

AI-MEDIATED FEEDBACK AND IELTS SPEAKING PERFORMANCE: AN INVESTIGATION INTO STUDENT PROGRESS AND PERCEPTIONS

Duong Thu Ha

Hanoi National University of Education, Hanoi, Vietnam

Corresponding author: Duong Thu Ha, e-mail: hadtl@hnue.edu.vn

Received April 20, 2026. Revised June 14, 2026. Accepted July 20, 2026.

Abstract. This study examines the potential impact of a human-AI collaborative feedback system on the IELTS speaking abilities of upper-intermediate English majors at a university in Vietnam. An explanatory sequential mixed-methods design was used, with the research focusing on the students' oral performance over a 4-week formative assessment period. The quantitative findings, derived from a pre-test and post-test conducted with a single group as a quasi-experimental study and from a 15-item Likert-scale questionnaire completed by the participants ($N = 76$), were supplemented by the qualitative responses to three open-ended survey questions. The quantitative data indicated a statistically significant improvement only in the area of Lexical Resource, whereas the overall band scores did not show a significant change, likely due to a temporary compromise between linguistic complexity and oral spontaneity. The survey results revealed that the learners regarded score breakdowns based on criteria as highly valuable and also placed importance on justifications of their transcripts that were supported by evidence in order to encourage self-regulated learning. At the same time, the qualitative responses emphasized that ongoing involvement by human experts is still necessary for overcoming affective barriers and for ensuring evaluative reliability. The study therefore proposes a Human-in-the-Loop (HITL) framework that combines the accuracy of automation with the guidance of pedagogy and thus complements standalone AI evaluation tools in the context of language assessment.

Keywords: AI-mediated feedback, Google Gemini, IELTS Speaking, Tertiary education, English majors

1. Introduction

In the way that English is currently taught, achieving a high degree of oral fluency is a key aim for students, particularly third-year students who are aiming to reach the C1 level. This phase in their studies represents a major change in the curriculum, since third-year English students must satisfy a compulsory graduation requirement of C1/IELTS 7.0 (this is equivalent to Level 5 in Vietnam's Foreign Language Proficiency Framework). It is for this reason that courses like "Listening – Speaking 6" serve as important routes to developing advanced oral skills. Yet the assessment of speaking abilities is a time-consuming process and usually leads to a "feedback gap", that is to say that students do not get timely and individualized feedback (Farooqi, 2025, p. 715). As Farooqi (2025, p. 715) and Vu et al. (2025, p. 30) note, even though formative feedback is essential for the development of speaking skills, it is still a great challenge for teachers to give high-quality, personalized assessments in large classes.

The application of Artificial Intelligence (AI) in the teaching of English as a Foreign Language (EFL) has been extensively researched worldwide, with the results showing that AI-based tools are able to improve both students' speaking skills and their readiness to communicate (Dakhil et al., 2025, p. 11). In Vietnam, researchers have observed that students have a favorable attitude towards using AI assistants for the purpose of language learning (Lam, 2024; Nguyen, 2024). Furthermore, there is empirical evidence to show that AI practice tools can lessen speaking anxiety and increase proficiency among undergraduate students (Duong & Suppasetseree, 2024, p. 314; Nguyen, 2025, p. 46). Nevertheless, a significant problem remains because of the complete absence of human interaction and the danger of students becoming overly reliant on AI methods

that work independently. Specifically, unmediated use of AI can result in a condition known as “false mastery”, wherein learners rely on automated shortcuts and therefore do not carry out deep cognitive processing or independent thinking (Bull, 2026, p. 89). Additionally, purely digital tools do not provide the same degree of socio-emotional understanding or pedagogical empathy as human teachers, which highlights the need for a model in which the instructor takes on a mediating role (Baskara, 2025, p. 443; Vo & Bui, 2024, p. 216).

What is novel about this research is the fact that it employs a human-in-the-loop collaborative feedback approach by means of an AI agent known as the “IELTS Speaking Auto-Grader”. The agent has been developed on the Google Gemini platform and has been set up via careful prompt engineering in such a way that it can analyze student recordings and give diagnostic feedback in accordance with the IELTS nine-band scale. Unlike earlier studies which have concentrated on general AI chatbots (Baskara, 2025) or standalone transcription programs (Park & Shin, 2023; Tat, 2025), this method involves structured evaluation carried out with the help of the instructor. The approach addresses the limitations mentioned by Chen and Sun (2025) and Wang and Zhou (2025), who advocated for a more systematic use of automated proficiency tools by instructors. By having the instructor act as a mediator, the study bridges the gap between the efficiency of technology and the reliability of teaching methods.

