UNCOLLCS
Proceeding of Undergraduate Conference on Literature, Linguistic, and Cultural Studies
E-ISSN: 2985-9476
Published by Fakultas Ilmu Budaya, Untag Surabaya

The impact of ChatGPT formative feedback on EFL learners' narrative essay writing

Mas Zia Ahdaan Ta'ajuddin Hidayat
Universitas 17 Agustus 1945 Surabaya, Indonesia
masziaahdaant@gmail.com

Pariyanto Pariyanto
Universitas 17 Agustus 1945 Surabaya, Indonesia
pariyanto@untag-sby.ac.id

Abstract This study uses quantitative research design, and aims to analyze the impact of ChatGPT formative feedback in improving the quality of EFL learners' narrative essay writing. The 16 participants are English Literature students of Universitas 17 Agustus 1945 Surabaya. In rating the pre-feedback and post-feedback narrative essays, the researcher serves as the first rater, accompanied by another rater, as well as ChatGPT automated feedback as means of comparison. In rating the essays, a narrative essay rubric from City University of New york is employed as a guide. The result of this research shows a significant improvement in the rating of the essays after receiving feedback from ChatGPT. This is based on the Sig. value of paired sample test of .000, meaning a significant improvement is evident. Moreover, the interrater reliability between the raters is also high, as Cronbach's Alpha shows a result of .912, indicating a value close to 1, meaning the rating process was reliable.

Keywords: *ChatGPT, EFL Learners, Formative feedback, Essay, Interrater Reliability*.

INTRODUCTION

The advancement in technology has brought us the invention of a large language model called ChatGPT by OpenAI, which was made public in November 30, 2022. As a large language model, ChatGPT covers the functions of automated generation of text, responding to questions, and doing tasks such as summarization and translation (Agomuoh, 2023). These functions are capable of being utilized within the realm of education, as they can be used to enhance writing skills, since ChatGPT has the capability of delivering feedback on style, grammar, and coherence (Aljanabi et al., 2023). This is supported by the previous studies stating that ChatGPT is able to generate essays that are of high quality on diverse range of topics (Hoang et al., 2023). Usage of ChatGPT in a realm of education also brought forth some conundrum in evaluating academic tasks such as students' essay writing (Stokel-Walker, 2022; Whitford, 2022). It is often a belief that writing skill is hard to master, especially writing in a new language (Blanchard & Root, 1998; Mundriyah & Parmawati 2016). Writing skill is the hardest skill to be learned because it is a production skill, which requires feedback (Harmer, 1991; Mohammed, S. I., 2021).

Feedback as a means to enhance learning is one of the most crucial things to have (Banihashem et al., 2022). The generally accepted meaning which heavily adheres to what a feedback means, is an information provided by an agent, e.g., peer, teacher, self, or technology about reactions to a person's performance or a product. A feedback that is formative, also acknowledged as formative feedback, can bring a positive and lasting influence on students' performance and learning. In English writing, formative feedback strengthens students' awareness on the areas which they can strive to improve. Formative feedback strategies such as teacher, peer, or even self-assesment can be used to address students' need for explanation (Bozorgian & Yazdani, 2021; Zhang et al., 2023), concerning aspects of their English writing such as grammar, word choices, organization, and content. It is no easy task, however, as in Indonesia, it is such as a challenge to assess individual student's ability in writing essays in English, as it is too time-consuming to be done within a short period of time, especially due to the high number of students in each class (Applebee & Langer, 2011; Graham, 2019). Since the demand for too much time can be such a restriction, it leaves some way for self-assessment as a form of feedback. Self-assessment can be beneficial to help students be self-aware of the potential weakness they have in their English writing. This duty should fall within the responsibilities of the teacher, as they are expected to be experts in the subject matter, thus being able to provide a competent positive input on the writing. However, due to the nature of individual assessment, this can be very demanding time-wise. Therefore, peer and self-assessment seem plausible to be an alternative. The involvement of students in the process of feedback can augment their awareness, active engagement, and motivation for learning (Arguedas et al., 2016). This situation, however, calls upon the possibilities of student not being able to properly address their own mistakes in writing, as they might not be aware of what the errors are in the first place.

