**INTERNATIONAL PREMIER JOURNAL OF LANGUAGES & LITERATURE (IPJLL)**
**VOLUME 4, ISSUE 3 (s3), 2026 p-ISSN: 3007-2336 e-ISSN: 3007-2344**

# EFFECTIVENESS OF AI VERSUS HUMAN FEEDBACK ON SSC STUDENTS' L2 ENGLISH LANGUAGE WRITING PROFICIENCY: A COMPARATIVE ANALYSIS

| | |
|---|---|
| Shehzad<br>linguisticskhan@gmail.com | PhD Scholar, English (Linguistics), Southwest University, Chongqing, China. |
| Sawairoon Hayat<br>sawairoonk@gmail.com | MPhil Research Scholar, English (Linguistics), National University of Modern Languages, Peshawar, Pakistan. |
| Aisha Hussain<br>aishahussain93@gmail.com | MPhil English Research Scholar, Qurtaba University of Science and Technology Peshawar, Pakistan. |
| Aizaz khan<br>aizazkhanhassanzai@gmail.com | MPhil Research Scholar, English (Linguistics), Northern University, Nowshehra, Pakistan. |
| Maryam Salam<br>maryam.ali.eng@gmail.com | Lecturer, Department of English, University of Rawalpindi, Pakistan. |

**Abstract**

The integration of artificial intelligence (AI) into English language teaching provides new possibilities for timely and efficient feedback on students' writing. This study compares human teacher feedback and AI-generated feedback in assessing the L2 English language writing proficiency of Secondary School Certificate (SSC) students in Pakistan. The data was collected from purposively sampled student of SSC who secured more than 95% marks in their English paper in SSC examination conducted by BISE Mardan. A quantitative comparative research design was employed using 26 student compositions evaluated through the ESL Composition Profile developed by Jacobs et al. (1981), which measures content, organization, vocabulary, language use, and mechanics. The students were asked to write an English essay regarding their exam experience. The essay was evaluated from subject specialist in English with over 10 years of teaching experience. The same essay was evaluated from ChatGPT (AI) via predefined scoring profile. The human teacher showed common Pakistani SSC assessment practices, and focused on content coverage and communicative effectiveness, while AI applied standardized criteria with greater focus on grammatical accuracy and coherence. Descriptive statistics, paired-samples t-tests, Pearson correlation, and intraclass correlation coefficients (ICC) were used for analysis. Results showed that human teachers assigned significantly higher scores ($M = 68.42$, $SD = 8.31$) than AI feedback ($M = 61.15$, $SD = 9.02$), $t(25) = 5.84$, $p < .001$. Despite these differences, strong agreement was observed ($r = .82$; $ICC = .76$). The findings suggest that AI can effectively supplement, but not replace, human feedback in Pakistani ESL classrooms.

**Keywords:** *AI Feedback, ESL Assessment, Human Teacher Feedback, L2 Writing, SSC Students Pakistan.*

**Corresponding Author:** Aizaz khan (MPhil Research Scholar, English (Linguistics), Northern University, Nowshehra, Pakistan
**Email:** aizazkhanhassanzai@gmail.com

## 1. Introduction

Writing ability, along with the other skills of second language (L2) acquisition, is a complex skill that learners will need to master and it is very demanding because it is a skill that involves their linguistic accuracy, coherence, organization, and rhetorical effectiveness. Feedback is a crucial element of writing instruction in English as a Second Language (ESL) contexts, in which students are encouraged to be aware of errors and develop their communicative competence (Ferris, 2003; Hyland, 2003; Hyland & Hyland, 2006). A good feedback helps learners refine their texts, refine language accuracy and cultivate higher order thinking skills which lead to better writing performance.

In the past, feedback to students' writing has been done by human teachers. Human feedback is appreciated for its contextual sensitivity, pedagogical expertise, and its capability to handle both linguistic and content related problems (Ferris, 2014). Individual comments, encouragement and suggestions from teachers can help to increase learners' participation and confidence. Teachers, however, may not have the time and energy necessary to offer timely and detailed feedback due to factors such as class size, workload, and time constraints (Lee, 2017).

