently, little empirical evidence is available on the impact of ChatGPT on students' writing skills (Su et al., 2023). Though there are discussions on its utility as a feedback tool (Bonner et al., 2023), few empirical studies make such claims. Thus, the present study looked into the impact of ChatGPT as a feedback tool on the academic writing skills of undergraduate English as a second language (ESL) students in a relatively crowded classroom at a university. To achieve the aim, the study employed a mixed methods intervention design with ChatGPT (as a feedback tool) as the independent variable and writing skills as the dependent variable. It was hypothesized that the employment of ChatGPT as a feedback tool would significantly impact students' writing skills. Considering that the study was set in relatively crowded classes, the findings can be generalized to similar ESL/English as a foreign language (EFL) settings. Additionally, the attempt to use ChatGPT as a feedback tool can be of significant pedagogic value and lead to further explorations globally.

Literature review

The utility and positive impact of formative feedback in the writing classroom are well-established (Olsen & Huns, 2023). When operationalized in the form of self-and peer assessment (SA and PA), formative feedback leads to reflection, self-regulation, self-monitoring, and revision on the parts of students (Lam, 2018). SA and PA can be used to augment learning in large-size writing classrooms often encountered in developing and under-developed countries in the Global South (Fathi & Khodabakhsh, 2019; Mathur & Mahapatra, 2022). However, research on feedback in large writing classes is limited (Rodrigues et al., 2022). It has been reported that smarter techniques must replace traditional ways to offer personalized dialogic feedback to students (Kohnke et al., 2023). With the proliferation of AI-driven tools such as Grammarly, QuillBot, Copy.ai, Word-Tune, ChatGPT, and others, it has become easier for students to obtain feedback on their writing (Marzuki et al., 2023; Zhao, 2022). They have advanced automated writing evaluation (AWE) and feedback in writing (Gayed et al., 2022). While the literature on Grammarly as a feedback tool in the writing classroom is well-established (Fitriana & Nurazani, 2022; Koltovskaia, 2020; Thi & Nikolov, 2022) and its positive impact on writing has been investigated empirically (Chang et al., 2021), the use of AI chatbots like ChatGPT for leveraging feedback in writing classes is a relatively new area and requires further investigation (Barrot, 2023). Since ChatGPT is a large generative language model, its potential to help students with writing is immense. It is more student-friendly and can provide more need-based assistance than other AWE tools, as suggested by

Guo et al. (2022) and Rudolph et al. (2023). It can support student writing by providing appropriate directions related to content and organization as they write (Allagui, 2023). Since it can automatically train itself and learn from previous conversations (Chan & Hu, 2023), students can receive tailored feedback suitable for individual needs.

Like earlier AI chatbots, ChatGPT can be used for generating ideas and brainstorming (Lingard, 2023). Recently, it has been accepted that ChatGPT can make writing easier and faster (Stokel-Walker, 2022). This potential, when exploited by teachers, can be converted into a dependable feedback tool. Wang and Guo (2023) discuss ChatGPT supporting students with learning grammar and vocabulary. As pointed out by Rudolph et al. (2023), irrespective of students' ability to use language accurately to ask questions, ChatGPT can provide feedback and information. In another study by Dai et al. (2023), students received corrective feedback from ChatGPT. Mizumoto and Eguchi (2023) also highlight similar findings when they tried ChatGPT as an AWE tool. In a study conducted in Saudi Arabia, Ali et al. (2023) discuss the positive impact of its use on learners' motivation. It could be due to its ability to provide reliable explanations (Kohnke et al., 2023) without the student having to go through the anxiety of asking the query in a classroom (Su et al., 2023). Since coming into existence in the last part of 2022 (OpenAI, 2022), ChatGPT has gained immense popularity among language educators. It has been reported as capable of producing high-quality texts (Gao et al., 2022), offering feedback on text organization, language use and recommending corrections (Ohio University, 2023), logically organizing content, adding appropriate supporting details and conclusion (Fitria, 2023). While Yan (2023) has reported benefits to students' writing skills through its use, he has also warned that its use can threaten academic honesty and ethicality in writing.

Theoretically, the utilization of ChatGPT for leveraging formative feedback can be placed within a framework comprising two theories. The first one is the theory of feedback as a dialogic process advocated by Winstone and Carless (2020). According to them, when feedback involves interactions, it leads to students clarifying their expectations, obtaining desired information and guidance, and making progress in learning. ChatGPT facilitates dialogue by responding to the user's queries regarding various aspects of writing. It offers suggestions when sought and functions as a support-on-demand tool. Additionally, it admits mistakes and rectifies itself thereby making dialogue meaningful. The second one is Barrot's (2023) theory of ChatGPT as a reliable writing tool that can provide immediate, need-based, and tailored feedback to students as they move through different stages of writing. The current study was built on these two theories as its aim was to utilize ChatGPT as a formative feedback tool involving SA and PA, and assess its impact on students' academic writing skills.

Research questions

The study addressed the following two research questions:

- In an intensive academic writing course, when the instructional hours and tasks are held constant, does the employment of ChatGPT as a feedback tool have any significant impact on undergraduate ESL students' writing skills?

- How do the experimental group students perceive the impact of ChatGPT as a feedback tool on their writing skills?

Methodology

Mixed methods intervention design

A mixed methods intervention design (Creswell & Plano-Clark, 2017) was adopted for the study. Since the study aimed to assess the impact of ChatGPT as a feedback tool and involved an intervention, a quasi-experimental design was an automatic choice for such a study, as with intervention studies in education (Gopalan et al., 2020). However, considering that qualitative data add to the validity of the claims and the strength of the study by providing in-depth details about how students used ChatGPT for self-and peer feedback during classroom assessments, it was logical to choose a mixed methods design. Figure 1 shows the design of the study.

