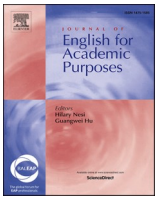
Journal of English for Academic Purposes 75 (2025) 101505

Contents lists available at ScienceDirect

Journal of English for Academic Purposes

journal homepage: www.elsevier.com/locate/jeap



Evaluating an AI speaking assessment tool: Score accuracy, perceived validity, and oral peer feedback as feedback enhancement

Xu Jared Liu [a,*], Jingwen Wang [b,**], Bin Zou [c]

[a] English Language Centre, School of Languages, Xi'an Jiaotong-Liverpool University, No. 111 Ren'ai Road, Suzhou Industrial Park, Suzhou, Jiangsu Province, 215123, P.R. China
[b] Department of Languages, Cultures and Linguistics, Faculty of Humanities and Arts, University of Southampton, B65, Avenue Campus, Highfield Road, Southampton, SO17 1BF, UK
[c] Department of Applied Linguistics, School of Humanities and Social Sciences, Xi'an Jiaotong-Liverpool University, No. 111 Ren'ai Road, Suzhou Industrial Park, Suzhou, Jiangsu Province, 215123, P.R. China

ARTICLE INFO

Handling Editor: Guangwei Hu

Keywords:
Adaptive Comparative Judgment (ACJ)
AI-assisted speaking assessment
English for Academic Purposes (EAP)
Human-computer collaboration
Oral peer feedback

ABSTRACT

Artificial Intelligence (AI) has significantly transformed language learning approaches and outcomes. However, research on AI-assisted English for Academic Purposes (EAP) speaking classrooms remains sparse. This study evaluates "EAP Talk", an AI-assisted speaking assessment tool, examining its effectiveness in two contexts: controlled tasks (Reading Aloud) that elicit non-spontaneous speech, and uncontrolled tasks (Presentation) that generate spontaneous speech. The research assessed accuracy and validity of EAP Talk scores through analysing 20 Reading Aloud and 20 Presentation recordings randomly selected from a pool of 64 undergraduate students. These recordings were graded by five experienced EAP teachers using Adaptive Comparative Judgment (ACJ) – a comparative scoring method – and the traditional rubric rating approach. Acknowledging the limitation of EAP Talk in providing scores without detailed feedback, the study further investigated its perceived validity and examined oral peer feedback as a complementary enhancement strategy. Semi-structured interviews with four students were conducted to investigate their perceptions of the AI-assisted assessment process, focusing on the benefits of EAP Talk in enhancing learning, its limitations, and the effectiveness of oral peer feedback. Scoring concordance analysis shows that EAP Talk performs well in the controlled task but less so in the uncontrolled one. Content analysis on the interview data reveals that EAP Talk facilitates student confidence and positively shapes learning styles, while oral peer feedback markedly improves speaking skills through effective human-computer collaboration. The study calls for more precise AI assessments in uncontrolled tasks and proposes pedagogical strategies to better integrate AI into EAP speaking contexts.

* Corresponding author.
** Corresponding author.
E-mail addresses: Xu.Liu@xjtlu.edu.cn (X.J. Liu), Jingwen.Wang@soton.ac.uk (J. Wang), Bin.Zou@xjtlu.edu.cn (B. Zou).

https://doi.org/10.1016/j.jeap.2025.101505
Received 2 September 2024; Received in revised form 16 March 2025; Accepted 18 March 2025
Available online 7 April 2025

1. Introduction

Artificial Intelligence (AI) has extensively influenced various educational sectors, particularly in language teaching and learning (Sharadgah & Sa'di, 2022; Yang et al., 2022). It introduces tools that enhance teachers' and learners' skills and knowledge, enabling them to effectively adapt and implement pedagogical ideas (Pokrivcakova, 2019). AI is instrumental in developing and refining tangible resources such as task design, brainstorming, and planning, and it significantly reduces teachers' workloads by automating feedback, especially in large classes (Dakakni & Safa, 2023; Liang et al., 2021). Furthermore, it facilitates the creation of personalised learning materials tailored to individual learners' interests and needs (Chen et al., 2021; Moussalli & Cardoso, 2020). Consequently, AI-enhanced language learning tools have positively impacted intangible aspects of language learning, such as student motivation and attitudes (Jeon, 2022; Ji et al., 2022).

Recognising these evident advantages, research into the application of AI in language education contexts has increasingly gained attention. Recent studies have mainly focused on the effects of AI-assisted tools on the speaking outcomes of English as a Foreign Language (EFL) learners, investigating how these tools shape student learning methodologies and the challenges posed by integrating AI in classrooms (Fathi et al., 2024; Khasawneh, 2023; Zou et al., 2024). Additional research has explored learners' emotional and social attitudes towards AI tools and their subsequent influence on speaking practices (Huang & Zou, 2024; Zou et al., 2023). However, a considerable gap remains in understanding the reliability of AI-assisted tool outputs, such as automatic assessment scores and feedback. Additionally, their advantages and limitations in influencing learning behaviours and experiences in English for Academic Purposes (EAP) classrooms warrant further investigation.

To address this research gap, the current study employs EAP Talk, an AI-assisted prototype application, to rigorously examine its integration into the speaking assessment process in university-level EAP classrooms. Initially, the study evaluates the validity and accuracy of scores generated by EAP Talk, assessing how well these scores correlate with those from human markers. We then explore students' perceptions of using such AI-assisted tools, identifying both advantages and disadvantages to determine the appropriateness of their implementation in EAP settings. Additionally, to improve the feedback effectiveness of AI tools, we implement a pedagogical strategy that incorporates oral peer feedback alongside AI feedback, examining the efficacy of this AI-human feedback combination.

To validate the scoring quality of EAP Talk, we conduct a tripartite verification, comparing AI-generated scores with those derived from Adaptive Comparative Judgment (ACJ) and analytical rubric rating. ACJ employs human judges who evaluate paired audio responses and make dichotomous decisions regarding their relative quality, thus producing standardised quality estimates along a proficiency continuum (Han & Xiao, 2022). This method is selected for its high reliability, which supplements the limitations of rubric ratings and provides comprehensive validation evidence for EAP Talk scores. Additionally, we carry out semi-structured interviews to examine students' perceptions on the integration of EAP Talk into practice and the perceived usefulness of oral peer feedback in EAP speaking classrooms.