The field of language learning and assessment has been revolutionized by recent advancements in Artificial Intelligence (AI), especially those related to natural language processing (NLP) and generative AI technologies. The advent of Artificial Intelligence (AI) in recent years, particularly in natural language processing (NLP) and generative AI technologies, has significantly reshaped language learning and assessment practices. AI-based tools like ChatGPT and automatic writing grading can offer instant and personalized suggestions for grammar, vocabulary, flow, and sentence structure. The technologies provide opportunities for learners to self-direct their learning and engage in ongoing revision, which will lead to more efficient writing instruction (Zawacki-Richter et al., 2019). AI-assisted writing support has become more available to students and educators, with the advent of generative AI.

With the introduction of ChatGPT, academic researchers have recently become more interested in its use in education. Kasneci et al. (2023) point out that LLM have immense potential to transform the learning landscape and enable individualized learning and assessment. Similarly, Tlili et al. (2023) proposed that ChatGPT is a revolutionary tool with the potential to transform the educational landscape, but also poses a number of issues of reliability, ethics and academic integrity. Likewise, Baidoo-Anu and Ansah (2023) highlighted the potential of AI technologies to enhance learning outcomes through personalized feedback and adaptive learning experiences.

While the potential is great, researchers have warned that AI-generated feedback can be lacking in the contextual understanding, affectual support, and pedagogical judgment that a human teacher can bring. While there is no doubt that human evaluators excel at the higher-order concerns of argumentation, idea development, and rhetorical appropriateness, the new research by Stevenson and Phakiti indicates that their performance is not superior overall, particularly in terms of accuracy. (Stevenson & Phakiti, 2019). New comparative research has also confirmed that the benefits of AI and human feedback are complementary and not opposites.

While there is a growing body of literature on the use of AI in writing, past research has largely been conducted in higher education settings and with university students. However, the attention has been limited to the primary school children with very little focus placed on secondary school children, especially those in developing countries like Pakistan where English is a second language and literacy is a serious concern in education. In addition, few empirical studies have compared the feedback provided by AI and human teachers for SSC students are available.

So, the current study aims to compare the effectiveness of AI given feedback to human teacher given feedback in improving the English writing skill of Secondary School Certificate (SSC) students. This study will use the writing samples of twenty-six high-achieving SSC students to offer empirical evidence about the strengths and weaknesses of both types of feedback.This study will involve the use of writing samples from twenty-six high-achieving SSC students, independently assessed by English teachers and an AI system. The results would be expected to give further insights to the extensive pool of knowledge on language learning with the aid of AI and have implications for the use of AI technologies in ELL in Pakistan.

## 2. Literature Review

The existing research and works have shown that AI-assisted feedback is effective in improving writing. According to Mahapatra (2024), ChatGPT had a positive impact on the academic writing skills of ESL students, aiding in revision and fostering student independence. Similarly, Polakova & Ivenz (2024) found that grammatical accuracy and writing quality of students improved a lot in the presence of AI-generated feedback. Yan and Zhang (2024) also noted that L2 writers were active users of the ChatGPT-generated automated corrective feedback, which led to more effective revision processes. Li et al. (2024) also found that ChatGPT was very reliable when evaluating EFL writing and provided feedback similar to that of experienced teachers.

For instance, Mahapatra (2024) was one of the first empirical studies investigating the pedagogical effectiveness of ChatGPT in English language writing. Given the use of generative artificial intelligence in the educational context and the absence of empirical research on its effects on academic writing in language learning, the researcher aimed to investigate the effect of ChatGPT on the academic writing proficiency of ESL learners. This research used a mixed methods intervention design that included tertiary level students learning English as a second

language. Three writing tests were used to gather data at various points of the intervention as well as focus group discussions to examine what learners experiences and perceptions were about using ChatGPT. The intervention enabled students to use ChatGPT to improve their writing throughout the process as a formative feedback tool. The results of the quantitative data from the writing test showed that the learners' writing skills increased from the start of the intervention until the end. The results showed significant improvements in grammatical correctness, word use, coherence and organisation of written texts. In addition, qualitative results obtained from focus group discussions showed that students experienced ChatGPT to be a good learning companion that could offer them immediate, easy, and personalized support. The participants mentioned that the AI system was helpful for revision practice and promoted self-study.