Participants

An intact class of students randomly assigned to two sections was chosen for the study. This aligns with the sampling principle used for quasi-experimental studies in education and applied linguistics (Perry Jr., 2011). These were first-year science and engineering students in an elite private-run university in India. None of them had used ChatGPT to improve their writing skills before the intervention. However, the participants had the experience of using mobile phones, laptops, the Internet, and AI tools such as Grammarly. Most of them were from financially well-off backgrounds and had exposure to television, media, and books on various topics from an early age. They were informed about the nature of participation and what they were expected to do before the start of the intervention. Their participation was voluntary, and they were given the choice to opt out of the study at any point during the intervention. Finally, 78 people in the experimental group (EG) and 56 in the comparison group (CG) consented to participate in the study. Since they were in their first year, most students in CG did not know students from the EG. All of them studied in English medium schools before joining the undergraduate program, had English as their second language, and aged between 18 and 19 years. Only those who attended all the intervention classes, i.e., six hours and took the pre-, post-, and delayed post-tests,
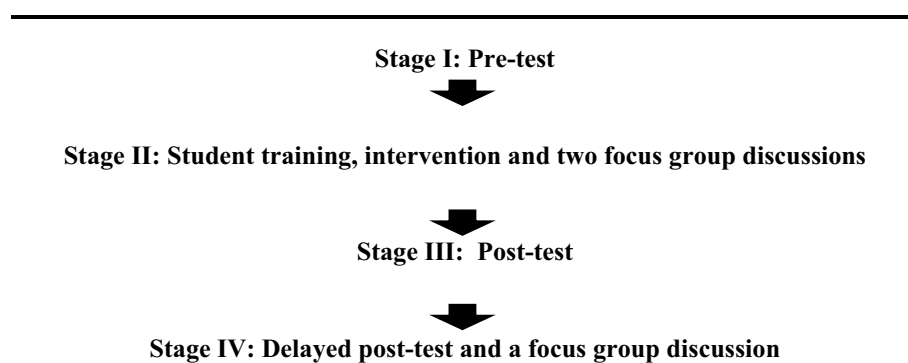

Stage I: Pre-test

Stage II: Student training, intervention and two focus group discussions

Stage III: Post-test

Stage IV: Delayed post-test and a focus group discussion

Fig. 1 Design of the study

were included in the reported data. Thus, 35 students from the EG and 37 students from the CG were included in the study. Six students volunteered when invited to participate in focus group discussions (FGDs). However, only five students participated in all three FGDs. Table 1 presents the inclusion and exclusion criteria.

Methods of data collection

Data were collected for the study primarily through three tests of writing. Each test comprised three tasks focusing on process, comparison, and cause-effect (see "Appendix"). Since all students studied science before joining the undergraduate program, they were familiar with scientific contexts used in the tasks. The rationale behind the task choice was that the writing genres were frequently used for academic purposes by the target group of students and were also part of the regular syllabus. Rubrics were created by the researcher and another teacher, who also taught the course to other groups of students, for each writing task type and were validated by two experts in applied linguistics. The evaluation criteria in the rubrics included content, organization, grammar, and vocabulary. The researcher and the other teacher evaluated students' writing performance using the rubrics. For all write-ups, Cohen's Kappa was found to be more than 0.8, indicating a strong inter-rater reliability. The qualitative data for the study were collected through FGDs with five EG students. The FGDs focused on obtaining views about and experiences of using ChatGPT for self-and peer feedback purposes. They were also encouraged to share artefacts as screenshots during the FGDs.

Procedure of data collection

Data were collected in various phases spanning almost an academic semester. First, students in the CG and EG took a pre-test. In the next phase, the EG was prepared through a short training program on using ChatGPT for SA and PA purposes in the writing classroom. In fact, Mathur and Mahapatra (2022) have recommended training for learners before using a digital tool in the classroom. After that, the intervention was undertaken in which the EG was taught process, comparison and cause-effect writing for six hours, and ChatGPT was used for SA and PA. Figure 2 presents the details about the intervention.

Table 1 Inclusion and exclusion criteria

Criterion	Condition for inclusion	Condition for exclusion
Age	Between 18 and 19 years	Less than 18 and more than 20
Academic qualification	Undergraduate first year in engineering and science	Any other year of undergraduation and any other disciplines
English as a second language	Who has English as their second language	Who has English as their first or foreign language
Number of classes attended	6	Anything fewer than 6
Participation in the classroom	Who participated in all classroom activities on self-and peer assessment and offered feedback	Who did not fully participate in classroom activities and did not offer adequate feedback

```
┌──────────────────────────────────────────────────────────┐
│        One-hour training on using ChatGPT for SA and PA     │
└──────────────────────────────────────────────────────────┘
                             ↓
┌──────────────────────────────────────────────────────────┐
│     Teaching of process, comparison and cause-effect writing │
└──────────────────────────────────────────────────────────┘
                             ↓
```

| Analysis of the genre and its features through tasks | → | Completion of text production tasks based on specific components of the genre | → | Completion of text production tasks in the form of paragraphs |