The main purpose of this research is to investigate the possible effect of an instructor-mediated, human-in-the-loop AI feedback model on the oral performance of third-year English majors, as well as to examine their views on this method. In order to do this, the study uses an explanatory sequential mixed-methods design, combining quantitative data (from pre- and post-test IELTS scores and a 15-item Likert-scale survey) with qualitative insights from the open-ended questions completed by 76 participants. Based on this mixed-methods framework, two research questions were formulated:

RQ1. To what extent is AI-mediated feedback associated with changes in the IELTS speaking performance of third-year English majors?

RQ2. What are the students’ perceptions of the utility and reliability of the AI-mediated feedback provided in the speaking evaluation reports?

2. Literature review

2.1. AI-automated feedback in speaking instruction

The use of automated feedback has altered the method in which second language speaking is taught by providing learners with immediate, diagnostic feedback a feature which is usually absent in traditional classroom situations (Farooqi, 2025, p. 716; Wang & Zhou, 2025, p. 211). In the case of speaking, AI systems evaluate a number of factors such as phonological accuracy and discourse coherence (Wang & Zhou, 2025). Research carried out by Zou et al. (2023) indicates that EFL learners feel they have made significant progress in speaking because they have received personalized written feedback and specific recommendations. On the basis of socio-cultural theory (Vygotsky, 1978), AI acts as a digital “More Knowledgeable Other” that operates within the learner’s Zone of Proximal Development and so offers automated assistance which allows students to become aware of their linguistic errors while carrying out independent practice. This type of continuous teaching method means that people can have more frequent speaking practice without having to increase the amount of administrative work teachers have to do (Luckin, 2017). Furthermore, the automated environment provides a low-pressure, non-judgmental atmosphere which helps to reduce learners’ inhibitions and as a result encourages their oral performance, their attention to structural accuracy, and their general readiness to communicate (Dakhil et al., 2025, p. 11; Duong & Suppasetseree, 2024, p. 314).

2.2. Human raters vs. AI evaluation in the IELTS context

When standardized systems such as IELTS are used to evaluate oral proficiency, there are significant contradictions between the scores assigned by automated methods and those given by human judges. While AI models demonstrate high consistency in assessing lexical density,

grammatical complexity and phonological features (Chen & Sun, 2025; Park & Shin, 2023), human assessors remain superior in judging pragmatic appropriateness and interactive coherence. In Vietnam, individual automated tools frequently fail to detect the subtle elements of conversational context or tend to penalize natural communicative fillers (Vu et al., 2025, p. 37), indicating that it is necessary to have an instructor-led approach in order to compare the results of automated systems with well-established teaching standards (Tat, 2025). In order to overcome the limitations of technology, current educational practices support a HITL framework. In this type of collaborative system, the AI system carries out the analysis of basic linguistic data, while human teachers act as evaluative validators by correcting mechanical errors and giving feedback that is based on the context of the situation, thus achieving a balance between the scalability of automation and human pedagogical empathy.

2.3. Student perceptions of AI in the Vietnamese tertiary context

Vietnamese students who are learning English show a high level of engagement when they use AI tools. They find AI helpers useful for personalized learning and for getting immediate corrections, which in turn helps to lower their Affective Filter and boosts their confidence as a result of doing a number of repeated, low-stakes exercises (Lam, 2024, p. 1303; Nguyen, 2024, p. 186; Vo & Bui, 2024, pp. 216–217). Yet Bull (2026, p. 89) warns that AI is a “mixed blessing” because it might lead to a reduction in academic effort if students rely completely on the automated outputs without any “pedagogical curation” or guidance from teachers.

When it comes to evaluating how students interact with automated assessment systems, the latest literature on computer-assisted language learning (CALL) has described student perception in terms of three closely related dimensions: pedagogical impact, perceived utility, and evaluative trust. The pedagogical impact deals with the effect that automated feedback has on learners’ psychological state, specifically its capacity to reduce affective filters and keep up a structural motivation for carrying out self-directed practice (Dakhil et al., 2025; Zou et al., 2023). As for perceived utility and clarity, the question is about the transparency and the level of detail in the feedback given. Learners are more likely to accept the feedback when it changes from providing general scores to giving criterion-based breakdowns that are backed by concrete evidence, for example through the use of exact performance transcripts (Farooqi, 2025). Finally, evaluative trust and future expectations influence the long-term feasibility of incorporating AI. Even though students value the quick results that AI offers, a major part of their perception involves a certain amount of professional skepticism concerning the machine’s ability to grasp pragmatic nuances, and this implies that a human-verified approach is required to guarantee evaluative reliability (Bull, 2026; Wang & Zhou, 2025).