2. Literature review

2.1. Applications of AI-assisted tools in L2 speaking classrooms

Research on AI tools in L2 speaking classrooms can be broadly categorised into two areas: exploring multiple AI resources and assessing the effectiveness of specific AI-assisted tools. For example, Madhavi et al. (2023) recognised the role of technological advancements in facilitating the shift from teacher-centred to student-centred approaches in L2 classrooms. Their research primarily examined learners' challenges with Information and Communication Technology (ICT) while articulating the advantages of AI tools in enhancing speaking skills. Shazly (2021) noted that although AI technologies support EFL learners' linguistic abilities, they do not clearly reduce speaking-related anxiety. However, these studies often overlook the reliability and validity of the feedback or materials produced by AI applications and fail to propose specific strategies for maximising the efficiency of AI tools in speaking instruction. Huerta-Macías (1995, p. 10) emphasised that "the trustworthiness of a measure consists of its credibility and auditability". Echoing this perspective, Brown and Hudson (1998) advocated for alternative assessments as a comprehensive method, emphasising the necessity for users to thoroughly evaluate their experimental procedures to ensure result trustworthiness and consistency. In this study, ACJ functions as an alternative assessment to examine the trustworthiness of AI-generated scores, thereby validating EAP Talk as a reliable measure. Thus, to encourage a positive washback effect of AI-enabled outputs on teaching practices and learners' behaviours, such as addressing specific weaknesses identified through AI-generated scores and feedback, these printouts must be underpinned by a reliable and valid assessment. Accordingly, despite technological advances in assessment methods, providing reliable and valid outputs remains essential for successful language acquisition, requiring rigorous evaluation through various methodologies.

Research on specific AI tools often reveals similar limitations: these studies primarily focused on students' perceptions of AI tools while neglecting the quality of the feedback these tools provided on students' productions. For instance, Lin and Mubarok (2021) analysed the impact of a mind map-guided AI chatbot, "Replika", in a university flipped English speaking classroom. Their study illustrated how Replika facilitated student interactions both in and outside the classroom yet failed to assess the efficacy of the feedback it provided. Likewise, He et al. (2024) investigated the use of EAP Talk in undergraduate EAP classrooms using the Technology Acceptance Model, noting that students valued its personalised feedback. Moreover, Lee and Jeon (2024) investigated the use of a self-developed voice-controlled conversational agent (VCA) as a language partner. Their study adopted Epley et al.'s (2007) three-factor theory of anthropomorphism – comprising elicited agent knowledge (humans' perception of human and non-human agents), effectance (motivation to interact with an agent), and sociality (humans' desire for social interaction) – which addresses the human tendency to attribute human traits to non-human entities. This framework was used to investigate Korean primary school

students' perception of this artificial agent. Their study revealed that the majority of these students perceived the VCA as human-like language partners. It is essential to recognise that the primary aim of assessing users' perceptions is to enhance these tools based on user feedback, thereby optimising their effectiveness and better aligning them with users' needs. Therefore, examining the effectiveness of AI-generated feedback is increasingly pivotal, as it profoundly affects users' attitudes towards using AI tools.

Furthermore, recent studies have demonstrated numerous benefits of AI-assisted language classrooms, including enhancing interactive opportunities, reducing learners' anxiety (Jeon, 2022), and promoting self-regulated learning styles (Qiao & Zhao, 2023). It is worth mentioning that Ngo and Hastie (2025) proposed integrating AI literacy into EAP programmes, defining it as the competencies to critically evaluate, appropriately communicate with, and ethically use AI applications. Their study reported that AI-integrated EAP modules positively transformed students' ability to critically evaluate generative AI production, enhanced their confidence in using various AI tools, and deepened their understanding of ethical AI application. Despite these advancements, documentation on AI-assisted EAP classrooms is still limited, with existing research primarily examining the impact of AI tools on language instruction (Ngo & Hastie, 2025), EAP writing pedagogies (van de Poel & Gasiorek, 2024), and users' attitudes towards these technologies (Zou et al., 2023, 2024). The effectiveness of feedback from AI speaking tools and viable pedagogical strategies for integrating AI into speaking instruction is rendered underexplored. As a result, this study aims to fill this gap by evaluating the effectiveness of AI-generated speaking feedback from learners' perspectives and proposing a pedagogical approach for AI-assisted speaking classrooms based on an analysis of these insights.

2.2. Adaptive Comparative Judgment

Comparative Judgment (CJ) is a technique originating from psychophysics, first proposed by Thurstone in 1927 to assess perceptual attributes such as "greyness" or "loudness" (Pollitt, 2012a). This method involves judges who evaluate two stimuli based on a holistic criterion to determine the better one. Through a series of such binary decisions, the evaluations undergo statistical analysis using the Bradley-Terry-Luce model (Bradley & Terry, 1952; Luce, 1959), an adaptation of the Rasch model, yielding standardised estimates for each stimulus. CJ has been adopted in educational assessment contexts as an alternative to traditional rubric scoring methods, praised for its high reliability and accuracy (Jones & Davies, 2024). Theoretically, the reliability of CJ stems from its ability to generate a consensus among judges. The validity of CJ-based methods is underpinned by the premise that humans are inherently more adept at making comparative judgments than absolute ones (Jones & Inglis, 2015).

The full potential of CJ was realised when Pollitt (2012a, 2012b) proposed enhancing it with computational adaptive algorithms to refine the pairing procedure. ACJ increases assessment efficiency over traditional CJ by pairing stimuli of similar estimated quality, thereby maximising the informational value of each comparison, as Pollitt (2012a) noted that comparisons between stimuli of closely matched quality produce more "information" than those between widely differing ones. Utilising this algorithm, ACJ first estimates the value of each stimulus from initial judgments, and then adaptively selects similar pairs for subsequent comparisons. Pollitt (2012b) emphasised that ACJ maintains all the benefits of traditional CJ, including high reliability, validity, and effective reduction of biases among judges, making it especially effective for evaluating complex portfolios or performances, such as L2 speaking production.