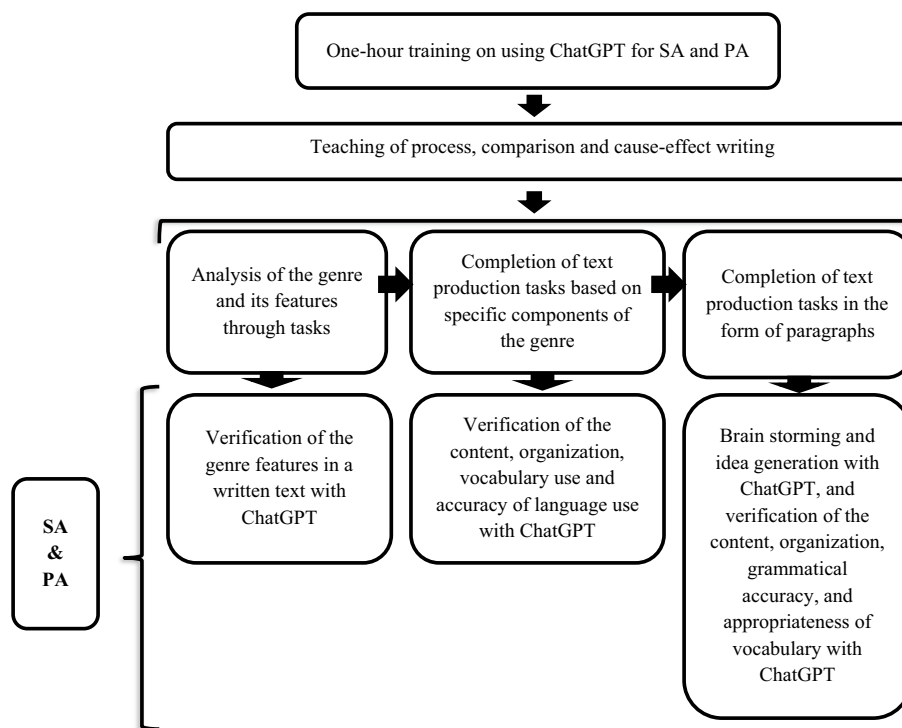| Verification of the genre features in a written text with ChatGPT | Verification of the content, organization, vocabulary use and accuracy of language use with ChatGPT | Brain storming and idea generation with ChatGPT, and verification of the content, organization, grammatical accuracy, and appropriateness of vocabulary with ChatGPT |

SA & PA

Fig. 2 Intervention plan

During the intervention, which was completed in a month, two FGDs were organized to obtain information about students' views on the employment of ChatGPT. The following prompts guided these discussions:

- We are using ChatGPT to assess our own and our peers' writing. How would you describe your experience with it?
- We are getting feedback on the topic sentence, supporting details, concluding sentence, use of signposts, appropriateness of content, and grammar. How beneficial do you think this exercise is?
- Do you have any suggestions for making the employment of ChatGPT more impactful?

In the fourth phase, a post-test was conducted for the CH and EG immediately after the end of the intervention. In the last phase, a delayed post-test was conducted for both groups and another FGD was conducted with five students from the EG. Both were organized almost two months after the post-test.

Analysis

The quantitative and qualitative data were analyzed separately and then triangulated to obtain answers to the research questions. Statistical analyses were performed for the quantitative data, which comprised pre-test, post-test, and delayed post-test scores for the CG and the EG. In fact, Rose et al. (2020) recommend using delayed post-tests for the robust assessment of the intervention impact. It is also a practice in writing related

Table 2 Descriptives

Tests	Mean	SD	N
Pre-test	12.581	0.833	37
Post-test	19.216	2.485	37
Delayed post-test	20.419	2.575	37

Table 3 Within-subjects effects

Cases	Sum of squares	df	Mean square	F	p	η^2
Score	1318.473[a]	2[a]	659.236[a]	330.704[a]	5.146e−37[a]	0.902
Residuals	143.527	72	1.993			

Type III sum of squares

[a] Mauchly's test of sphericity indicates that the assumption of sphericity is violated ($p < 0.05$)

interventions (Rezai et al., 2022). The data analysis involved a one-way RM ANOVA run on the EG's scores across three tests to compare the corresponding mean scores. A post-hoc Bonferroni test (Loewen & Plonsky, 2017) was run to control the overall error rate due to RM ANOVA. Then, a one-tailed *t*-test was computed to compare the pre-test, post-test and delayed post-test mean scores of the CG and the EG. Before the *t*-tests were run, Levine's test for equality of variance was calculated. The *f*-ratio value was 0.16078, and the *p*-value was 0.689658. The result was not significant at $p < 0.05$; thus, the homogeneity requirement is met. Also, a Shapiro–Wilk test was conducted to test normality. For the CG, it did not show a significant departure from normality, $W(35) = 0.98$, $p = 0.675$. For the EG too, no significant departure from normality was observed, $W(37) = 0.96$, $p = 0.168$. Grubb's test, performed to identify outliers, did not find any.

The qualitative analysis involved coding the transcripts of the FGDs. A phronetic iterative approach (Tracy, 2019) was adopted in this case. The approach was found suitable because the coding was inductive, and at the same time, it was guided by the themes that emerged through the review of literature and a few known patterns.

Findings

Positive impact of ChatGPT as a feedback tool

The impact of ChatGPT as a feedback tool on students' writing skills was positive and significant. The differences among the EG mean scores for the pre-test, post-test, and delayed post-test (see Table 2) indicate the trajectory of improvement in students' writing skills.

The claim is strengthened by the results from the one-way RM ANOVA computations. The results (F [2, 72] = 330.704, $p = 5.146e-37$, $\eta^2 = 0.902$) indicate a significant difference among some of the mean scores across the tests. A Bonferroni post-hoc test was performed to trace any significant differences between pairs due to the intervention. The differences between the pre-test and post-test ($p < 0.001$, $d = -3.300$) and pre-test and delayed post-test ($p < 0.001$, $d = -3.898$) scores were found to be statistically highly

Table 4 Between-subjects effects

Cases	Sum of squares	df	Mean square
Residuals	342.507	36	9.514

Type III sum of squares

Table 5 Post-hoc comparisons score

		Mean difference	SE	t	Cohen's d	p_{bonf}
Pre-test	Post-test	− 6.635	0.331	− 20.070	− 3.300	2.839e−30***
	Delayed post-test	− 7.838	0.331	− 23.708	− 3.898	1.075e−34***
Post-test	Delayed post-test	− 1.203	0.331	− 3.638	− 0.598	0.002**

Cohen's d does not correct for multiple comparisons

p-value adjusted for comparing a family of 3

**$p < 0.01$, ***$p < 0.001$

Table 6 Independent samples *t*-test of post-test scores

	t	df	p
Post-test scores	− 5.643	70	3.297e−7

Student's t-test

Table 7 Independent samples *t*-test of delayed post-test scores

	t	df	p
Score	− 9.371	70	5.544e−14

Student's t-test

Table 8 Group descriptives

	Group	N	Mean	SD	SE
Score	CG	35	15.400	1.897	0.321
	EG	37	20.419	2.575	0.423

significant. A significant difference was observed between the post-test and delayed post-test scores ($p < 0.01$, $d = − 0.598$) (see Tables 3, 4 and 5).