While these are all positive results, Mahapatra (2024) highlighted that ChatGPT is not a replacement for human teachers. Some issues were raised with factual inaccuracies and excessive reliance on AI-generated ideas. Overall, the study found that generative AI has significant promise in helping learners improve their language skills, provided that it is used responsibly and within a meaningful pedagogical context. The study was conducted with university students, and only looked at the impact of AI feedback as opposed to teacher feedback, meaning that it did not address the relative effectiveness of the two. For this reason, the present study builds on the study of Mahapatra by comparing Artificial Intelligence generated and Human feedback among secondary school students in Pakistan.

Likewise, Yan and Zhang (2024) studied students' interactions with the automatic written feedback produced by ChatGPT. This study intended to get a grasp of how cognitive, behavioral, and emotional processes of second language writers related to generative artificial intelligence feedback on their L2 writing. The researchers used a mixed method multiple case study approach, gathering qualitative and quantitative data from ELLs and analyzing the revision processes of writing with ChatGPT. The study examined learner interaction with the feedback provided by ChatGPT and investigated students' use of suggestions in the rewritten compositions. Results showed that everyone interacted with the automated feedback and grammar items improved, as well as lexical choice and sentence structure. The researchers noted that the students exhibited significant behavioral engagement with their texts in terms of revising them based on the recommendations provided by the AI. Also observed was cognitive engagement, where students critically assessed the feedback and one student added comments of his own to implement suggestions. Emotional engagement also was predominantly positive, with learners stating that they appreciated the immediacy and comprehensiveness of the feedback.

However, it has been found that some students had problems in understanding some of the suggestions and sometimes accepted the revisions without critically evaluating them. Therefore, Yan and Zhang (2024) highlighted the need for teacher mediation and guidance to ensure that the learning benefits of AI feedback are maximized. They believe that AI-based

corrective feedback should be used in conjunction with teacher feedback, not in its place. Although the study provided valuable insights into learners' engagement with ChatGPT, it did not directly compare AI feedback with human evaluators. Thus, the present study is an attempt to fill this gap by comparing the students' English writing proficiency of AI-generated feedback with teacher-generated feedback in SSC students.

To explore the potential of ChatGPT to improve English as a Foreign Language (EFL) writing evaluations, Li et al. (2024) carried out a preliminary study. The first objective of the research was to investigate the reliability and efficacy of ChatGPT 3.5 and 4 in assessing students' essays, when compared to English teachers. To evaluate consistency of scoring and the quality of feedback given by the AI systems, the researchers used Generalizability Theory and qualitative analysis. The study included several essay samples created by the EFL learners, and the scores assigned by the human raters were compared to the scores assigned by ChatGPT. The results revealed that ChatGPT showed a high level of reliability and consistency in assessing written compositions. In particular, ChatGPT-4 showed higher agreement with teacher ratings compared to ChatGPT-3.5. Moreover, qualitative analysis showed that AI-generated comments were useful at finding grammatical mistakes, word usage problems and sentence level errors. But, it was determined that the more experienced teachers gave more context-sensitive and nuanced comments on the development of ideas, coherence, organization, and rhetorical appropriateness.

### 2.1. Research Gap

Based on their findings, the researchers noted that ChatGPT has a great deal to offer as an evaluator, but shouldn't be used in place of human assessors. Rather, AI must be used as an additional tool that can help teachers alleviate some of their load and offer prompt feedback. One of the weaknesses of this study was its narrow focus on higher education settings. Furthermore, it highlighted the assessment of writing, not the comparative effects of AI and human feedback on writing. Therefore, the present study aims to make a contribution to this growing body of knowledge by comparing the AI-generated feedback with teacher-generated feedback of SSC students in Pakistan.

## 3. Methodology

This study used a quantitative comparative research design with an aim of comparing the effectiveness of two kinds of feedback on the English writing skills of Secondary School Certificate (SSC) students—artificial intelligence (AI) generated feedback and human teacher feedback. Twenty-six high achieving SSC second language English students, who were purposively selected, participated in the study. Participants included students who completed a writing task in which they wrote essays on a given topic. The written compositions were independently assessed by the experienced English teachers and by the AI system, ChatGPT, based on the ESL Composition Profile developed by Jacobs et al. (1981) that examines five areas of composition: content, organization, vocabulary, language use, and mechanics, with a total score of 100 points. The human evaluator and the AI system also gave written feedback

on students' compositions, which was in addition to giving marks. The quantitative data derived from the scores was analyzed using descriptive statistics such as mean and standard deviation, and a paired-samples t-test was used to see if there were statistically significant differences between the scores given by the human teachers and the AI system.