In response to this, many approaches have emerged, among which is computer-assisted feedback. In recent times, technology has advanced enough that a personalized feedback generated by computer tools have proved a promising value in improving feedback practices by enabling measurable, timely, and tailored feedback. (Banihashem et al., 2023; Deeva et al., 2021; Drachsler, 2023; Drachsler & Kalz, 2016; Pardo et al., 2019; Zawacki-Richter et al., 2019; Rüdian et al., 2020). If such a useful computer-assisted feedback could be a positive addition to teacher feedback, then it should prove suitable in rectifying the time constraint challenge faced by teachers in giving feedback to the students.

In the context of feedback and computer tools, the recent stride forward in the field of technology has provided us with the novel invention of Artificial Intelligence, also acknowledged as A.I. One of the most recent Artificial Intelligence tools is known as "ChatGPT", which has raised numerous talks about its possibility to influence the current state of education (Ray, 2023). The introduction of ChatGPT has initiated discussions on the number of ways A.I. can positively affect education as a whole (Bond et al., 2024; Darvishi et al., 2024). As of recent, it can be argued that ChatGPT is capable of making the process of writing simpler and quicker (Stokel-Walker, 2022).

It is possible to take advantage of ChatGPT in the realm of education, as it provides what is called as A.I.-generated feedback (Farrokhnia et al., 2023. It assists students in their writing by offering relevant guidance on content and structure as they compose their work. (Allagui, 2023). There are many forms of writing which ChatGPT can provide feedback on.

One of the more prevalent forms of writing is essay. An essay is a short piece of writing about a particular topic or subject. There are many types of essays, but essays often fall within four main categories: argumentative, expository, descriptive, and narrative. As one

of the most commonly written essays, narrative essay is a writing that recounts a personal experience, usually one that taught the author a life lesson
that is important. According to SI, Mohammed (2021), narrative essays encompass three mandatory elements; character, theme, and dialogue.
Within the context of essay writing and AI-generated feedback, a previous study conducted by Hasman et al., (2023) on the impact of implementation of ChatGPT in students' essay writing skills, has stumbled upon this topic. Upon inspecting the results, it is evident that due to the intervention of ChatGPT during pre-test and post-test, the total scores of the essays have increased from 846 to 1053. A significant improvement is apparent, enforced by a p-value of less than 0.05 (0.015).
Likewise, earlier study conducted by Mahapatra (2024) about the impact of ChatGPT on ESL students' academic writing skill, further enforces this notion. The results shown in the research have demonstrated the beneficial effects of feedback given by ChatGPT in the field of academic writing, as well as the overall positive attitudes toward the use of the AI tool.
In addition to that, students' perception of using the OpenAI ChatGPT application in improving writing skills are further explored in J Zebua & Katemba (2024)'s study. The results showed that most respondents held a generally favorable view of using OpenAI ChatGPT to improve writing skills.
Considering the background of the current research, it is hoped that it can showcase whether or not the use of ChatGPT formative feedback can be of help for students in raising their awareness on their own essay writing, as well as rectifying the students' mistakes in writing. In the context of ChatGPT usage in the realm of education, this research also carries a potential contribution in the matter, as it delivers insights about the quality of AI-generated feedback within the domain of EFL college students' essay writing.

LITERATURE REVIEW

ChatGPT

ChatGPT is a language model developed by OpenAI which uses deep learning to produce human-like text based on the input it receives (OpenAI, 2020). ChatGPT encompasses the functions of text generation, answering questions, and doing tasks such as summarization and translation (Agomuoh, 2023). These functions are capable of being utilized within the realm of education, as they can be used to improve writing skills, since ChatGPT is trained to deliver feedback on style, grammar, and coherence (Aljanabi et al., 2023).
In the context of writing within an educational setting, ChatGPT, such as the free version of 3.5, has garnered significant attention due to their capability and potential in revolutionizing the automated generation of educational materials and language assessments. (Meida, et al. 2023). This premise is in line with its ability to provide feedback on students' writing quality.