However, ACJ has seen limited application in educational assessment research compared to CJ, particularly in the field of L2 assessment where it is primarily used for evaluating textual rather than spoken outputs. Paquot et al. (2022) demonstrated strong reliability and a promising correlation with original proficiency levels in their investigation of crowdsourced ACJ for text-based L2 proficiency assessments. Similarly, Sherman et al. (2022) applied ACJ to evaluate student projects incorporating design thinking into an English composition course, thereby emphasising advancements in students' rhetorical awareness. Although research on spoken stimuli is scarce, the efficacy of ACJ in assessing spoken products has been initially established. Newhouse and Cooper (2013) compared the efficacy of analytical rubric marking, ACJ, and individual teacher assessments in Italian language production, confirming the high reliability and strong correlation of ACJ with established assessment methods. Given the proven advantages of ACJ and its underexplored potential in spoken production assessment, this study opts to utilise ACJ to validate scores from EAP Talk, thereby contributing to the broadened applications of this method in educational contexts.

2.3. Peer feedback in L2 classrooms

Oral peer feedback, a core activity in language classrooms, was defined as "a communication process through which learners engage in dialogues about performance and standards" (Liu & Carless, 2006, p. 280). The goals of peer feedback are comprehensive, designed to provide immediate checks on performance or understanding, enhance students' self-regulation and evaluative skills, facilitate collaborative learning, and develop self-confidence, accountability, and critical thinking (Gielen et al., 2010; Han & Xu, 2020; Kumar et al., 2023; Schünemann, 2017; van Popta et al., 2017). Peer feedback can be delivered in written or oral formats, and its effectiveness largely depends on students' feedback literacy, which Sutton (2012, p. 31) defined as "the ability to read, interpret, and use written feedback effectively".

Students with high feedback literacy view teacher feedback as an active process and engage deeply with the information provided (Molloy et al., 2020). However, Carless and Boud (2018, p. 1316) considered it "unrealistic and inefficient" for teachers to provide extensive comments to a large number of learners. In response to these practical teaching constraints and pedagogical challenges, peer feedback has been adopted. Timed and guided peer feedback sessions conducted in class promote dynamic collaboration and encourage students to make judgments based on their own insights. Carless (2020) emphasised the importance of actionable responses, noting that such sessions allow students to immediately act on feedback and address their speaking weaknesses.

Extensive research has investigated the impact of peer feedback on L2 writing (Ruegg, 2015; Storch, 2017; Yu & Lee, 2016).

However, there is comparatively less research concerning the effects of peer feedback on L2 speaking in EAP classrooms, particularly in higher education contexts. The existing studies that examined peer feedback on enhancing speaking skills have predominantly focused on specific forms. For example, Rodríguez-González and Castañeda (2016) examined how guided peer feedback in L2 Spanish-speaking practice influenced language complexity, global accuracy and fluency from an ontological perspective. Notably they found no improvement in students' language competence. Moreover, Chien et al. (2020) focused on EFL contexts, comparing English speaking ability, motivation, high-order thinking skills and English learning anxiety between students using peer assessment-based spherical video-based virtual reality (SVVR) and those using non-peer-assessment-based SVVR approaches. Their findings revealed that students who received feedback through the peer-assessment-based SVVR method outperformed those receiving that from the non-peer-assessment-based SVVR method, showing greater motivation, enhanced critical thinking skills and reduced English learning anxiety. Additionally, Wu and Miller (2020) and Yeh et al. (2019) investigated mobile-assisted and online blog-based peer feedback in English for Specific Purposes (ESP) and EFL classrooms respectively, demonstrating the positive effects of these peer feedback forms on students' speaking performance. The current study purports to expand the existing literature by assessing the influence of peer feedback as supplementary support in technology-enhanced EAP speaking classrooms.

2.4. Oral peer feedback on speaking performance and sociocultural perspectives

Sociocultural theory in L2 acquisition stresses the importance of engaging learners in a "dialectic interaction of two ways of creating meaning in the world (interpersonally and intrapersonally)" to master linguistic properties and contextual communicative skills (Lantolf & Pavlenko, 1995, p. 110; Vygotsky & Cole, 1978). This approach focuses on the interplay between participants and meaning-making, mediated through negotiation within social and cultural contexts (Carless & Young, 2023; Esterhazy, 2018). Ranta and Lyster (2007) categorised oral corrective feedback into two main types: reformations, which include recasts and explicit corrections, and prompts, which are implicit signals that encourage learners to self-repair. Building on the principles of sociocultural theories and the dimensions of such feedback types, Lyster et al. (2013) observed that "learners feel comfortable testing their linguistic hypotheses following corrective feedback from their classmates" (p. 29) and confirmed that "peer interaction provides a context where learners engage in interactional moves that are conducive to L2 development" (p. 27).

The decision to implement oral peer feedback in AI-assisted L2 classrooms is grounded both theoretically and practically. In EAP classrooms, oral peer feedback is crucial for developing student feedback literacy, which Carless and Boud (2018) identified through four key aspects: (1) appreciating feedback processes, (2) making academic judgments, (3) managing affect, and (4) taking action. These aspects are particularly pertinent in teaching environments tailored to meet specific student needs. Furthermore, sociocultural theory advocates for oral over written feedback, emphasising the importance of meaning-making through mediated negotiation and interaction among learners. Oral peer feedback also encourages students to consider or respond to feedback promptly, in line with Carless and Boud's (2018) emphasis on active engagement in feedback literacy. Drawing on these sociocultural approaches, this paper proposes a pedagogical strategy that addresses potential deficiencies in technology-enhanced EAP speaking classrooms.

To address the research gaps identified, this study is guided by the following research questions:

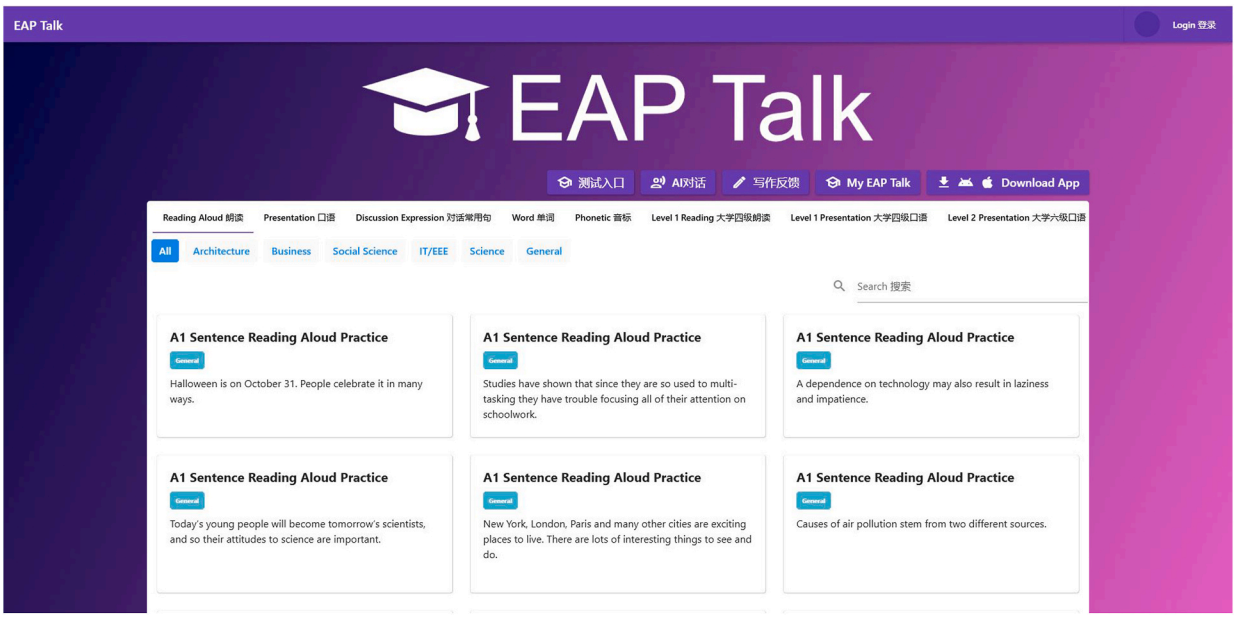


Fig. 1. The user interface of EAP Talk.

- RQ 1: How closely do scores from EAP Talk align with outcomes assessed by human raters using both ACJ and rubric scoring, thereby exploring the criterion-related validity of EAP Talk?
- RQ 2: What are students' perceptions of using EAP Talk in their speaking learning process, specifically regarding its perceived validity (i.e., whether the EAP Talk accurately measures what it claims to measure)?
- RQ 3: How does oral peer feedback, when used as supplementary support to feedback generated by EAP Talk in the classroom, enhance students' speaking proficiency?

3. Methodology

3.1. Introduction to EAP Talk

EAP Talk (https://www.eaptalk.com) is a free online AI-powered English speaking assessment system, tailor-made for EAP speaking contexts. It supports a variety of EAP-oriented speaking activities including Reading Aloud, Presentations, Discussion Expressions, and general tasks like Word, Phonetic, and AI Chatting (refer to Fig. 1). Developed with cutting-edge technologies such as natural language processing, speech recognition, and statistical modelling, EAP Talk (2.0) provides users with automated, intelligent feedback (e.g., pronunciation errors) to prompt self-regulated speaking practices. The system is accessible via its official website, mobile applications for iOS and Android, and a WeChat mini-programme.

3.2. Research context

This study was conducted over a 14-week intermediate-level EAP module at a transnational university, an institution offering educational programmes where students study in a country different from that of the awarding institution. The module aimed to develop first-year undergraduate students' language skills and academic competence through a mix of offsite and onsite seminars. Initially, students engaged in English for General Academic Purposes and foundational courses, with discipline selection occurring toward the end of the academic year. In this context, three student groups ($n = 64$) participated in weekly onsite speaking activities, utilising EAP Talk for 20–30 min of each 100-min session over nine weeks. To facilitate effective peer feedback, students were paired and participated in a brief, 10-min training session at the start of each seminar, led by an EAP teacher. During the training, students practiced orally evaluating two samples: a Reading Aloud task and a Presentation task. They used the criteria from their final speaking exams' rubric which included specific core requirements for Reading Aloud (e.g., content) and Presentation (e.g. time, use of visual aids, structure and idea development) tasks, as well as fluency, cohesion, vocabulary, grammar, and pronunciation components. After the training, pairs exchanged feedback orally upon reviewing their samples stored in EAP Talk. During these exchanges, discrepancies in judgment about peers' speaking performance often emerged due to differing interpretations of the rubric. To address the discrepancies, the teacher monitored the feedback sessions, intervening to clarify misunderstandings and explain critical aspects of the rubric, such as pronunciation, fluency, and discourse-level organisation. Unlike written peer feedback, which requires time for note-taking and often further oral clarification, oral peer feedback allows for immediate preparation and delivery during the review of audio recordings or even direct demonstration of correct pronunciation and expression. This immediacy optimises feedback uptake and improvement. Oral peer feedback ensures a more efficient and focused feedback process, as it is both more immediate and directly addresses students' specific needs. As such, it is more likely to be utilised by students to address their weaknesses than written feedback, which may not always be read or acted upon.

3.3. Selection of student recordings

Twenty recordings from each of the Reading Aloud and Presentation tasks were randomly selected from the 64 undergraduate students participating in the summative assessment via EAP Talk. This platform prompted students to record responses for these two types of tasks. In the Reading Aloud task, students were required to read a 22-word sentence, with their responses varying from 13 to 25 s. For the Presentation task, students expressed their opinions on assigned questions, with response durations ranging from 15 to 102 s.

3.4. Operationalisation of ACJ

Five judges ($M = 30.00$ years, $SD = 2.28$), all university EAP teachers with experience ($M = 2.60$ years, $SD = 0.80$) in higher

Table 1
Profiles of the ACJ judges.

Pseudonym	Gender	Age	L1	Qualification	University teaching experience
Jane	female	28	Mandarin	MSc in TESOL	4 years
Mike	male	31	Mandarin	MSc in Applied Linguistics	2 years
Lyn	female	28	Mandarin	MSc in Media and Communication	3 years
James	male	34	Mandarin	MA in TESOL	2 years
Anna	female	29	Mandarin	PhD in Applied Linguistics	2 years

education teaching and student coursework assessment, were recruited to assess recordings using ACJ. Table 1 provides comprehensive profiles for each judge.

ACJ was implemented through the web-based platform "RM Compare" (https://compare.rm.com). Judges accessed their individual interfaces via a unique invitation link. During the assessment sessions, two recordings were displayed side-by-side for comparison. Judges evaluated these recordings and selected the one that demonstrated better speaking proficiency by clicking the "A" (left side) or "B" (right side) button. All decisions were automatically logged in the RM Compare system for subsequent statistical analysis. Before each judging session, judges received approximately 15 min of training to familiarise them with the ACJ principles, the functionalities of the RM Compare system, and the detailed operation procedures.