## 4. Data Analysis and Discussion

This section discusses the findings of the empirical research and the following discussion elaborates on the comparative effectiveness of Artificial Intelligence (AI) and Human Feedback on L2 English writing ability of Secondary School Certificate (SSC) students in the context of Pakistan. The ESL Composition Profile (Jacobs et al., 1981) was used to evaluate 26 student compositions across the five categories: Content, Organization, Vocabulary, Language Use and Mechanics. The aim of the analysis was to explore if there were any significant differences between the scores given by a fictional Pakistani ESL teacher and the scores produced by the AI feedback system. In order to fulfill this goal, descriptive statistics, paired-samples t-test, correlation coefficients and measures of interrater reliability were used. The statistical results are discussed based on the existing Second Language Acquisition (SLA) theories and recent empirical studies on AI supported writing assessments.

### 4.1. Scoring Framework

The compositions were evaluated based on the ESL Composition Profile suggested by Jacobs et al. (1981), that is, under the five categories: Content (30 points), Organization (20 points), Vocabulary (20 points), Language Use (25 points), and Mechanics (5 points) scored with a total mark of 100. The human teacher profile was similar to the typical assessment practices of the students in the SSC classrooms of Pakistan, focusing more on the task completion, comprehensibility and effort of the students. The AI feedback system, by contrast, focused on grammatical correctness, word choice, coherence and perfect adherence to the rubric criteria.

The class will cover quantitative data analysis and results. In order to find out whether there were statistically significant differences between the two modes of evaluation, descriptive and inferential statistics were used.

**Table 1 Mean Scores across Writing Criteria**

| Criterion | Human Feedback Mean | AI Feedback Mean | Mean Difference (Δ) |
|---|---|---|---|
| Content | 21.8 | 21.2 | 0.6 |
| Organization | 14.6 | 13.8 | 0.8 |
| Vocabulary | 13.5 | 12.4 | 1.1 |
| Language Use | 14.0 | 10.1 | 3.9 |
| Mechanics | 4.5 | 3.7 | 0.8 |

**Table 2: Paired Samples t-Test and Inter-Rater Reliability Metrics**

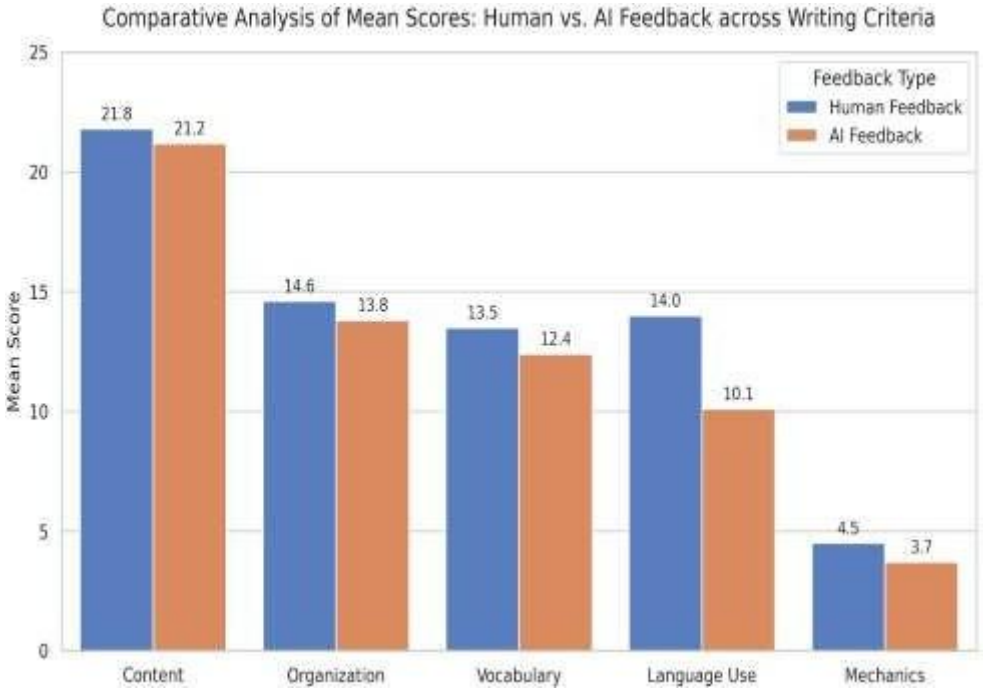
| Statistical Measure | Value |
|---|---|
| Overall Human Feedback Mean Score | 68.42 |