Feedback

Feedback is a formative input which addresses a product, work, or performance of someone or something. The main goals of a feedback are to enhance knowledge and skill acquisition, as well as reduce errors (Nelson et, al. 2021). Feedback is a powerful tool to improve learning outcomes. Feedback might come from the works, instructors, other students, or a machine. In education, feedback plays a crucial role by informing students

about their assessment performance and the reasons for their grades (Pariyanto, 2021; Silalahi and Pariyanto, 2021; Pariyanto and Pradipta, 2020). It is instrumental in learning as it provides students with guidance on enhancing their academic abilities. (Donnelly & Kirk, 2010).

Second Language Acquisition

According to Richards, J, (1985) in Khasinah (2014), Second Language Acquisition refers to the process by which people develop proficiency in a second or foreign language. In essence, Second Language Acquisition (SLA) is the acquisition of a second or additional languages after the first language acquisition is under way or completed (Fromkin, V 2002). In the context of SLA, Ellis (2008) posits that behaviorists view language learning as environmentally controlled by various stimulus and feedback that language learners are exposed to as language input.

Narrative Essay

J, Sexton (2019) defines narrative essay as a writing that recounts a personal experience, usually one that taught the author a life lesson that is important. According to SI, Mohammed (2021), narrative essays encompass three mandatory elements; character, theme, and dialogue.

Narrative Essay Rubric

A rubric is a multi-purpose rating guide for assessing student products and performances (K, Wolf. 2007). Knoch (2011) describes the two vital points of a rubric, which are the criteria themselves and the descriptions of how well they were met. In the context of narrative essay, rubric is used as a guide in the rating process during the assessment of the essays. Referring to the narrative essay rubric from City University of New York (2021), there are several unique criteria of writing quality included, along with the descriptions for parameter:

1. Ideas (20 points)
    a. Essay merely presents events without a clear purpose or theme.
    b. While a theme occasionally surfaces, readers are often left questioning the purpose of the events and the story itself.
    c. Essay has a theme that connects all events but doesn't effectively relate it to the reader.
    d. Essay has a focused purpose indicating the story's significance to the reader, centered around a central theme that comments on human experience or society.

2. Style (20 points)
    a. Rare and ineffective use of narrative techniques.
    b. Limited narrative techniques fail to emphasize the theme.
    c. Includes some narrative and figurative language but primarily follows a linear narrative. Attempts to convey thematic importance.
    d. Engages readers with narrative techniques and figurative language (foreshadowing, dialogue, imagery, etc.) effectively enhancing understanding of the theme.

3. Plot (20 points)
    a. Events are described vaguely without a logical conclusion or clear beginning/ending.

   b. Some events seem disconnected from the conclusion, lacking description. Beginning and ending are predictable with unanswered questions.
   c. Events build logically to support the theme, adequately described with an engaging beginning and effective ending.
   d. Events interact cohesively with the theme, major and minor events intertwined to a natural conclusion. Beginning informs readers, and the ending addresses the theme, resolving loose ends satisfactorily.

4. Voice (20 points)
   a. Fails to connect or recognize the reader. Personal details don't engage readers.
   b. Attempts to persuade readers of the essay's importance but with ineffective details.
   c. Provides reasons for the theme's importance, detailing events and experiences.
   d. Recognizes the audience through characters and events, vividly detailing experiences to enhance reader understanding.

5. Creativity (10 points)
   a. Essay lacks creativity and fails to engage readers.
   b. Attempts at surprises or imagery often confuse readers.
   c. Includes surprises, conflicts, and imagery to some extent.
   d. Incorporates surprises, conflicts, and imagery effectively to deepen reader understanding.

6. Transitions and Sentence Structure (5 points)
   a. Lacks transitions and varied sentence structure, causing confusion and hindering readability.
   b. Poor transitions between paragraphs and lack of sentence variety limit reader engagement.
   c. Well-organized paragraphs with effective transitions and varied sentence structures maintain reader engagement.
   d. Naturally builds upon transitions and narrative devices, using varied sentence structures to ensure clear organization and reader engagement.

7. Conventions (5 points)
   a. Missing MLA format, numerous spelling, and punctuation errors indicate lack of proofreading.
   b. Attempts MLA format but does not follow guidelines, with many errors causing confusion.
   c. Maintains proper MLA format and standard English, addresses spelling and punctuation errors with some mistakes.
   d. Consistently maintains proper MLA format and standard English, effectively addresses spelling and punctuation errors.

The combinations of rating parameter above accumulate to a total of 100 points.