Since the findings from the one-way RM ANOVA offered little information about the comparative performance between the CG and the EG, two one-tailed independent sample *t*-tests were computed for both groups' post-test and delayed post-test scores. A significant difference was found between the post-test scores of the CG and the EG: $t(70) = − 5.643$, $p = 3.297e−7$, with the EG's ($M = 19.216$, SD $= 2.485$) mean score significantly higher than that of the CG ($M = 16.371$, SD $= 1.695$). The difference continued to be statistically significant for the delayed post-test scores of the CG and the EG: $t(70) = − 9.371$, $p = 5.544e−14$, with EG's ($M = 20.419$, SD $= 2.575$) mean

score significantly higher than that of the CG ($M = 15.400$, SD $= 1.897$) (see Tables 6, 7 and 8).

Students' positive perception of the impact

The FGDs were transcribed and coded. Table 9 presents the analysis.

The codes were classified under three main themes: content, organization, and grammar and two sub-themes: positive and specific and negative. When asked about their views on the impact of ChatGPT as a feedback tool, students in FGDs spoke in terms of content, organization, and grammar.

> *The way it guides us in obtaining the required information, arranging our ideas, and writing correctly is surprising. I was never aware that it could be a writing buddy. (S2, FGD 1)*
>
> *You see, when we were asking it to help us with organizing the ideas in writing, it gave us some directions. I guess that's quite helpful. It's like someone is constantly there to oversee your writing process. (S4, FGD 1)*
>
> *It's actually better than Grammarly in the sense that it explains the grammar issues when asked. You have a choice, and you can also learn from it. (S1, FGD 2)*

Table 9 Coded FGD data

Themes and Sub-themes	Codes	References
1. Content 1.1 Positive	Generation of more ideas on the topic	22
	Focused information on the topic	21
	Promotion of learner autonomy	19
	Promotion of collaboration	15
	Faster writing speed	13
	Creation of strong topic sentences	13
	Time-saving in brainstorming	8
1.2 Negative	Less motivation to think	4
	More machine dependence	3
2. Organization 2.1 Positive and specific	Easy to stay focused	22
	Strong connection between the main idea and the supporting details	17
	Appropriate use of signposts	11
	Reaffirming the main idea in the conclusion	8
	Enabling collaborative organization	7
2.2 Negative	Imposing a pattern on writing	2
	Hindering creative organization	2
3. Grammar 3.1 Positive and specific	Improving accuracy in the sentence structure	22
	Reliable source of grammar	21
	Providing explanations for language errors	17
	Making vocabulary choices more accurate	9
	Leading to discussions among peers	8
	Promotion of explicit knowledge of grammar	7
3.2 Negative	Reduction in attention to grammatical accuracy	5
	Increase in machine-dependence	3

They seemed to be happy about how ChatGPT helped them generate ideas and focused information on the given topic and work independently.

> *The best part about ChatGPT is that you ask for information, and you get it. You can go as specific or detailed as you wish. This reduced our thinking time invested to get ideas and information. (S5, FGD 3)*
>
> *I felt that it was a handy support tool for writing without being dependent on anyone. When we write, we usually have queries regarding several aspects of writing. With ChatGPT, you have a reliable support system with you. I like that freedom. (S3, FGD 2)*

They also highlighted how it promoted collaboration among peers, facilitated faster task completion, helped them create strong topic sentences and reduced brainstorming time.

> *When you use ChatGPT in a classroom with your classroom, you're doing it with several people. So much talk going on simultaneously! It's kinda cool. The conversations are so meaningful and without noticing, we are working together and writing. (S2, FGD 3)*
>
> *I absolutely love how we play with it together and how that fun is so productive. The process took much less time and there was this constant focused chatter which helped complete the tasks. We didn't miss anything significant, for example, when cheating the topic sentence, because one or two people are working with me and giving me feedback on the topic and the strength of the controlling idea. (S4, SGD 2)*

However, it was also pointed out on a few occasions that ChatGPT could lead to a lack of motivation to think and more machine dependence.

> *It might be a concern that my dependence might discourage me to do things on my own when writing. What if I won't want to write on my own? It can be scary, but I don't know. (S5, FGD 3)*

In their comments on the impact of ChatGPT as a feedback tool on organization in writing, it was an overwhelming claim that ChatGPT made it easier to stay focused on the topic when writing.

> *We felt that it keeps us on our toes. You know it's so easy to get diverted and include details unrelated to the topic. When you as ChatGPT, it tells you where you skid off the track. (S1, FGD 3)*
>
> *It's unreal! You share the topic sentence and the supporting details and it tells you if and how you have adhered to the controlling idea in your details. Isn't that cool? I don't think we can do that so accurately on our own. (S3, FGD 3)*

Students also mentioned that through feedback, they could add appropriate supporting details related to the main idea, use proper signposts, and write strong conclusions. More collaboration among peers when organizing the content was also highlighted.

> *When you asked us to write that paragraph on AI use in education, I created a topic sentence and added the required details along with a concluding sentence. When I asked GPT to tell me if my concluding sentence is a good one for the topic sentence,*

the feedback surprisingly good. I did the same when sharing feedback on my friends'
writing. (S5, FGD 2)
I'm happy that I have improved my signpost use. In fact, most of my classmates have
too. It made me conscious about the choice of signposts. The thread connecting infor-
mation and ideas suddenly felt more robust. Sometimes, the explicit explanation
with examples helped. (S4, FGD 3)

Though a minor, nonetheless, ChatGPT was claimed to impose a pattern on writing and hinder creativity in content organization.