3. Research methodology

3.1. Research design

The study employs an explanatory sequential mixed-methods design; it begins by observing quantitative changes in oral performance at two key assessment phases (T1 and T2) using a one-group quasi-experimental pretest-posttest method, and then administers a mixed perception questionnaire which includes both quantitative items (using a 5-point Likert scale) and qualitative ones (with open-ended questions). The fact that the methods are carried out in this sequential order means that the retrospective perceptual data can serve to provide a contextual background and act as a supplement to the quantitative trends in IELTS performance.

3.2. Participants

The final group was made up of 76 third-year students majoring in English who were studying at the Faculty of English of a large public university in northern Vietnam. With regard to their demographic composition, the group comprised 67 women (88.2%) and 9 men (11.8%), the average age being approximately 21 years. All the participants shared the same academic objective, which was to attain C1 level proficiency in English, specifically so as to achieve a total score of at least 7.0 on the IELTS test in order to satisfy the university’s required graduation exit

criteria. They were both attending two different classes for the “Listening – Speaking 6” course, which was taught by the same instructor. In order to maintain the structure of the one-group quasi-experimental design, the two classes were merged and treated as one cohort ($N = 76$) for all the statistical analyses. This sample size reflects the intensive and long-term nature of the pedagogical intervention, which included the continuous monitoring and the application of double scoring by both human raters and AI for the oral presentations over a period of four weeks.

3.3. Research setting

The study occurred during the second semester of the academic year 2025–2026 and was carried out in the “Listening – Speaking 6” class, the students in this class using the IELTS material from their coursebook as the basis for their speaking prompts. The pedagogical intervention was applied in an identical manner to the whole group, and since the same instructor taught both of the parallel classes, there was consistency in the instructions given, in the pace of the curriculum, and in the use of the AI-mediated feedback system.

3.4. Data collection tools

To gather a comprehensive dataset that integrates performance metrics with psychological and pedagogical perceptions, three primary instruments were deployed:

(1) **IELTS speaking tasks:** For the IELTS speaking section, standardized mock tests consisting of all three parts (Part 1, Part 2 and Part 3) were used as the measure at both the pre-intervention stage (Time 1 – T1) and the post-intervention stage (Time 2 – T2). The tests recorded live oral ability in the four official categories: Fluency and Coherence, Lexical Resource, Grammatical Range and Accuracy, and Pronunciation.

(2) **AI evaluation reports:** The diagnostic reports were produced by a Gemini-based agent specifically designed as the “IELTS Speaking Auto-Grader”. This agent transcribed the audio files and then prepared the feedback in line with the IELTS nine-band criteria. In order to identify and correct any possible inconsistencies in the scoring, the instructor looked at each AI-generated report and made adjustments to it by referring to the actual audio recordings and the official Band Descriptors before the reports were issued. The course instructor checked the AI-drafted report on three specific grounds: (1) acoustic and phonological accuracy: correcting the transcription errors in the speech-to-text process which arose due to interference from the learner’s first language; (2) pragmatic and coherence validation: altering the scores in instances where the AI had treated natural, non-linear communicative pauses as dysfluency; and (3) pedagogical curation: enhancing the mechanical linguistic measures and providing vocabulary scaffolds that were suitable for the students’ level of proficiency.

For instance, in the case where the AI had automatically given Student 42 a penalty for pronunciation due to interference from the student’s first language (for example, by omitting the final consonants in “recharge”), the teacher became involved. The instructor observed that the student had maintained both overall intelligibility and a natural intonation in their sentences and therefore cancelled the penalty recommended by the automated draft, then changed the Pronunciation score to 6.0 and gave specific feedback.