METHOD

This research employed a quantitative research design. According to Adedoyin (2020), quantitative research is the study of phenomena using numerical data and statistical, analytical, or computational tools. A quantitative research method was employed in this study in order to quantify and analyze quantitative data, exploring the efficacy of which ChatGPT AI-generated feedback influence the English writing of English Literature college students.

The sources of the data will include the results of the narrative essays written by the 16 participants of English Literature students in Universitas 17 Agustus 1945 Surabaya, 5 of which are females, and 15 of which are 8th semester students, with only 1 participant in 6th semester. All the participants are required to having passed Paragraph Writing and Essay Writing. The data also includes the revised version after receiving feedback from ChatGPT. The researcher collects and analyzes the data from the pre-feedback and post-feedback essays.

The research instruments involved in this study are: narrative essays, ChatGPT, narrative essay rubric, researcher, Microsoft Excel, and SPSS.

The data collection involves giving a task to the 16 participants, which is to write short narrative essays narrating their childhood memories, which will then be submitted to ChatGPT along with the prompt "please provide a formative feedback to improve the quality of the following essay".

Data analysis procedure in this study involved several steps. First, upon collecting the essays, they are then put into ChatGPT with the above prompt. Second, after receiving feedback from ChatGPT and obtaining the revised version from the participants, the data are put in Excel spreadsheet for organization. Third, the researcher and another human rater, as well as ChatGPT rate the essays based on the narrative essay rubric from The City University of New York (2021). Fourth, the data are put in SPSS to calculate the results to be presented. Fifth, upon obtaining all the results, conclusions are drawn.

RESULT AND DISCUSSION

The analysis of the impact of ChatGPT AI-generated formative feedback on the essay written by EFL English Literature students is conducted through a task of writing a narrative essay about their childhood memories. The results reveal notable improvements in some aspects of their writing.

It has been found that, initially, based on the narrative essay rubric, EFL students of varying demographics demonstrated some expected common issues such as lack of clear paragraphing, inadequate attempt in putting imagery in their narrative essays, and frequent grammatical errors. However, after the usage of ChatGPT's formative feedback, the quality of the paragraphs, imagery, and grammatical errors within their narrative essays have somewhat improved. The rating process is done by the researcher, who is assisted by another rater. Automated rating from ChatGPT is also employed, so as to minimize subjectivity, as well as serving as an alternative. The raters refer to the narrative essay rubric for the several criteria for writing narrative essays.

Pre-feedback Essay Scores

Table 1. Pre-feedback Essay Ratings

	Rater 1	Rater 2	ChatGPT
Essay 1	51	49	48
Essay 2	64	45	64

Essay 3	67	65	68
Essay 4	87	73	78
Essay 5	68	54	67
Essay 6	76	59	65
Essay 7	74	69	61
Essay 8	71	63	74
Essay 9	85	86	88
Essay 10	67	62	65
Essay 11	58	60	62
Essay 12	77	58	65
Essay 13	77	60	67
Essay 14	73	66	65
Essay 15	66	57	63
Essay 16	54	48	56

Post-feedback Essay Scores

Criteria	Rater 1	Rater 2	ChatGPT
Ideas	15.0625	12.6875	14.375
Style	13.1875	12.75	12.75
Plot	14.4375	12.8125	13.75
Voice	14.5625	11.4375	14.0625
Creativity	6.1875	5.875	6.4375
Transitions and Sentence	3.25	2.6875	2.5625
Conventions	3	2.6875	2.0625
Mean	69.6875	60.9375	66.0000
Total	1115	975	1056

Table 2. Post-feedback Essay Ratings

	Rater 1	Rater 2	ChatGPT
Essay 1	53	58	48
Essay 2	72	50	64
Essay 3	74	70	68
Essay 4	89	77	78
Essay 5	74	59	67
Essay 6	80	67	65
Essay 7	82	76	61

Essay 8	75	66	74
Essay 9	87	88	88
Essay 10	74	70	65
Essay 11	68	62	62
Essay 12	81	66	65
Essay 13	79	64	67
Essay 14	76	70	65
Essay 15	71	61	63
Essay 16	61	53	56