I'm not sure, but sometimes it feels like I'm under a spell and I'm arranging informa-
tion as directed. Though it's my responsibility to choose and accept, I may be getting
too lazy to use my own creativity to place things in order. (S2, FGD 3)

When talking about the impact of ChatGPT on grammar, students highlighted Chat-GPT as a reliable source of grammar, a tool for improving accuracy in sentence structure and obtaining explanations for language errors.

I've been using Grammarly for a while, but it provides explanations for the gram-
mar queries or when it identifies an error. Nothing like knowing the details about
the issue. We kinda get sucked into curiosity by asking it questions on sentence struc-
ture, tense use and other aspects of grammar. It provides detailed explanations with
examples. Good alternative to Grammarly, dictionary and other such stuff. (S1,
FGD 3)

On several occasions, students shared artefacts showing how ChatGPT directs them to use of zero conditionals in scientific writing. Thus, when asked to verify 'At first, we take a bowl full of water and heat it. When it boils, we will stop the stove. Then, we take it out.' by a student, ChatGPT provides a polished version: 'Initially, a bowl filled with water was taken and heated. Once the boiling point is reached, the stove was promptly turned off. Subsequently, the bowl was carefully removed.' On some other occasions, students were encouraged by ChatGPT responses and sought further explanations related to the correct use of punctuation, articles, and sentence structures in formal contexts. In the above-mentioned case, ChatGPT explains that the use of past tense and passive voice are common in scientific writing, especially in the methods section and that passive voice adds to objectivity and clarity in writing.

The minor patterns included the claims related to helping with vocabulary choices, peer discussions and contributing to explicit knowledge of grammar. Two relatively minor patterns indicating negative impacts also emerged from the FGDs: lack of attention among students in terms of maintaining grammatical accuracy in writing and an increase in machine dependence.

Discussion

This mixed methods intervention study assessed the impact of ChatGPT as a formative feedback tool on academic writing skills of undergraduate ESL students, which comprised the first research question. In addition, it also captured students' views on the impact of ChatGPT on their writing skills, which was the focus of the second research question. The answer to the first question was obtained through the quasi-experimental

study, and the FGD data was analyzed to find the answer to the second question. The findings from the experiment corroborated those from the FGDs. Few studies before this (published in mainstream journals) have empirically explored the impact of ChatGPT on academic writing skills. Thus, the current study bears significance. The study contributes to several areas of research on academic writing. First, it demonstrates the utilization of ChatGPT for formative feedback purposes in an academic writing classroom. Second, it is conducted in a natural setting without many intrusive measures. Last, the mixed methods intervention design employed for the study is relatively rare in academic writing research.

Positive impact on writing skills

Though the CG and the EG students demonstrated improved writing skills during the period under consideration, the EG's performance was significantly better than their CG peers across two post-intervention tests. Though there is little research on the impact of ChatGPT on students' academic writing skills, the findings are consistent with those reported by researchers who focused on the impact of AI-driven AWE tools on tertiary-level students' writing skills (Marzuki et al., 2023; Zhao, 2022). Many existing studies utilized these tools for formative purposes (Rudolph et al., 2023) and found similar results. Thus, the employment of generative AI only strengthens those claims. The positive impact was evident in students' achievements in terms of generation of focused ideas, better connection among ideas and sentences, and improved grammatical accuracy, which were also claimed by previous researchers (Allagui, 2023; Kohnke et al., 2023; Su et al., 2023; Wang & Guo, 2023). The delayed post-test results in this study add to the generalizability of the positive impact claims (Rose et al., 2020). In fact, the literature on delayed post-tests highlights the sustainability of the impact and retention of writing skills (Rezai et al., 2022). The continued positive impact could be a result of various factors. First, planned training was organized for students before the intervention was undertaken, as advised by researchers who undertook similar interventions (Mathur & Mahapatra, 2022). Second, students were engaged in SA and PA, which have been proven effective strategies in writing classrooms (Mathur & Mahapatra, 2022). Third, the metalinguistic explanations as part of the corrective feedback could have made a difference, as claimed by many previous researchers (Kohnke et al., 2023). Fourth, the instant and personalized nature of the obtained feedback could have strongly impacted the continued positive impact (Gayed et al., 2022). Fifth, self-correction was made easy (Dai et al., 2023) Last, a major factor that might have shaped EG students' performance in the study is their pre-intervention proficiency levels. They had to pass a challenging test of English to join the institution. Thus, future researchers may investigate language proficiency as a variable when determining the impact of ChatGPT. The fear regarding the loss of creativity and the imposition of a pre-decided pattern is an addition to the fear regarding the loss of ethicality and such issues reported by Yan (2023). Several other factors, like the affordances of ChatGPT as a writing tool and its ability to engage students in a dialogic feedback process, backed by theories of Barrot (2023) and Winstone and Carless (2020), also strengthened the theoretical foundations underpinning the use of ChatGPT. Through the empirical evidence on the utility of ChatGPT as a formative feedback tool in the academic writing classroom, this study establishes that the integration

of ChatGPT into the academic writing instruction can yield positive impacts when the instructor is aware of the affordances of ChatGPT and knows how to guide students about its utilities. It also proves that the dialogic nature of ChatGPT can be fully put to use when students are adequately prepared.

Students' positive outlook

The findings from the analysis of FGD data indicate an overall positive attitude towards the impact of ChatGPT. This finding is consistent with the literature on students' views on the impact of ChatGPT and other AI-driven tools on their writing (Marzuki et al., 2023; Yan, 2023). The overall positive attitude can be explained by the enthusiasm for using ChatGPT in the classroom. The findings on how ChatGPT aids content generation align with Gayed et al.'s (2022), Guo et al.'s (2022) and Marzuki et al.'s (2023) claims. In fact, this confirms assertions in the education literature on ChatGPT. Staying focused on the topic is an added advantage which could be new to the literature.