(3) **Perception questionnaire:** The survey instrument was based on those developed by Nguyen (2024) and Lam (2024) and was then modified to suit the particular situation of Vietnamese undergraduates. Although adapting items from previous studies provided a starting point, the construct validity and the suitability of the items were additionally examined by two experienced TESOL academics before the questionnaire was used. The questionnaire consisted of two main sections:

The quantitative section (questions 1 to 15) consisted of 15 closed-ended items which were rated using a 5-point Likert scale. For each participant, composite scale scores were determined before the subscale means and standard deviations were calculated. Analysis of reliability produced internal consistency coefficients for the various dimensions: for Impact on Learning and Performance ($\alpha = .795$), for Perceived Utility and Clarity ($\alpha = .863$), and for Reliability and Future Expectations ($\alpha = .828$).

The qualitative section (Items 16–18) included three open-ended questions which asked students to identify the helpful aspects of the reports (Item 16), to point out the main limitations (Item 17), and to give suggestions for future improvement (Item 18). These open-ended responses are used to complement and provide context for the quantitative findings.

3.5. Data collection procedure

The operational protocol spanned a 4-week formative pedagogical cycle, executed through four distinct procedural phases:

In phase 1 – the baseline diagnostic assessment (T1) – the participants did an IELTS Speaking recording, which was then processed by the AI agent and checked by the instructor using the official band descriptors. The instructor also gave the students a 45-minute orientation in which they were taught how to interpret criteria scores, find errors in the transcript, and apply corrective feedback.

In phase 2, the 4-week feedback-action cycle involved students keeping weekly self-study logs as a way of applying the feedback, which included reviewing the T1 speech transcripts, re-recording sections in order to correct the pronunciation and intonation patterns that had been flagged, and engaging in weekly timed triads for practice.

In phase 3, at the end of the four-week cycle, a second complete speaking assessment was recorded under the same conditions and then evaluated using the same HITL protocol.

In phase 4, the quantitative data were analyzed using SPSS. Paired-samples t-tests were carried out to examine the changes in the IELTS scores (comparing T1 with T2). In order to provide additional information to the p-values, Cohen's d effect sizes and 95% Confidence Intervals (95% CIs) for the mean differences were calculated and presented. Lexical Resource was chosen as the primary outcome of interest. Since the comparisons for the subscales were exploratory in character and the study used a single-group design, no multiple comparison corrections (for example, Bonferroni or Holm) were made for the secondary criteria, and the findings relating to these are therefore interpreted with caution as preliminary trends. For the perception questionnaire, a composite score was calculated for each participant to obtain the subscale Means and Standard Deviations. The qualitative answers given to the open-ended questions were analyzed by means of inductive thematic coding to supplement the results from the quantitative survey.

4. Findings and discussion

4.1. Changes in students' speaking performance

Paired-samples t-tests were carried out for all four IELTS criteria and for the overall band scores in order to examine the changes in score between Time 1 and Time 2. Table 1 presents the comparative means, standard deviations, t-statistics, p-values, Cohen's d effect sizes, and 95% confidence intervals.

Table 1. Paired samples t-test results for speaking performance (N=76)

Criteria	Time 1 (M±SD)	Time 2 (M±SD)	Mean Diff [95% CI]	t	p
Overall Band Score	6.91 ± 0.63	6.93 ± 0.50	0.02 [-0.09, 0.13]	-0.35	.724
Fluency & Coherence	6.86 ± 0.69	6.79 ± 0.57	-0.07 [-0.19, 0.05]	1.11	.272
Lexical Resource	6.87 ± 0.68	7.01 ± 0.58	0.14 [0.01, 0.27]	-2.15	.035
Grammatical Range & Accuracy	6.54 ± 0.67	6.67 ± 0.59	0.13 [-0.01, 0.27]	-1.89	.063
Pronunciation	6.97 ± 0.61	7.01 ± 0.54	0.04 [-0.06, 0.14]	-0.83	.410

Table 1 shows a statistically significant difference that was not adjusted for, in the Lexical Resource measure, this increasing from 6.87±0.68 at Time 1 to 7.01±0.58 at Time 2, $t(75) = -2.15$, $p = .035$, with a small effect size (Cohen's $d = 0.25$, 95% CI [0.01, 0.27]). This change took place after the 4-week self-regulated practice cycle, during which the students worked with the vocabulary banks and received transcript feedback from their T1 reports. It should, however, be pointed out that if an alpha level adjustment for multiple comparisons (for example, using Bonferroni or Holm) had been strictly applied to all five tests, the p-value would have gone

above the adjusted threshold, so the finding should be treated with caution and seen as a possible trend that needs to be confirmed in controlled trials.