Criteria	**Rater 1**	**Rater 2**	**ChatGPT**
Ideas	15.625	12.685	15.75
Style	14.5625	12.75	13.75
Plot	14.875	12.8125	14.875
Voice	16	11.4375	15.125
Creativity	6.625	5.875	6.875
Transitions and Sentence	3.625	2.6875	3.25
Conventions	3.4375	2.6875	2.8125
Mean	74.75	66.0625	72.4375
Total	1196	1057	1159

Pre-feedback and Post-feedback Paired Sample Test Results

The results of the essays prior to and after receiving formative feedback from ChatGPT, as well as the reliability between raters are put as Paired Sample and Cronbach's Alpha in SPSS tool, and are presented below:

Rater 1

Table 3. Paired Sample Test Rater 1

Paired Samples Test

		Paired Differences					t	df	Sig. (2-tailed)
		Mean	Std. Deviation	Std. Error Mean	95% Confidence Interval of the Difference Lower	Upper			
Pair 1	PreFeedbackHuman1 - PostFeedbackHuman1	-5.06250	2.56824	.64206	-6.43102	-3.69398	-7.885	15	.000

Rater 2

Table 4. Paired Sample Test Rater 2

Paired Samples Test

		Paired Differences					t	df	Sig. (2-tailed)
		Mean	Std. Deviation	Std. Error Mean	95% Confidence Interval of the Difference Lower	Upper			
Pair 1	PreFeedbackHuman2 - PostFeedbackHuman2	-4.81250	2.19754	.54938	-5.98348	-3.64152	-8.760	15	.000

ChatGPT

Table 5. Paired Sample Test ChatGPT

Paired Samples Test

		Paired Differences					t	df	Sig. (2-tailed)
		Mean	Std. Deviation	Std. Error Mean	95% Confidence Interval of the Difference Lower	Upper			
Pair 1	PreFeedbackChatGPT - PostFeedbackChatGPT	-6.43750	7.81425	1.95356	-10.60142	-2.27358	-3.295	15	.005

Interrater Reliability

Table 6. Mean Scores between the Raters

PreFeedbackHuman1	69.6875	10.02476	16
PreFeedbackHuman2	60.9375	10.09600	16
PreFeedbackChatGPT	66.0000	8.86942	16

Table 7. Cronbach's Alpha of the Raters

Cronbach's Alpha	N of Items
.912	3

Upon observing the results of Paired Test Samples above, it is evident that the general improvements of the narrative essays written by the participants are significant after

receiving formative feedback from ChatGPT, and revising based on the received feedback. A Sig. value of less than .005 is apparent for the rating results of Rater 1 and 2, indicating a statistically significant improvement. Likewise, ChatGPT rating results have a Sig. value of .005, which is still within the range of a highly significant statistic for improvement.

Furthermore, based on the Reliability Statistics referring to Cronbach's Alpha, it can be observed that the interrater reliability between three raters is high (.912). According to IBM Corp. (2021), the closer it is to 1, the higher the reliability is between the subjects. A value of 0.7 or higher is generally accepted for research purposes. This result indicates that the consistency across the ratings from the three raters is excellent.

Overall Impact of ChatGPT's Feedback

Based on the ratings from Rater 1, the overall average score of the essays improved from 69.6875 in the pre-feedback phase to 74.75 in the post-feedback phase. This indicates a general positive impact of ChatGPT's formative feedback on students' essay writing. As the reliability among raters is high, only a single source of rating for the essays (Rater 1) is used, since the general improvement is evident. Among all the criteria, voice showed the most significant improvement, followed by style and ideas. This suggests that while ChatGPT's feedback was effective across various aspects, it was particularly influential in enhancing the stylistic elements of students' writing.

The significant improvement in voice suggests that ChatGPT's feedback was particularly effective in helping students develop a more engaging and polished writing style. This, in turn, contributed to the overall improvement observed in the other criteria, as a well-developed writing style can enhance the overall quality and impact of an essay.

CONCLUSION

Based on the results obtained from the study, it is evident that the formative feedback provided by ChatGPT has had a significant positive impact on the narrative essay writing skills of EFL learners. This is clearly demonstrated by the increase in the mean scores, which rose from 69.6875 before the feedback to 74.7500 afterward. This substantial improvement is further supported by statistical analysis, where the Sig. value obtained from the SPSS tool was found to be less than .001. Considering that a Sig. value of .005 is the minimum threshold for statistical significance, the results indicate that the improvements observed are not due to random chance but are a direct result of the formative feedback provided by ChatGPT.