On the other hand, the promotion of learner autonomy and peer collaboration are similar to findings in the AWE literature (Dai et al., 2023; Rudolph et al., 2023). The advantages of faster task completion and creating a more robust topic, and concluding sentences are attractive, as they are relatively new to the AWE and ChatGPT literature. They may need more microscopic investigation.

In terms of organization, the perceived impact confirms the findings of Allagui (2023) and Marzuki et al. (2023), who highlight the help with the organization of content in their study. Thi and Nikolov's (2022) study on Grammarly reports negative perceptions related to organization, with which the current study disagrees. The reasons for this kind of perception may have to do with their ability to ask appropriate questions. It will be interesting to examine whether students with low proficiency levels can benefit from ChatGPT in terms of the organization of their content. Another overwhelming opinion that emerged from the study is that ChatGPT improves grammatical accuracy, which is in line with the claims by Ohio University (2023) and Wang and Guo (2023). Though it is an expected finding that strengthens results reported in Grammarly studies (Fitriana & Nurazani, 2022), the highlight here is the attention paid by students to the explicit metalinguistic feedback, which is also found in AWE literature (Kohnke et al., 2023). However, the lack of attention to accuracy-related details may need longitudinal-focused inquiries.

Conclusion

The study was an attempt to investigate the extent to which the employment of ChatGPT as a feedback tool impacts ESL students' academic writing skills. It could be one of the earliest intervention-based empirical inquiries into the impact of ChatGPT on students' academic writing skills. Conducted as a mixed methods intervention study, the study's findings were on expected lines. The significant positive impact coupled with students' positive perception of the same can add a fresh perspective to the literature on ChatGPT and other AWE tools. The findings of the study add to the literature on AWE, especially the use of generative AI. They strengthen and further theories of feedback as a dialogic tool and ChatGPT as a formative feedback tool that can be harmoniously integrated into large writing classes. The findings of the study demonstrate that interactions

facilitated by ChatGPT during the process of writing has a direct positive impact on student learning. It positively influences how students seek feedback, how they engage with it and how they make improvements in their academic writing. It enables students to overcome the anxiety involved in asking for and receiving the desired kind of feedback. In large classes, where feedback-related dialogue is a thorny issue, ChatGPT can be a potential game-changer. It can provide tailored feedback that goes beyond the barriers of language, time, and place. It may be safely claimed that students who may not be very proficient in English can ask for and receive help in their own language on ChatGPT. The study also makes a case for ChatGPT as a formative feedback tool that can drive writing forward through SA and PA. In a way, ChatGPT breaks the barrier between SA and PA and strengthens the role of a teacher as a facilitator because many time-consuming tasks in large size writing classrooms such as monitoring content, organization, vocabulary use and grammatical accuracy can be easily performed by ChatGPT.

The study establishes the potential of ChatGPT as a pedagogic tool for writing classrooms, especially in many Global South countries where students have access to portable computing devices and the Internet. It can be easily integrated into the regular teaching of academic writing skills in institutions of higher education. A major factor that needs mentioning is the teacher's attitude towards ChatGPT and their ability to use it in a constructive manner in a large size classroom. The latter one includes self-and student-training. It is true that many webinars and workshops on the use of ChatGPT are being conducted for teachers working in Global South countries like India. However, without proper reflective planning and an analysis of the need for shunning traditional feedback strategies, the use of ChatGPT may not be as impactful. Thus, teacher education programs need to orient teachers towards utilizing ChatGPT in their writing classrooms.

Methodologically, using mixed methods intervention design, a potent way of conducting educational experiments, can be a significant addition to the applied linguistics literature. Though it is one of the first attempts to empirically investigate the impact of ChatGPT on students' academic writing skills, the study has a few limitations. First, it focused on only three writing genres. Second, the intervention lasted for only six hours. Third, artefacts from students were not included in the study. Since the study used an intact classroom, which adds to the authenticity and validity of sampling, it was impossible to go beyond the prescribed syllabus, focus on more components and continue the intervention for more hours. The artefacts could have added to the study, but privacy and copyright issues were difficult to overcome. It may be difficult to ignore how the limitations could have shaped the findings. The findings could be true for only three genres, but a major genre like argumentative writing got ignored in the process. Thus, any generalization should be carefully worded. Also, the absence of artefacts hurts the validity of claims. The findings are entirely based on the tests and students' opinions which could have been strengthened by the collection and analysis of artefacts from the classroom. Future researchers can investigate the impact of ChatGPT using an extended intervention period. More investigation is required to compare its impact on various writing genres and across various micro features. Another exciting area could be the impact of corrective metalinguistic written feedback given through ChatGPT on students' writing

skills. With ethical issues looming large, the future of writing research will nonetheless be dominated by generative AI writing tools like ChatGPT.

Appendix

Process writing tasks

1. Write a paragraph in 150 words describing the procedure for conducting an experiment using two spring balances to verify Newton's third law of motion.
2. Write a paragraph in 150 words describing the procedure for conducting an experiment in the laboratory to study Archimedes' principle.
3. Write a paragraph in 150 words describing the procedure for plotting a cooling curve that indicates the relationship between a hot object and the time taken by it to cool down.

Comparison writing tasks

1. Write a paragraph in 150 words comparing the features of iOS and android operating system.
2. Write a paragraph in 150 words comparing the properties of metals and non-metals.
3. Write a paragraph in 150 words comparing renewable and non-renewable resources.

Cause-effect writing tasks

1. Write a paragraph in 150 words describing the impact of regular exercise on our mental health.
2. Write a paragraph in 150 words describing the impact of climate change on the environment.
3. Write a paragraph in 150 words describing the impact of technology addiction on mental health.