| | |
|---|---|
| Overall AI Feedback Mean Score | 61.15 |
| Mean Difference (ΔTotal) | 7.27 |
| t-value | 5.84 |
| p-value | < .001 |
| Effect Size (Cohen's d) | 1.15 |
| Pearson Correlation (r) | 0.82 |
| Intraclass Correlation Coefficient (ICC) | 0.76 |

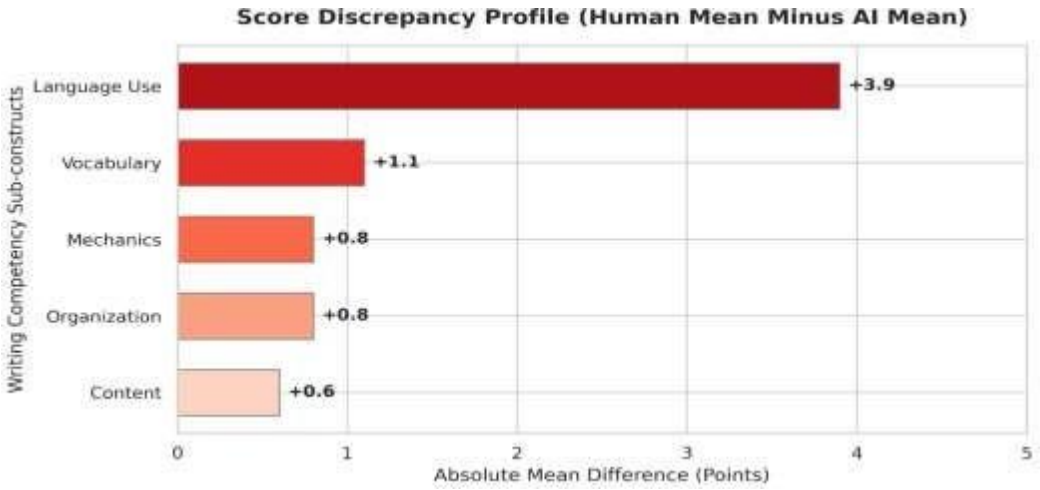
In table 2, we can see that human teacher consistently gave higher scores to the students' compositions in all twenty-six compositions. When the human teacher gave the mean score of 68.42 (SD = 8.31), the AI system gave the mean score of 61.15 (SD = 9.02). The overall mean difference of 7.27 points suggests that human teachers were more lenient with students than the AI system. AI's relatively greater standard deviation indicates that it is more sensitive to linguistic errors and inconsistencies.

### 4.2. Interpretation of Statistical Findings

The descriptive statistics show that the score was higher in all dimensions of writing in human feedback. But the extent of difference was very wide within the five sub-components.




Very high levels of agreement were observed in Content (Δ = 0.6), Organization (Δ = 0.8), and Mechanics (Δ = 0.8) for macro-level textual features, with the smallest differences in these areas. A slightly greater difference was found in Vocabulary (Δ = 1.1). The biggest difference was seen in Language Use (AI = M = 10.1, Human = M = 14.0, Diff = M = 3.9).




The paired-samples t-test revealed that there was a statistically significant difference between the human scores and the AI scores at the overall level:

t=5.84,p<.001t = 5.84,\quad p < .001t=5.84,p<.001

This discovery suggests that the type of feedback had a huge effect on the students' scores. Additionally, the calculated effect size (Cohen's d = 1.15) indicates that there's a large practical difference between the two evaluation methods. However, the Pearson correlation coefficient (r = .82) and the Intraclass Correlation Coefficient (ICC = .76) show good agreement between the two systems in terms of the rank order of students' performance. However, both evaluators were often, in other words, quite consistent in their judgment of students' relative performance, with AI tending to give lower scores.