For n recordings, there are $\binom{n}{2} = \frac{n(n-1)}{2}$ unique possible pairings. With 20 recordings selected per item type, the total number of possible pairwise comparisons was calculated to be 190. Each judge was tasked with assessing 38 pairs per task. To better understand the cognitive processes of judges during decision-making and to gather data supporting the scoring validity of ACJ – specifically, whether judges' criteria effectively measure the speaking constructs — we invited four judges to participate in Think Aloud sessions with the second author of this article (the author whom the judges had not come to know). Unfortunately, one judge was unavailable due to personal reasons. During these sessions, judges articulated the criteria used while evaluating the initial 15 pairs for the Reading Aloud task. Recognising that this process can be tedious and potentially lead to cognitive fatigue, we adjusted the approach for the Presentation task. Given its more complex evaluative dimensions, judges assessed only the first 10 pairs to ensure sustained concentration and accuracy. The Think Aloud sessions were conducted in the judges' L1 to facilitate smooth expression and elicit deeper insights. The Think Aloud Protocols (TAPs) were initially transcribed using iFlytec (https://www.iflyrec.com/), translated into English, and subsequently checked by the first author to ensure accuracy. The judges were then required to complete the remaining comparisons independently within one week.

3.5. Rubric scoring

Two EAP teachers, each with approximately ten years of experience in assessing L2 speaking, rated all 40 recordings (20 for each task type) using the TOEFL iBT Integrated Speaking scale rubric. This rubric is composed of four five-point subscales: general description (relevance and support), delivery (speech fluidity and pronunciation), language use (grammar and vocabulary accuracy), and topic development (organisation and cohesion). However, specific dimensions, such as topic development, were excluded for the Reading Aloud task. This exclusion reflects the Reading Aloud task's design to elicit non-spontaneous speech, as the content is pre-provided and identical for all speakers. Consequently, judges cannot assess speakers' ability to generate original discourse, which precludes evaluation of discourse-level competencies such as general description and topic development. Moreover, the TOEFL scale was chosen due to its international recognition and established validity in assessing speaking proficiency. Each rater was required to score speech samples according to the subscales, with the final scores for each recording that were calculated by averaging these subscores. These scores were subsequently correlated with AI scores and ACJ results to establish criterion validity evidence for EAP Talk.

3.6. Semi-structured interview

To investigate users' perceptions of EAP Talk and the perceived effectiveness of oral peer feedback in enhancing their learning outcomes, four participants were recruited for individual semi-structured interviews. These participants were randomly selected from the same pool of 64 to ensure that the sample was representative of the broader group (Adeoye, 2023). As noted, all these participants were students in EAP courses who had used EAP Talk for nine weeks and participated in extensive oral peer feedback training provided by their teachers. Consequently, they had developed a thorough understanding of EAP Talk's utility in facilitating speaking proficiency and had gained significant experience in conducting oral peer feedback. Details of the four students' backgrounds and their after-class use of EAP Talk are presented in Table 2. Furthermore, all participants were native Mandarin speakers with intermediate English proficiency (CEFR B1) and were expected to reach CEFR B2 by the end of the module.

Before initiating the interviews, ethical approval was obtained from the first author's institution. Participants were instructed to read the information sheet and sign the consent form before their interviews. All of them were informed that their interviews would be audio recorded and anonymised. To facilitate more in-depth insights into their experiences with EAP Talk both inside and outside the classroom, interviews were conducted in the participants' L1. The sessions were facilitated by the second author, who had no prior relationship with the participants via an online meeting platform. Audio recordings were transcribed verbatim using iFlytec software and the transcripts were subsequently translated by one author and calibrated by another.

Table 2
Students' backgrounds and their use details of EAP Talk after seminars.

Pseudonym	Gender	Discipline	Frequency of EAP Talk use (per week)	Duration of EAP Talk use (per time)
Leo	Male	Business	3 times	around 30 min
Angela	Female	Advanced Technology	2 times	around 30 min
Jean	Female	Advanced Technology	2 times	around 10 min
Craig	Male	Science	5 times	around 25 min

3.7. Data analysis

To address RQ1 on the accuracy of EAP Talk scores, we first collected reliability and validity evidence for ACJ results, subsequently using these as benchmarks for validating EAP Talk scores. Two statistical evidence types were provided for the reliability of ACJ: infit statistics, which measure consistency among judges and recordings, and Scale Separation Reliability (SSR) for assessing inter-rater reliability. These indices were automatically analysed in the RM Compare system using the Bradley-Terry-Luce model. We further established the validity of ACJ through criterion-related validity, demonstrated by correlations between ACJ results and rubric scores of both task types. Additionally, we analysed TAPs using manifest content analysis to enhance our understanding of scoring validity. The first and second authors developed the coding scheme collaboratively based on the criteria of various speaking rubrics, achieving an inter-coder reliability of 0.93, with any discrepancies resolved through discussion. Finally, we correlated EAP Talk scores for assessed tasks with both ACJ and rubric scores to evaluate their alignment with human assessments.

To address RQ2 on the perceived validity of EAP Talk and RQ3 regarding the effectiveness of oral peer feedback, we analysed interview transcripts (totalling 18,403 words) using Pavlenko's (2007) autobiographic narrative inquiry approach, specifically focusing on linguistic biographies. This method explores speakers' life histories to understand how languages are acquired, used, or abandoned, drawing on cognitive, textual, and discursive theoretical frameworks. Accordingly, we analysed interview data within Pavlenko's frameworks, with the first author synthesising relevant anecdotes into generic, summative themes. To ensure coding credibility, we employed Dovetail (https://dovetail.com/#analysis), a thematic analysis software which automates themes in qualitative texts. This tool is particularly suitable for organising narrative data retrieved from interviews. The second author independently assessed the alignment of generic categories, considering detailed aspects of the participants' interactions, academic backgrounds, and possible emotions expressed through tones. Upon establishing an inter-coder reliability of 0.89, any discrepancies were resolved through discussion among both authors.

4. Results

4.1. RQ 1: how closely do scores from EAP Talk align with outcomes assessed by human raters using both ACJ and rubric scoring, thereby exploring the criterion-related validity of EAP Talk?