References

Ali, J. K. M., Shamsan, M. A. A., Hezam, T. A., & Mohammed, A. A. Q. (2023). Impact of ChatGPT on learning motivation: Teachers and students' voices. *Journal of English Studies in Arabia Felix, 2*(1), 41–49. https://doi.org/10.56540/jesaf.v2i1.51

Allagui, B. (2023). Chatbot Feedback on students' writing: Typology of comments and effectiveness. In O. Gervasi, B. Murgante, A. M. A. C. Rocha, C. Garau, F. Scorza, Y. Karaca, & C. M. Torre (Eds.), *International conference on computational science and its applications* (pp. 377–384). Springer. https://doi.org/10.1007/978-3-031-37129-5_31

Anderson, J. A., & Ayaawan, A. E. (2023). Formative feedback in a writing programme at the University of Ghana. In A. Esimaje, B. van Rooy, D. Jolayemi, D. Nkemleke, & E. Klu (Eds.), *African perspectives on the teaching and learning of English in higher education* (pp. 197–213). Routledge.

Barrot, J. S. (2023). Using ChatGPT for second language writing: Pitfalls and potentials. *Assessing Writing, 57*, 100745. https://doi.org/10.1016/j.asw.2023.100745

Bonner, E., Lege, R., & Frazier, E. (2023). Large language model-based artificial intelligence in the language classroom: Practical ideas for teaching. *Teaching English with Technology, 23*(1), 23–41. https://doi.org/10.56297/BKAM1691/WIEO1749

Bozorgian, H., & Yazdani, A. (2021). Direct written corrective feedback with metalinguistic explanation: Investigating language analytic ability. *Iranian Journal of Language Teaching Research, 9*(1), 65–85. https://doi.org/10.30466/IJLTR.2021.120976

Butterfuss, R., Roscoe, R. D., Allen, L. K., McCarthy, K. S., & McNamara, D. S. (2022). Strategy uptake in writing Pal: Adaptive feedback and instruction. *Journal of Educational Computing Research, 60*(3), 696–721. https://doi.org/10.1177/07356331211045304

Chan, C. K. Y., & Hu, W. (2023). Students' voices on generative AI: Perceptions, benefits, and challenges in higher education. *International Journal of Educational Technology in Higher Education, 20*, 43. https://doi.org/10.1186/s41239-023-00411-8

Chang, T. S., Li, Y., Huang, H. W., & Whitfield, B. (2021). Exploring EFL students' writing performance and their acceptance of AI-based automated writing feedback. In *2021 2nd International conference on education development and studies* (pp. 31–35). https://doi.org/10.1145/3459043.3459065

Creswell, J. W., & Plano Clark, V. L. (2017). *Designing and conducting mixed methods research* (3rd ed.). Sage Publications.

Dai, W., Lin, J., Jin, F., Li, T., Tsai, Y., Gasevic, D., & Chen, G. (2023). Can large language models provide feedback to students? A case study on ChatGPT. https://doi.org/10.35542/osf.io/hcgzj

Fathi, J., & Khodabakhsh, M. R. (2019). The role of self-assessment and peer-assessment in improving writing performance of Iranian EFL students. *International Journal of English Language and Translation Studies, 7*(3), 1–10.

Fitria, T. N. (2023). Artificial intelligence (AI) technology in OpenAI ChatGPT application: A review of ChatGPT in writing English essay. *ELT Forum: Journal of English Language Teaching, 12*(1), 44–58. https://doi.org/10.15294/elt.v12i1.64069

Fitriana, K., & Nurazni, L. (2022). Exploring English department students' perceptions on using Grammarly to check the grammar in their writing. *Journal of English Teaching, 8*(1), 15–25. https://doi.org/10.33541/jet.v8i1.3044

Gao, C. A., Howard, F. M., Markov, N. S., Dyer, E. C., Ramesh, S., Luo, Y., & Pearson, A. T. (2022). Comparing scientific abstracts generated by ChatGPT to original abstracts using an artificial intelligence output detector, plagiarism detector, and blinded human reviewers. *BioRxiv.* https://doi.org/10.1101/2022.12.23.521610

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

Golzar, J., Momenzadeh, S. E., & Miri, M. A. (2022). Afghan English teachers' and students' perceptions of formative assessment: A comparative analysis. *Cogent Education, 9*(1), 2107297. https://doi.org/10.1080/2331186X.2022.2107297

Gopalan, M., Rosinger, K., & Ahn, J. B. (2020). Use of quasi-experimental research designs in education research: Growth, promise, and challenges. *Review of Research in Education, 44*(1), 218–243. https://doi.org/10.3102/0091732X20903302

Guo, K., Wang, J., & Chu, S. K. W. (2022). Using chatbots to scaffold EFL students' argumentative writing. *Assessing Writing, 54*, 100666. https://doi.org/10.1016/j.asw.2022.100666

Haristiani, N. (2019). Artificial Intelligence (AI) chatbot as language learning medium: An inquiry. *Journal of Physics: Conference Series, 1387*(1), 1–7. https://doi.org/10.1088/1742-6596/1387/1/012020

Huisman, B., Saab, N., van den Broek, P., & van Driel, J. (2019). The impact of formative peer feedback on higher education students' academic writing: A meta-analysis. *Assessment & Evaluation in Higher Education, 44*(6), 863–880. https://doi.org/10.1080/02602938.2018.1545896

Imran, M., & Almusharraf, N. (2023). Analyzing the role of ChatGPT as a writing assistant at higher education level: A systematic review of the literature. *Contemporary Educational Technology, 15*(4), ep464. https://doi.org/10.30935/cedtech/13605

Kohnke, L., Moorhouse, B. L., & Zou, D. (2023). ChatGPT for language learning and teaching. *RELC Journal*. https://doi.org/10.1177/003368822311628

Koltovskaia, S. (2020). Student engagement with automated written corrective feedback (AWCF) provided by Grammarly: A multiple case study. *Assessing Writing, 44*, 100450. https://doi.org/10.1016/j.asw.2020.100450

Lam, R. (2018). Feedback in writing portfolio assessment. In R. Lam (Ed.), *Portfolio assessment for the teaching and learning of writing* (pp. 59–72). Springer.