On the other hand, the changes in Fluency and Coherence ($p = .272$) and in Grammatical Range and Accuracy ($p = .063$) were not statistically significant. The small, non-significant difference in fluency could be due to a temporary cognitive compromise in which learners, when applying the newly learned vocabulary, increased their cognitive processing while speaking spontaneously. The improvement in Pronunciation ($+0.04$, $p = .410$) was negligible, which might indicate the limitations of text-based feedback methods when it comes to dealing with suprasegmental features. The overall Band Scores did not change between T1 and T2 ($p = .724$).

4.2. Students' perceptions of AI-mediated speaking feedback

In order to assess the pedagogical reception of the human-AI collaborative framework, the attitudinal data from the 15 Likert-scale items were divided into three clear thematic categories that addressed the research questions. For the sake of best spatial efficiency and clarity in the quantitative data presentations (Tables 2, 3, and 4), the survey items are shown using shortened versions of the main constructs. Importantly, these concise labels have a direct semantic relationship to the full statements used in the validated research instrument. Before determining the subscale group scores, a composite scale score was calculated for each participant.

The perception survey's first aspect was concerned with the way in which the AI reports affected the students' motivation and performance. Table 2 gives a detailed account of the mean scores and standard deviations for the items relating to Impact on Learning and Performance.

Table 2. Students' perceptions of the Impact on Learning and Performance (N=76)

Abbreviated items	Mean	SD	Interpretation
Q1. Active use of suggestions for second recording	3.86	0.83	High
Q2. Reduced anxiety in evaluation	3.84	0.85	High
Q3. Increased motivation to practice consistently	4.00	0.85	High
Q4. Expanded lexical and grammatical range	3.89	0.92	High
Q5. Enhanced confidence in speaking progress	3.63	0.86	High
Group Average	3.84	0.86	High

Note. Q1 to Q5 correspond to Items 1 to 5 of the perception questionnaire.

Table 2 shows that students strongly agreed with the view that AI-mediated feedback has a positive effect on learning (mean = 3.84, standard deviation = 0.86). It is worth noting that the detailed sub-score breakdowns (Q3, mean = 4.00) increased students' motivation to carry out practice, while the targeted vocabulary suggestions (Q4, mean = 3.89) directly contributed to the observed improvement in Lexical Resource. Although the students actively followed the AI's recommendations (Q1, mean = 3.86) and felt less anxious when evaluating their work (Q2, mean = 3.84), their confidence in their progress was the lowest (Q5, mean = 3.63), which suggests that it takes continued exposure to turn immediate linguistic improvements into long-term self-efficacy.

The next dimension looked at how clear and useful students found the digital reports, and Table 3 gives the mean ratings for this category.

Table 3. Students' perceptions of the Perceived Utility and Clarity of AI reports (N=76)

Abbreviated items	Mean	SD	Interpretation
Q6. Ease of understanding and navigation	3.95	0.75	High
Q7. Identifying weaknesses via criteria breakdown	4.12	0.83	High
Q8. Convincingness of evidence-based justifications	4.11	0.72	High
Q9. Comprehensiveness vs. traditional oral comments	3.76	0.99	High
Q10. Benefits of quick turnaround time	3.89	0.76	High
Q11. Clarity of IELTS Band Descriptors	4.04	0.79	High
Group Average	3.98	0.81	High

Note. Q6 to Q11 correspond to Items 6 to 11 of the perception questionnaire.

The figures in Table 3 show that students agreed to a high degree about the usefulness and clarity of the AI-produced reports (mean = 3.98, standard deviation = 0.81), this being the highest collective mean of the three dimensions. Particularly, the results for the criteria-based aspect (Q7, mean = 4.12) and those for the evidence-based justifications (Q8, mean = 4.11) reached the top end of the “High” range, proving that specific transcripts together with clear score justifications were effective in helping learners understand the difference between their performance and the standard assessment criteria. Moreover, the students greatly liked the way the reports matched the official IELTS Speaking Band Descriptors (Q11, mean = 4.04) and the fast turnaround time of the reports (Q10, mean = 3.89), both of which enabled them to reflect on their performance and make adjustments immediately. Although the ease of understanding and navigating the reports was also rated very highly (Q6, mean = 3.95), the comparison with traditional oral comments (Q9, mean = 3.76) produced the most differing responses (standard deviation = 0.99), which shows that even though the digital reports offered thorough criteria-based clarity, some learners still regarded the interpersonal and affective support found in teacher-led feedback as valuable.