Before addressing RQ1, we established the reliability and validity of ACJ to use its results as a benchmark for verifying EAP Talk scores. Tables 3 and 4 summarise the distribution of acceptable items and judge performance metrics. According to Pollitt's (2012a) guidelines, the threshold for infit values should be no more than the mean infit value plus two standard deviations. Results indicate that the majority (95 %) of the recordings in the Reading Aloud task, and all recordings in the Presentation task, were consistently judged. Furthermore, all judges showed significant consistency in their decision-making across various task types.

Additionally, the SSRs for the Reading Aloud and Presentation tasks reached 0.95 and 0.91, respectively, indicating a high level of inter-rater reliability for the scales produced by ACJ. These results exceed the standards recommended by Verhavert et al. (2019), which establish an SSR above 0.70 for low-stakes research settings and over 0.90 for high-stakes assessments.

The correlation coefficients among EAP Talk scores, ACJ results, and rubric scores across different task types are presented in Tables 5 and 6. To provide criterion-related validity evidence for ACJ, the average Pearson's r between ACJ results and rubric scores from two raters across these tasks was 0.72 ($p < 0.05$), demonstrating a significant correlation. This indicates that the two human scoring methods – ACJ and rubric rating – assessed the recordings with notable consistency.

Furthermore, to collect evidence of scoring validity, TAPs documented a total of 100 decisions made by four judges conducting a think-aloud judgment on 15 pairs from the Reading Aloud task and 10 pairs from the Presentation task. Content analysis identified 385 codings linked to 27 criteria, detailed in Tables 7 and 8. On average, each judge employed four assessment criteria per decision. As shown in Table 7, the 20 criteria for assessing the Reading Aloud task were grouped into two main categories, aligning with two components of the TOEFL iBT rubric. The ACJ judges omitted the other two components due to the controlled conditions of the Reading Aloud task, which restrict speakers from developing their own content on a specified topic, as previously mentioned. In the Presentation task, judges used 17 criteria divided into three TOEFL iBT scale categories. The absence of "General Description" might stem from the researchers' interpretation of its significant overlap with "Topic Development" on the TOEFL scale.

A notable pattern in Table 7 shows that the majority of the codes (91.45 %) associated with the Reading Aloud task focused on pronunciation (articulation accuracy), which is a fundamental construct this task aims to assess. For the Presentation task (see Table 8), judges emphasised not only pronunciation (48.94 %) but also the speakers' ability to extensively develop and clearly and coherently articulate their ideas (36.88 %). This focus corresponds well with the intended constructs of the Presentation task type.

Consequently, the evidence substantiates the validity of ACJ results as a scoring benchmark for this study. Nonetheless, correlations vary across tasks and scoring approaches: for the Reading Aloud task, EAP Talk scores showed a correlation of 0.75 with ACJ results and an average correlation of 0.67 with rubric scores (from the two raters) - values that are acceptable though not ideal. For the

Table 3
Recordings were consistently evaluated by judges.

Task Type	Threshold	Item infit range	Percentage of recordings were judged consistently (n = 20)
Reading-Aloud	≤1.65	0.47–1.80	95 %
Presentation	≤1.49	0.60–1.44	100 %

Table 4
Distribution of judge infit values.

Task Type	Threshold	Judge infit range	Percentage of judges performed consistently (n = 5)
Reading Aloud	≤1.43	0.85–1.38	100 %
Presentation	≤1.54	0.70–1.29	100 %

Table 5
Correlations among three scoring methods (Reading Aloud task).

	EAP Talk	ACJ	Rubric Scoring (first rater)	Rubric Scoring (second rater)
EAP Talk	–	0.75	0.59	0.74
ACJ		–	0.89	0.62
Rubric Scoring (first rater)			–	0.77

Note: *Correlation is significant at the 0.01 level.*

Table 6
Correlations among three scoring methods (Presentation task).

	EAP Talk	ACJ	Rubric Scoring (first rater)	Rubric Scoring (second rater)
EAP Talk	–	0.40	0.53	0.31
ACJ		–	0.71	0.65
Rubric Scoring (first rater)			–	0.49

Note: *Correlation is significant at the 0.01 level.*

Table 7
Criteria judges referred to in assessing Reading Aloud task.

General Categories	Criteria judges referred to	No. of codings	%
Delivery	Articulation accuracy	34	91.45 %
	Speech smoothness & fluidity	31	
	Intonation	24	
	Rhythm	19	
	Intelligibility	15	
	Articulation clarity	14	
	Speech naturalness	14	
	Pause	11	
	Sentence stress	10	
	Flow of speech	10	
	Word stress	8	
	Speaking rate	8	
	Hesitation	7	
	Comprehensibility	4	
	Self-correction	2	
	Tone	1	
	Self-repetition	1	
	Accentedness	1	
General Description	Sentence completion	16	8.55 %
	Sentence mastery	4	

Presentation task, correlation were markedly lower: 0.40 with ACJ results and an average of 0.42 with rubric scores, indicating a very weak relationship. These findings suggest that EAP Talk effectively assesses controlled tasks such as pronunciation proficiency, requiring only minor adjustments. However, its performance in evaluating spontaneous speech is inadequate, pointing to a need for significant enhancements to the model design.

4.2. RQ2: what are students' perceptions of using EAP Talk in their speaking learning process, specifically regarding its perceived validity?

4.2.1. General perceptions

Among the four students, three expressed positive attitudes towards using EAP Talk for regular speaking practice in class. Specifically, Leo noted that it could enhance listening skills and suggested its beneficial use at the beginning of class for routine English practice. Similarly, Angela highlighted the availability of multiple learning materials in EAP Talk, while Craig pointed out its advantages in identifying mistakes. In contrast, Jean's lack of positive feedback may be attributed to her less frequent use of EAP Talk. The following excerpts illustrate these varied perspectives.

Table 8
Criteria judges referred to in assessing Presentation task.