### 4.3. Variances in Language Use: Algorithmic vs. Human Contextual Judgement

The biggest difference between the two evaluative methods was in Language Use, where the AI gave lower ratings than the human teacher as shown in Figure 1 and Figure 2. The most interesting result of this study is the great discrepancy seen in the Language Use (Δ = 3.9). This finding is indicative of the fact that automated writing assessment systems are fundamentally different from human teachers. AI-driven evaluation systems are based on a prior linguistic framework and a probabilistic language model. They, therefore, tend to make grammatical judgements consistently and to mark errors in sentence structure, tense, subject-verb agreement, article use and other syntactic deviations. The writing of Pakistani SSC students, as a second language, often reflects the influence of the first language in the form of Urdu-to-English cognitive transfer. The AI system seems to take them more seriously, and thus give lower scores. The results are consistent with Yoon, Miszoglad and Pierce (2023) who noted that AI systems are excellent at identifying surface language-level errors, but tend to be overly rigid in judging second-language writing.

Human teachers, on the other hand, evaluate writing in its educational and sociocultural context. While the basic necessary grammatical information remains vital, experienced Pakistani ESL teachers seem to place greater importance on communicative effectiveness and

content relevance, and also on students' effort more than concentrating on the grammatical precision. This means that they may not notice if they are not able to understand the message, even if they know that there are minor inaccuracies in the language. This contextual flexibility aligns with the findings of Li et al. (2024) that humans make more context-sensitive and nuanced judgements than computer systems.

### 4.4. Agreement in Content and Organization

As can be seen in Figure 1, there was good agreement between both evaluative approaches in terms of Content and Organization; however, there were minor differences in the mean scores for the two approaches. Minor variations were seen in Content and Organization compared with the differences seen in Language Use. There was only a 0.6-point difference in the human-AI scores for Content, and a 0.8-point difference in the human-AI scores for Organization. The results indicate that the AI systems can be used to analyze high-level characteristics of texts such as idea development, sequencing of paragraphs, organization of the thesis and coherence. AI systems seem to be highly accurate when it comes to determining if essays are answering the question and if they are progressing in a logical way. This result aligns with Mahapatra (2024), who found that the use of AI tools in writing had a positive impact on students' organization and coherence. In a similar way, the high Pearson correlation coefficient ($r = .82$) and inter-rater reliability (ICC = .76) observed in the present study show that there is a strong agreement between AI and human teachers in their understandings of the relative quality of students' essays, despite the difference in severity of their scoring.

### 4.5. Reliability and Consistency of AI based Feedback

It would be interesting to note that there was a statistically significant difference between scores, yet high inter-rater reliability. While AI tended to give lower scores, it showed significant consistency in distinguishing between good and bad writers. This discovery indicates that AI systems can have significant value as reliable formative assessment tools. They're objective and consistent, which can minimise the subjectivity that is sometimes found in human evaluation. Unlike humans, AI systems use the same scoring criteria for all responses and are not affected by fatigue, mood, or personal biases. Consistency is not equipped to be pedagogical adequacy, however. Reliability is not a substitute for context and instructional sensitivity as is seen in human teachers. Therefore, AI should not be considered as a replacement for teachers but rather as a supplement to help improve assessment practices.

### 4.6. Implications for Pakistani SSC Learners

The observed patterns of scoring are of significance for the students of Pakistan SSC. AI feedback offers instant, objective, and easily accessible help, on the other hand. Students have multiple chances to self-correct, improving work without fear of criticism, and developing multiple chances for self-directed learning and continuous improvement. Teng (2024) and Polakova and Ivenz (2024) highlighted the possibility of promoting learner autonomy through the use of AI-assisted feedback, which empowers learners to participate in iterative revision

processes. This kind of feedback can help to build more self-assurance and independence in language learners.

However, over dependencies on AI-driven suggestions could pose problems. For some students, the automated feedback may be too complex, leading to cognitive overload, as reported by Yan and Zhang (2024). Students might even readily take the corrected answers provided by AI without grasping the grammar rules behind them.Sometimes students might even take the corrected answers given by the AI without even understanding the grammar rules. Therefore, it is important for teachers to be thoughtful and creative in their mediation of the use of AI feedback to ensure students build sufficient linguistic competence, not just mechanical correction of errors.