Finally, Table 4 looks at students’ opinions regarding the long-term reliability of the system and the expectations they have about its ongoing use in the curriculum.

Table 4. Students’ perceptions of Feedback Reliability and Future Expectations (N=76)

Abbreviated items	Mean	SD	Interpretation
Q12. Accuracy of AI scores vs. actual ability	3.64	0.95	High
Q13. Confidence in instructor-curated feedback	3.82	0.78	High
Q14. Preference for digital reports over verbal feedback	3.72	0.95	High
Q15. Desire for future AI feedback integration	3.93	0.84	High
Group Average	3.78	0.88	High

Note. Q12 to Q15 correspond to Items 12 to 15 of the perception questionnaire.

As for the reliability and future integration of AI-mediated feedback, Table 4 indicates that the students kept a high level of agreement ($M = 3.78$, $SD = 0.88$). The greatest degree of agreement among the students was seen in question Q15 ($M = 3.93$), since they clearly wanted to keep receiving these automated reports in the future speaking modules. This preference is also shown by Q14 ($M = 3.72$), which shows that when carrying out independent speaking practice, learners prefer the detailed and permanent character of digital reports to temporary verbal comments. Interestingly, however, Q12 on the perceived accuracy of the AI scores received the lowest mean score of all the questions ($M = 3.64$). This shows that although the students find the feedback useful, they are still somewhat doubtful about the machine’s ability to fully grasp the complexities of human speech. The lack of trust in this area is compensated for by Q13 ($M = 3.82$), in which students stated that the teacher’s role in checking and selecting the feedback is essential to their sense of security. The findings thus imply that AI should be regarded as a powerful supplementary tool rather than as a complete substitute for human teacher involvement.

4.3. Qualitative feedback from open-ended survey responses

In order to gain a more detailed insight into the way in which the participants experienced the human-AI collaborative feedback process, the written answers to the three open-ended survey questions (questions 16 to 18) were analyzed by means of an inductive thematic coding approach based on Braun and Clarke’s (2006) qualitative framework. This inductive coding process led to the official identification of three main macro-themes: (1) the advantages of the AI-mediated feedback, (2) the difficulties and cognitive trade-offs, and (3) suggestions for future improvement. The qualitative comments provided are helpful for interpreting the quantitative test results and survey ratings.

4.3.1. Strengths of the AI-mediated feedback

The qualitative remarks strongly confirm the high average scores achieved in the perceived utility section. Several students highlighted the benefit of receiving detailed, criteria-based feedback instead of just general comments. For instance, Student 10 said that “instead of giving

general comments, the reports clearly point out my specific weaknesses, such as pronunciation errors, a limited vocabulary, or the overuse of simple sentence structures.” It appears that this degree of specificity is a key reason for students taking the initiative to carry out practice, since Student 18 noted that “the vocabulary suggestions are very helpful every time I have difficulty finding better words.” Thus, the qualitative observations lend support to the quantitative findings by providing an explanation for the positive change observed in the Lexical Resource scores following the self-study cycle.

4.3.2. Challenges and cognitive trade-offs

Although the general response was positive, the students identified a significant practical issue which accounts for the small, non-significant differences in the Fluency and Coherence scores ($p = .272$). A number of the participants mentioned that they noticed a great deal of mental strain when attempting to use newly learned vocabulary in spontaneous speech. As Student 19 remarked: “I had been asked to use more advanced vocabulary, but when I tried to do so I sometimes became less fluent because it took me longer to think.” Student 14 also said that even though they knew how to extend their answers in response to the recommendations, the pressure of speaking in real time caused them to “run out of ideas or feel nervous”. These observations indicate that although the AI feedback does provide the necessary linguistic tools, students still need to gain more experience over time in order to incorporate these improvements without losing their oral fluency.

4.3.3. Suggestions for future improvement

When asked for their suggestions, the main issue mentioned was the need for more concrete modeling resources. The students wished for more practical examples and model answers so that they could bridge the gap between receiving feedback and improving their speaking skills. For example, Student 10 proposed that “short action plans or step-by-step strategies” should be included, and Student 52 stated that teachers could “add potential model responses to guide us on how to improve specifically based on our previous performance”. Furthermore, the qualitative data revealed a clear agreement about the complementary role of technology. As Vo and Bui (2024, p. 216) pointed out, the participants emphasized that although the AI agent provided the figures and pointed out the errors, it did not offer the emotional support and subtle encouragement that only a human instructor can give. This is the reason why, even though the digital reports are very useful, the students still consider instructor mediation to be the main source of evaluative reliability (Baskara, 2025, p. 443).