General Categories	Criteria judges referred to	No. of codings	%
Delivery	Articulation accuracy	21	48.94 %
	Speech smoothness & fluidity	18	
	Intelligibility	9	
	Intonation	4	
	Sentence stress	3	
	Hesitation	3	
	Comprehensibility	3	
	Speech naturalness	3	
	Rhythm	2	
	Pause	2	
	Flow of speech	1	
Topic development	Idea development	29	36.88 %
	Coherence & cohesion	23	
Language use	Sentence structure complexity	10	14.18 %
	Vocabulary complexity	6	
	Grammaticality	2	
	Collocation use accuracy	2	

> "… We all can use it. It would be a useful tool if we used it at the beginning of class." (Leo)
>
> " … because EAP Talk can provide us with insights to shape our answers and also enrich our language resources." (Angela)
>
> "EAP Talk lets us record our voice and listen back to spot mistakes. We can also see scores, highlighted errors, and good words, all of which help improve our speaking practice." (Craig)
>
> "I don't use it very often – I mainly stick to the Reading Aloud section and rarely touch the others. However, I have used the AI-driven section for class presentations." (Jean)

4.2.2. Satisfactions towards EAP Talk

Leo and Craig acknowledged that EAP Talk generally met their expectations; however, they identified its limitations in preparing for international English tests and providing detailed feedback on pronunciation and grammar errors. For example, Leo elaborated that

> "The feedback from the app is somewhat general. Compared to apps from some domestic educational companies, it lacks specific feedback like speaking tips for liaison. Including detailed feedback on liaison could really help users enhance their fluency."

The other two students expressed greater dissatisfaction. Angela noted that the system could not recognise sounds well and provided only limited feedback. She recommended that

> "EAP Talk can offer a question-answer format or provide various answer templates that students can customize to craft their own responses."

Furthermore, Jean's dissatisfaction stemed from EAP Talk not aligning with her learning style, as she explained:

> "I prefer reading full articles or books from start to finish, and I'm not really into practicing English by topics. Instead, I like to improve my speaking by reading English books or watching British and American TV shows."

4.2.3. Comparison with other automated language learning tools

In terms of competitiveness with other tools, two students stressed the limited application context of EAP Talk. For instance,

> "EAP Talk is specifically designed for our EAP courses, focusing on English for Academic Purposes … " (Jean)
>
> "This app was developed by our university, so it might be more tailored to students here." (Craig)

Leo observed the low popularity of EAP Talk, indicating that

> "The first issue is popularity. Although the app is available and downloadable on mobile app stores, its download numbers are underwhelming. This lack of user engagement leads to limited data on user experience … "

Moreover, Angela pointed out that EAP Talk lacked engaging and interactive features. She clarified that

> "I think EAP Talk lacks multiple functions. Unlike other apps I've used, which offer short videos like public lectures, TED Talks, or movie clips, users can't dub or imitate the speech and tone from these videos. Including such features could make the learning process more enjoyable."

4.2.4. Usefulness of EAP Talk in enhancing learning

Opinions on the effectiveness of EAP Talk as a language learning tool varied. Angela and Craig appreciated its usefulness in enhancing speaking skills, especially for preparing language test materials and transitioning from lexical features to sentential structures. Conversely, Leo criticised its lack of innovation and recommended more personalised interfaces for scoring and feedback to improve its utility. In contrast, Jean, after facing repeated compatibility issues with the app, decided against its long-term use.

4.2.5. Learning styles and confidence development

Of the four students, three felt that EAP Talk did not match their preferred speaking learning styles, while one student found it aligned with his approach. Leo criticised the lack of flexible response samples for diverse speaking questions, expressing a preference for learning through imitation. Similarly, Angela noted the limited diversity of answer options. In contrast, Jean disliked the structured question types of the platform and favoured using diverse English learning apps post-class for a broader range of reading and imitation materials.

Notably, notwithstanding their varied opinions on heterogenous aspects of EAP Talk, all four students reported that it had enhanced their confidence. For example,

> "I've become much more confident. Before coming to university, I only knew how to do exercises and was scared to talk to foreigners. Now, I can communicate with international teachers, and my speaking has improved a lot." (Leo)
>
> "My confidence grew as I had more to say and clearer thoughts. My logic became more organised, which made me feel more assured." (Angela)
>
> "Yes, my confidence was improved. After all, I experienced Presentation (the speaking test task of this semester), and there is a section of Presentation in EAP Talk." (Craig)

Jean's response, though tentative, still reflected a positive influence on her confidence.

> "It might have boosted my confidence a bit, but I didn't really think much about it."

4.3. RQ3: how does oral peer feedback, when used as supplementary support to feedback generated by EAP Talk in the classroom, enhance students' speaking proficiency?

Regarding the effectiveness of oral peer feedback as a supplement to EAP Talk's feedback, two students observed that the feedback from EAP Talk was insufficient, overly general, and sometimes vague. They emphasised their preference for and greater trust in feedback from human raters. This preference is further illustrated in the following excerpts.

> "EAP Talk just gives me a score and a general idea of my English proficiency. It only marks mispronounced words, and the feedback is often vague. Honestly, I don't trust the scores it gives me." (Angela)
>
> "I trust human raters and feedback more." (Leo)

Due to these concerns, all four participating students consistently favoured human feedback and displayed positive attitudes towards oral peer feedback, indicating that these sessions provided valuable support. For example, Angela emphasised the interactive nature of peer feedback:

> "My partner's feedback is usually spot-on and interactive, allowing me to make revisions right away."

Likewise, Craig emphasised that peer feedback significantly enhanced his discourse-level organisation.

> "Peer feedback usually focuses on the content of our conversations, while EAP Talk's automatic scoring targets pronunciation and grammar. For example, in one EAP class, my partner and I practiced speaking together. His English was better, so he varied his answers a lot and gave me plenty of useful feedback."

However, two students noted that the quality of peer feedback largely depends on the proficiency level of the peer providing it.

> "I think the effectiveness of peer feedback depends on the student's English level. For beginners, it might not be very helpful because they can't provide constructive feedback. But intermediate or advanced students can share useful ideas, and I can learn from their good examples. Sometimes I even ask my dad for advice since he's lived overseas for a while and speaks English well." (Leo)
>
> "I think oral peer feedback can be effective, but it really depends on the student's proficiency. If a student is good at English, they can provide better feedback." (Jean)

5. Discussion

This paper assessed the effectiveness of EAP Talk, an AI learning and assessment tool, as a supplementary aid in an EAP course. Recognising the limitations of EAP Talk's feedback, this study further explored the benefits of integrating oral peer feedback. We initially evaluated EAP Talk's precision in assessing student responses to both controlled and uncontrolled tasks. The results show that

while EAP Talk aligns closely with human scoring approaches in controlled Reading Aloud tasks, it lacks accuracy in assessing open-ended questions (Presentation tasks), a finding identified by several researchers (Brooks, 2009; Chen et al., 2022; Zou et al., 2024). However, this finding contradicts research by Sun and Zhang (2020) and Sun (2021), which found that the iFly, another similar AI speaking assessment system, tended to overestimate performance in controlled tasks while more closely matching human evaluations in open-ended tasks. These differences may result from varied methodologies and criteria used by different AI tools to analyse student responses. These inconsistencies highlight the substantial need for enhancing AI assessment tools. To further improve EAP Talk's accuracy in the era of iterative and transformative AI, developers are encouraged to adopt explainable AI concepts in education (Khosravi et al., 2022), incorporate inputs from a wider range of speech backgrounds, and enable customisation of marking criteria to meet user needs. For now, we advise using the AI tool as an auxiliary support, ensuring human raters validate the evaluations before sharing them with students.