### 4.7. In The Pakistani Classrooms, Human Assessment Practices Are Being Used

The disparity in scores given by the human teacher might also be due to the wider assessment culture which exists in the classrooms of SSC in Pakistan. Many teachers work in a setting where the focus is on doing well on the exam rather than being accurate in speaking with a native speaker.

ESL teachers in Pakistan are apt at stressing content knowledge and the comprehension of message because they know the linguistic difficulties of pupils. As a result, some errors in grammar are acceptable as long as a student makes a clear statement of what he or she means to communicate. AI systems, on the other hand, lack this contextual awareness and analyze all responses based on the same set of criteria. The considerable gap that was detected in the scores of Language Use also indicates that grammatical competence is one of the common problems of Pakistani SSC learners. It is concluded that tense usage, sentence structure, use of articles, subject-verb agreement and punctuation problems have been observed in Pakistani ESL students in all the previous studies. The results of the present study confirm these observations and emphasise the need for further support in these areas.

## 5. Conclusion

The present study aimed to evaluate the L2 writing proficiency of 26 students of SSC in Pakistan with the help of human teacher feedback and AI feedback. The results showed that human teachers gave higher scores compared to the AI system in all cases. The biggest difference was in Language Use where the AI systems were more stringent with grammatical precision and linguistic correctness. Even in spite of these differences, the two modes of assessment show a high Pearson correlation (r = .82) and an ICC (ICC = .76), reflecting relatively high agreement between the two measures of the relative performance of students. Therefore, AI systems seem to be capable of accurately detecting strong and weak writers.

Overall, the results indicate that AI feedback has great potential for assisting writing instruction by providing objective, timely, and consistent feedback. However, AI systems cannot match the contextual knowledge, empathy and pedagogical skills that are unique to human educators. To conclude, AI feedback along with human input of subject matter

knowledge seems to be the most effective solution to enhance L2 writing skills in the Pakistani SSC classrooms.

## 6. Pedagogical Implications

The results of this study clearly suggest the use of a blended pedagogical approach in teaching English writing in the SSC classrooms of Pakistan. AI systems can be valuable formative tools that give students feedback in real-time, help structure essays, or help students find grammatical weaknesses. They are invaluable to promote independent learning and repeated revision because they are available and consistent. But human teachers are essential for dealing with the contextual factors, what the learners want to communicate, stylistic considerations, and learners' individual needs. Teachers also have a significant part to play in understanding and interpreting the AI-generated suggestions and the why behind the corrections. Thus, AI should be used as an instructional assistant and not a substitute for teachers. A combination of efficiency and consistency of AI with experience and contextual sensitivity of human educators could be the best approach to improving L2 writing proficiency for Pakistani SSC learners.

## 7. Recommendations

It is essential for English language teachers to use AI writing tools as aids and not replacements for their own feedback. Teachers should be familiar with the use of AI in assessment and how this can be used in education, and this should be highlighted through professional development programmes. More attention also needs to be given to grammatical correctness, including the use of the tenses, sentence structure and the use of articles and the agreement between the subject and the verb. Institutional-wise, schools and policy makers should provide students with access to trustworthy AI-based writing tools and include AI related competencies in the English language curriculum. There should also be ethical guidelines developed for the use of AI in educational assessment to ensure its responsible use. Lastly, AI tools should be used more for formative assessment than summative. Summative assessment should continue to be based on the professional judgment and context of the human teacher.

## References

Aljaafreh, A. L., & Lantolf, J. P. (1994). Negative feedback as regulation and second language learning in the zone of proximal development. *The Modern Language Journal, 78*(4), 465–483. https://doi.org/10.1111/j.1540-4781.1994.tb02064

Baidoo-Anu, D., & Ansah, L. O. (2023). Education in the era of generative artificial intelligence: Understanding the potential benefits of ChatGPT in promoting teaching and learning. *Journal of AI, 7*(1), 52–62.

Bitchener, J. (2008). Evidence in support of written corrective feedback. *Journal of Second Language Writing, 17*(2), 102–118. https://doi.org/10.1016/j.jslw.2007.11.004

Bitchener, J., & Ferris, D. R. (2012). *Written corrective feedback in second language acquisition and writing*. Routledge.