5. Conclusions and pedagogical implications

5.1. Conclusions

The study looked at the changes in students’ performance and their perceptions after they had received feedback mediated by an instructor and an AI. Several important conclusions may be drawn from the findings.

As for oral performance (RQ1), the quantitative data show that over the course of a 4-week single-group study, the overall band scores stayed mostly unchanged, although there was a positive change in the Lexical Resource category ($p = .035$, Cohen’s $d = 0.25$). This indicates that detailed diagnostic feedback could help promote short-term lexical reflection and active learning. Yet since the single-group pretest-posttest design does not take into account external factors such as other instruction being given at the same time or the effects of repeated testing, and since this result would not remain significant after applying strict multiple-testing corrections, it should be regarded as an initial observational trend rather than as conclusive evidence of effectiveness.

With regard to student perceptions (RQ2), the participants as a whole expressed positive opinions about all the aspects of the survey, praising the structure, the detailed breakdown of the criteria, and the quick availability of the AI reports. At the same time, however, the students’ responses showed a clear preference for having instructors provide oversight to supplement the automated evaluations.

5.2. Pedagogical implications

The results of this study have important practical consequences for contemporary language teaching and for assessment methods that make use of technology. Firstly, educational institutions ought to use HITL frameworks that have been validated by teachers rather than relying on standalone AI grading systems, employing AI as a quick diagnostic aid while keeping it under the supervision of instructors. Secondly, to deal with the cognitive imbalance between the complexity of vocabulary and the speed of speaking, teachers should help students find a balance between expanding their vocabulary and practicing fluency, making sure that new vocabulary is introduced slowly so as not to overwhelm learners' working memory when they are producing speech spontaneously. Lastly, automated reports can be improved by incorporating active scaffolding, for example, by giving students concrete model answers together with the highlights of their errors in order to support their self-directed learning.

5.3. Limitations and suggestions for further research

When drawing conclusions from these findings, several methodological limitations should be taken into account. Firstly, the study used a one-group quasi-experimental pretest-posttest approach and included no control group. As a result, the changes seen between T1 and T2 might be due to other factors such as regular classroom teaching, maturation, or the fact that the test was taken more than once, rather than being solely attributable to the feedback intervention. Secondly, the four-week period was quite short and might therefore not be long enough to show broader longitudinal developments in speaking fluency or pronunciation. Thirdly, with regard to the scoring process, the course instructor was the only one who reviewed the AI's outputs and there was no independent or blind second human rater; hence, it was not possible to calculate inter-rater agreement figures. Fourthly, the sample consisted only of third-year English majors at one public university in northern Vietnam, which restricts the generalizability of the results. Finally, the qualitative data came from three open-ended survey questions. Even though these are helpful in supporting the quantitative findings, they do not offer the same level of depth as those obtained through semi-structured interviews.

Future studies should use randomized controlled trials (RCTs) including control groups, adopt multi-rater blinded scoring procedures together with measures of inter-rater reliability, lengthen the duration of the intervention to cover an entire semester, and also include in-depth qualitative interviews.

Notes on contributors: Duong Thu Ha (M.A.) is a lecturer at the Faculty of English, Hanoi National University of Education, Vietnam.

Conflict of interest: The author declares no conflict of interest.

Acknowledgments: During the preparation of this manuscript, generative AI tools were used solely to enhance language quality, improve grammar and polish the writing style. All research ideas, analytical frameworks, data collection, and critical interpretations were independently conceived and executed by the author.

REFERENCES

- Baskara, F. R. (2025). ChatGPT and Google Gemini in EFL education: a qualitative exploration of pedagogical efficacy among Indonesian sophomores. *JOLLT Journal of Languages and Language Teaching*, 13(1), 436–447. <https://doi.org/10.33394/joltt.v13i1.9926>
- Braun, V., & Clarke, V. (2006). Using thematic analysis in psychology. *Qualitative Research in Psychology*, 3(2), 77–101. <https://doi.org/10.1191/1478088706qp063oa>
- Bull, D. A. (2026). Artificial Intelligence as a “mixed Blessing” in online higher education: a human-centered AI-teaching presence (HC-AI-TP) perspective. *International Journal of Computer Science and Information Technology Research*, 14(1), 88–103. <https://doi.org/10.5281/zenodo.18458137>
- Chen, T., & Sun, S. (2025). Evaluating automated evaluation systems for spoken English proficiency: An exploratory comparative study with human raters. *PLoS ONE*, 20(3), Article e0320811. <https://doi.org/10.1371/journal.pone.0320811>