Lingard, L. (2023). Writing with ChatGPT: An illustration of its capacity, limitations & implications for academic writers. *Perspectives on Medical Education, 12*(1), 261–270. https://doi.org/10.5334/pme.1072

Loewen, S., & Plonsky, L. (2017). *An A-Z of applied linguistics research methods*. Bloomsbury Publishing.

Mahapatra, S. K. (2021). Online formative assessment and feedback practices of ESL teachers in India, Bangladesh and Nepal: A multiple case study. *Asia-Pacific Education Researcher, 30*(6), 519–530. https://doi.org/10.1007/s40299-021-00603-8

Marzuki, S., Widiati, U., Rusdin, D., Darwin, R., & Indrawati, I. (2023). The impact of AI writing tools on the content and organization of students' writing: EFL teachers' perspective. *Cogent Education, 10*(2), 2236469. https://doi.org/10.1080/2331186X.2023.2236469

Mathur, M., & Mahapatra, S. (2022). Impact of ePortfolio assessment as an instructional strategy on students' academic speaking skills: An experimental study. *CALL-EJ, 23*(3), 1–23.

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

Ohio University. (2023). *ChatGPT and teaching and learning*. https://www.ohio.edu/center-teaching-learning/resources/chatgpt

Olsen, T., & Hunnes, J. (2023). Improving students' learning—The role of formative feedback: Experiences from a crash course for business students in academic writing. *Assessment & Evaluation in Higher Education*. https://doi.org/10.1080/02602938.2023.2187744

OpenAI. (2022). ChatGPT: Optimizing language models for dialogue. *OpenAI*. https://openai.com/blog/chatgpt/

Perry, F. L., Jr. (2011). *Research in applied linguistics: Becoming a discerning consumer*. Routledge.

Rad, H. S., Alipour, R., & Jafarpour, A. (2023). Using artificial intelligence to foster students' writing feedback literacy, engagement, and outcome: A case of Wordtune application. *Interactive Learning Environments*. https://doi.org/10.1080/10494820.2023.2208170

Rezai, A., Naserpour, A., & Rahimi, S. (2022). Online peer-dynamic assessment: an approach to boosting Iranian high school students' writing skills: A mixed-methods study. *Interactive Learning Environments*. https://doi.org/10.1080/10494820.2022.2086575

Rodríguez, M. F., Nussbaum, M., Yunis, L., Reyes, T., Alvares, D., Joublan, J., & Navarrete, P. (2022). Using scaffolded feedforward and peer feedback to improve problem-based learning in large classes. *Computers & Education, 182*, 104446. https://doi.org/10.1016/j.compedu.2022.104446

Rose, H., McKinley, J., & Baffoe-Djan, J. B. (2020). *Data collection research methods in applied linguistics*. Bloomsbury Academic.

Rudolph, J., Tan, S., & Tan, S. (2023). ChatGPT: Bullshit spewer or the end of traditional assessments in higher education? *Journal of Applied Learning & Teaching, 6*(1), 342–363. https://doi.org/10.37074/jalt.2023.6.1.9

Stokel-Walker, C. (2022). AI bot ChatGPT writes smart essays—Should professors worry? *Nature*. https://doi.org/10.1038/d41586-022-04397-7

Su, Y., Lin, Y., & Lai, C. (2023). Collaborating with ChatGPT in argumentative writing classrooms. *Assessing Writing, 57*, 100752. https://doi.org/10.1016/j.asw.2023.100752

Tai, A. M. Y., Meyer, M., Varidel, M., Prodan, A., Vogel, M., Iorfino, F., & Krausz, R. M. (2023). Exploring the potential and limitations of ChatGPT for academic peer-reviewed writing: Addressing linguistic injustice and ethical concerns. *Journal of Academic Language and Learning, 17*(1), T16–T30.

Taskiran, A., & Goksel, N. (2022). Automated feedback and teacher feedback: Writing achievement in learning English as a foreign language at a distance. *Turkish Online Journal of Distance Education, 23*(2), 120–139. https://doi.org/10.17718/tojde.1096260

Thi, N. K., & Nikolov, M. (2022). How teacher and Grammarly feedback complement one another in Myanmar EFL students' writing. *The Asia-Pacific Education Researcher, 31*(6), 767–779. https://doi.org/10.1007/s40299-021-00625-2

Tracy, S. J. (2019). *Qualitative research methods: Collecting evidence, crafting analysis, communicating impact*. Wiley.

Wang, M., & Guo, W. (2023). The potential impact of ChatGPT on education: Using history as a rearview mirror. *ECNU Review Education*. https://doi.org/10.1177/20965311231189826

Winstone, N., & Carless, D. (2020). *Designing effective feedback processes in higher education*. Routledge.

Yamashita, T. (2021). Corrective feedback in computer-mediated collaborative writing and revision contributions. *Language Learning & Technology, 25*(2), 75–93.

Yan, D. (2023). Impact of ChatGPT on learners in a L2 writing practicum: An exploratory investigation. *Educational Information Technology*. https://doi.org/10.1007/s10639-023-11742-4

Zhang, F., Schunn, C., Chen, S., Li, W., & Li, R. (2023). EFL student engagement with giving peer feedback in academic writing: A longitudinal study. *Journal of English for Academic Purposes, 64*, 101255. https://doi.org/10.1016/j.jeap.2023.101255

Zhao, X. (2022). Leveraging artificial intelligence (AI) technology for English writing: Introducing wordtune as a digital writing assistant for EFL writers. *RELC Journal*. https://doi.org/10.1177/00336882221094089

Zhu, M., Liu, O. L., & Lee, H. S. (2020). The effect of automated feedback on revision behavior and learning gains in formative assessment of scientific argument writing. *Computers & Education, 143*, 103668. https://doi.org/10.1016/j.compedu.2019.103668

Publisher's Note

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

Santosh Mahapatra   The author is an Associate Professor at BITS Pilani Hyderabad Campus, India. He has published in reputed international journals, reviewed manuscripts for several international journals and worked on many national and international funded research projects. His research interests are learning technology, teacher development and educational language policy.