The study specifically focused on EFL English Literature college students at Universitas 17 Agustus 1945 Surabaya who had passed Paragraph Writing and Essay Writing classes. Conducted within the vicinity of The Faculty of Cultural Sciences at the University, the research delved into the impact of ChatGPT's feedback on improving students' English writing abilities. Due to the specific nature and context of the study, the findings may not be universally applicable. The scope was limited to a particular group of students within a single university, which may influence the generalizability of the results.

In conclusion, the data strongly suggests that ChatGPT formative feedback is an effective mechanism for improving the narrative essay writing skills of EFL learners at Universitas 17 Agustus 1945 Surabaya. The significant increase in mean scores, backed by the statistical evidence, highlights the potential use of AI in educational settings. However, the specific focus and context of the study should be considered when interpreting the results. Incorporating ChatGPT as a supplemental feedback tool could be a valuable strategy for educators aiming to foster better writing skills among their students, particularly within similar educational environments.

REFERENCES

Banihashem, S. K., Kerman, N. T., Noroozi, O., Moon, J., & Drachsler, H. (2024). Feedback sources in essay writing: peer-generated or AI-generated feedback? *International Journal of Educational Technology in Higher Education*, *21*(1). https://doi.org/10.1186/s41239-024-00455-4

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. doi: 10.30466/ijltr.2021.120976

Farrokhnia, M., Banihashem, S. K., Noroozi, O., & Wals, A. (2024). A SWOT analysis of ChatGPT: Implications for educational practice and research. *Innovations in Education and Teaching International*, *61*(3), 460–474. https://doi.org/10.1080/14703297.2023.2195846

Graham, S. (2019). Changing How Writing Is Taught. *Review of Research in Education*, *43*(1), 277–303. https://doi.org/10.3102/0091732X18821125

Pariyanto, P. (2021). NATIVE ENGLISH SPEAKING TEACHERS (NESTs) AND INDONESIAN ENGLISH TEACHERS (IETs) EFL STUDENTS' PERCEPTION AND PREFERENCES. *Anaphora : Journal of Language, Literary, and Cultural Studies*, 3(2), 112-121. https://doi.org/10.30996/anaphora.v3i2.4620

Pariyanto, P., & Pradipta, B. (2020). FACTORS INFLUENCING AN EFL LEARNER'S PROFICIENCY: AN ENGLISH TEACHER'S PERSPECTIVE. *Anaphora : Journal of Language, Literary, and Cultural Studies*, 2(2), 89-97. https://doi.org/10.30996/anaphora.v2i2.3369

Ray, P. P. (2023). ChatGPT: A comprehensive review on background, applications, key challenges, bias, ethics, limitations and future scope. *Internet of Things and Cyber-Physical Systems*, *3*(April), 121–154. https://doi.org/10.1016/j.iotcps.2023.04.003

Ramaprasad, A. (1983). On the definition of feedback. *Behavioral Science*, *28*(1), 4–13. https://doi.org/10.1002/bs.3830280103

Silalahi, P., & Pariyanto, P. (2021). USING WHATSAPP AS AN INSTRUCTIONAL TOOL TO ENHANCE READING AND WRITING SKILLS: INDONESIAN EFL LEARNERS' PERCEPTION. Anaphora : Journal of Language, Literary, and Cultural Studies, 4(1), 79-86. https://doi.org/10.30996/anaphora.v4i1.5289

Steiss, J., Tate, T., Graham, S., Cruz, J., Hebert, M., Wang, J., Moon, Y., Tseng, W., Warschauer, M., & Olson, C. B. (2024). Comparing the quality of human and ChatGPT feedback of students' writing. *Learning and Instruction*, *91*. https://doi.org/10.1016/j.learninstruc.2024.101894

Zebua, J. A. Z., & Katemba, C. V. (2024). Students' Perceptions of Using the OpenAI ChatGPT Application in Improving Writing Skills. *Journal of Language and Literature Studies*, *4*(1), 110–123. https://doi.org/10.36312/jolls.v4i1.1805