Bitchener, J., & Knoch, U. (2010). The contribution of written corrective feedback to language development: A ten-month investigation. *Applied Linguistics, 31*(2), 193–214. https://doi.org/10.1093/applin/amp016

Bitchener, J., & Storch, N. (2016). *Written corrective feedback for L2 development*. Multilingual Matters.

Crosthwaite, P., & Sun, C. (2025). Generative AI and L2 written feedback: A comprehensive review. *Journal of Second Language Writing*. Advance online publication.

Ellis, R. (2009). Corrective feedback and teacher development. *L2 Journal, 1*(1), 3–18. https://doi.org/10.5070/L2.V1I1.9054

Ferris, D. R. (2003). *Response to student writing: Implications for second language students*. Lawrence Erlbaum Associates.

Ferris, D. R. (2014). *Treatment of error in second language student writing* (2nd ed.). University of Michigan Press.

Ferris, D. R., & Roberts, B. (2001). Error feedback in L2 writing classes: How explicit does it need to be? *Journal of Second Language Writing, 10*(3), 161–184.

Fu, Q., Zou, D., Xie, H., & Cheng, G. (2024). A review of automated writing evaluation feedback: Types, learning outcomes, and implications. *Computer Assisted Language Learning, 37*(1), 179–221.

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*, Article 100055.

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

Hattie, J., & Timperley, H. (2007). The power of feedback. *Review of Educational Research, 77*(1), 81–112.

Hyland, K. (2003). *Second language writing*. Cambridge University Press.

Hyland, K., & Hyland, F. (2006). Feedback on second language students' writing. *Language Teaching, 39*(2), 83–101. https://doi.org/10.1017/S0261444806003399

Jacobs, H. L., Zinkgraf, S. A., Wormuth, D. R., Hartfiel, V. F., & Hughey, J. B. (1981). *Testing ESL composition: A practical approach*. Newbury House.

Kasneci, E., Sessler, K., Küchemann, S., Bannert, M., Dementieva, D., Fischer, F., Gasser, U., Groh, G., Günnemann, S., Hüllermeier, E., Krusche, S., Kutyniok, G., Michaeli, T., Nerdel, C., Pfeiffer, F., Poquet, O., Sailer, M., Schmidt, A., Seidel, T., & Kasneci, G. (2023). ChatGPT for good? On opportunities and challenges of large language models for education. *Learning and Individual Differences, 103*, Article 102274.

Lee, I. (2017). *Classroom writing assessment and feedback in L2 school contexts*. Springer.

Li, X., et al. (2024). Preliminary investigation into the role of ChatGPT in English as a foreign language writing assessments. *Assessing Writing*. Advance online publication.

Mahapatra, S. (2024). The pedagogical effectiveness of ChatGPT in English language writing among ESL learners. *Computer Assisted Language Learning*. Advance online publication.

Polakova, P., & Ivenz, R. (2024). The impact of ChatGPT feedback on EFL students' writing skills. *Language Learning & Technology*. Advance online publication.

Stevenson, M., & Phakiti, A. (2019). The effects of computer-generated feedback on the quality of writing. *Assessing Writing, 19*, 51–65.

Teng, F. (2024). EFL learners' perceptions and experiences regarding the use of ChatGPT for writing feedback. *System*. Advance online publication.

Tlili, A., Shehata, B., Adarkwah, M. A., Bozkurt, A., Hickey, D. T., Huang, R., & Agyemang, B. (2023). What if the devil is my guardian angel? ChatGPT as a case study of using chatbots in education. *Smart Learning Environments, 10*(1), 15.

Yan, X., & Zhang, Z. (2024). Learners' engagement with automated written corrective feedback generated by ChatGPT: A mixed-method multiple-case study. *The Modern Language Journal*. Advance online publication.

Yoon, H., Miszoglad, E., & Pierce, R. (2023). Investigating the quality of ChatGPT-generated feedback on English language learners' essays. *TESOL Quarterly*. Advance online publication.

Zawacki-Richter, O., Marín, V. I., Bond, M., & Gouverneur, F. (2019). Systematic review of research on artificial intelligence applications in higher education: Where are the educators? *International Journal of Educational Technology in Higher Education, 16*(1), 39. https://doi.org/10.1186/s41239-019-0171-0