- Dakhil, T. A., Karimi, F., Al-Jashami, R. A. U., & Ghabanchi, Z. (2025). The effect of Artificial Intelligence (AI)-mediated speaking assessment on speaking performance and willingness to communicate of Iraqi EFL learners. *International Journal of Language Testing*, 15(2), 1–18. <https://doi.org/10.22034/ijlt.2024.486564.1383>
- Duong, T., & Suppasetseree, S. (2024). The effects of an Artificial Intelligence voice Chatbot on improving Vietnamese undergraduate students' English speaking skills. *International Journal of Learning, Teaching and Educational Research*, 23(3), 293–321. <https://doi.org/10.26803/ijlter.23.3.15>
- Farooqi, S. H. (2025). Efficacy of AI-generated feedback by SmallTalk2Me for improving speaking skill of Saudi EFL learners. *Forum for Linguistic Studies*, 7(3), 714–728. <https://doi.org/10.30564/fls.v7i3.8294>
- Lam, K. N. (2024). Exploring students' perceptions of AI-powered assistants in enhancing English speaking proficiency. *International Journal of Innovative Science and Research Technology*, 9(9), 1299–1307. <https://doi.org/10.38124/ijisrt/IJISRT24SEP792>
- Lincoln, Y. S., & Guba, E. G. (1985). *Naturalistic inquiry*. SAGE Publications.
- Luckin, R. (2017). Towards Artificial Intelligence-based assessment systems. *Nature Human Behaviour*, 1, 0028. <https://doi.org/10.1038/s41562-016-0028>
- Nguyen, H. H. (2025). AI for English speaking practice: A study of effectiveness and engagement among Vietnamese university learners. *Australasian Journal of Educational Technology*, 41(3), 45–61. <https://doi.org/10.14742/ajet.10050>
- Nguyen, T. X. (2024). English majors' perceptions of AI tool application in English language learning at tertiary level in Vietnam. *Journal of Knowledge Learning and Science Technology*, 3(1), 179–193. <https://doi.org/10.60087/jklst.vol3.n1.p193>
- Park, S. Y., & Shin, Y. (2023). A comparative study between human raters and Chatgpt in l2 speaking performance assessment. *English language teaching*, 35(4), 63–83. <https://doi.org/10.17936/pkelt.2023.35.4.004>
- Pimentel, J. (2019). Some biases in likert scaling usage and its correction. *International Journal of Sciences: Basic and Applied Research (IJSBAR)*, 45, 183–191.
- Tat, C. (2025). Using AI transcription-translation tools to improve speaking skills for the IELTS Test. *EDC Proceedings 2025, 20th Education and Development Conference*, 83–94. <https://doi.org/10.52987/edc.2025.006>
- Vo, D. Q., & Bui, D. C. T. (2024). Enhancing English speaking skills among English major students: The role of artificial intelligence-based applications. *Proceedings of the 12th OpenTESOL International Conference 2024: Nurturing Open Minds, Shaping Inclusive Futures in Language Education*, 211–220.
- Vu, T. H. D., Le, T. P. T., Nguyen, T. H., & Tong, V. B. (2025). Comparative study of feedback on IELTS speaking performance: Pre-service teachers versus automatic speech recognition applications. *VNU Journal of Foreign Studies*, 41(3), 26–43. <https://doi.org/10.63023/2525-2445/jfs.ulis.5516>
- Vygotsky, L. S. (1978). *Mind in society: the development of higher psychological processes*. Harvard University Press.
- Wang, Z., & Zhou, Z. (2025). the application classification of Artificial Intelligence in second language oral learning and its impact on oral performance: A systematic review. *Lecture Notes in Education Psychology and Public Media*, 99, 210–218. <https://doi.org/10.54254/2753-7048/2026.ht29552>
- Zou, B., Du, Y., Wang, Z., Chen, J., & Zhang, W. (2023). An investigation into Artificial Intelligence speech evaluation programs with automatic feedback for developing EFL Learners' Speaking Skills. *SAGE Open*, 13(3), 1–8. <https://doi.org/10.1177/21582440231193818>