Overall, students expressed positive attitudes toward the integration of EAP Talk into their EAP courses. Two students particularly praised its benefits for shaping their learning behaviours. One emphasised the daily English practice it facilitated, noting its positive impact on long-term learning habits and listening skills. This aligns with findings by Qiao and Zhao (2023), who observed that AI-assisted speaking instruction ameliorates self-regulatory speaking abilities. Additionally, two students found that EAP Talk effectively helped them identify errors in their speaking practices, indirectly demonstrating the efficiency of the tool. These outcomes illustrate the potential for developing large-scale AI-based assessment systems, as Luckin (2017) suggested, benefiting a broad spectrum of stakeholders, including students, teachers, and parents. However, one participant's infrequent use of the tool and preference for recorded audio files emphasise individual differences in learning styles. While some learners thrived in structured and secure environments, others exceled in less controlled, more challenging settings.

Furthermore, two students confirmed that EAP Talk greatly improved their vocabulary learning, with one specifically appreciating its utility in preparing various speaking topics. In contrast, Jeon (2021) found that chatbot-assisted dynamic assessments provided only diagnostic insights without significantly aiding EFL learners' vocabulary acquisition. This variance could arise from individual learner differences, duration of vocabulary study, or research method constraints, among others. For example, one student was unable to conclusively evaluate the effectiveness of the tool, critiquing its lack of innovation and inadequate user data. Additionally, another student only appreciated its effectiveness over time but expressed concerns that compatibility issues nearly undermined her initial enthusiasm. Clearly, the stability and compatibility of systems play a critical role in their long-term adoption as digital learning tools.

However, the interview data have uncovered limitations in EAP Talk as well. Two students pointed out that despite its design being tailored to EAP contexts and assessments, EAP Talk lacks engaging functions and materials. Although this feedback may appear subjective, it starkly contrasts with findings by Lee et al. (2023) and Zhang et al. (2024), which suggested that AI-assisted language classrooms significantly enhance enjoyment in foreign language learning. This difference in psychological responses may be due to variations in the design of AI applications, the diversity of the learning materials used, or individual learner differences. Furthermore, these students criticised the feedback from the tool as too generic, highlighting a need for more detailed feedback from digital tools. This observation aligns with prior research that underscores the need for further exploration into the effectiveness of AI tools in fully developing L2 speaking skills (Fathi et al., 2024; Madhavi et al., 2023; Shazly, 2021; Skehan, 2009). Future research should investigate optimising AI-assisted speaking instruction and enhancing the specificity of AI-generated feedback to better serve students' needs.

Remarkably, while three out of four students noted that EAP Talk did not align with their learning styles or preferences, all reported enhanced confidence. This finding is supported by recent studies (Fathi, 2024; Rad, 2024). This boost in confidence is attributed to the simulated learning environment facilitated by EAP lecturers, where students interact with an AI figure that introduces topics and task requirements. Such interactions, which mimic a natural speaking environment, are likely instrumental in increasing student confidence. It becomes clear that reducing language anxiety involves more than just the AI tool; it results from a collective effort that includes AI technology, support from peers and lecturers, time management, and metacognitive strategies, among other factors.

Finally, the current research confirms that oral peer feedback serves as an effective supplementary aid, enhancing speaking skills and refining the use of EAP Talk in EAP speaking classrooms. The analysis reveals that oral peer feedback substantially improves constructive feedback in AI-assisted speaking activities. The effectiveness of this feedback largely hinges on the proficiency levels of peers. Despite standardised tests designed to equalise proficiency levels in EAP classrooms, discrepancies remain. Addressing this issue involves targeted training by lecturers and strategic pairing of students, allowing those with lower proficiency levels to benefit from their more skilled peers.

6. Conclusion

This study evaluated the validity of scores from EAP Talk, uncovering that the AI tool performs well in assessing controlled tasks but is less effective in evaluating spontaneous speech. Through comprehensive qualitative analysis, we examined students' perceptions of EAP Talk, an AI-driven English speaking assessment tool used in classrooms, and the effectiveness of oral peer feedback as supplementary support. Feedback from EAP Talk was generally satisfactory, but its limitations in providing adequate feedback for spontaneous speech were noted. Drawing on student feedback literacy and sociocultural learning theories, oral peer feedback was integrated to enhance the effectiveness of EAP Talk, particularly in uncontrolled tasks, and to facilitate dynamic interactions in EAP speaking classrooms. Given that speaking skills typically receive minimal formal emphasis compared to reading, listening, and writing (Bruce, 2015; Ferris & Tagg, 1996; Jamieson et al., 2013), this research emphasises the potential of AI-assisted approaches in advancing speaking instruction in higher educational contexts.

The current study also recognises several limitations. The qualitative data were collected through semi-structured interviews with only four participants, constraining the generalisability of the findings. The research examined learners' perceptions of using EAP Talk

in EAP speaking classrooms, its impact on language learning, and the effectiveness of oral peer feedback. However, these insights may not extend to other educational settings such as general ESL courses, primary or secondary English schools, or pre-sessional post-graduate EAP programmes. Notably, this research was specific to Year One EAP programmes at a transnational university, distinct from English programmes in other higher educational contexts. Additionally, the qualitative data primarily depended on students' perspectives of EAP Talk, derived from a single interview session, potentially introducing subjectivity. These perceptions could evolve over time or differ with various peer interactions.

Moreover, this study highlights two key pedagogical implications. First, drawing on Carless's (2020) emphasis on actionable student feedback literacy, it integrates sociocultural perspectives to assess the effectiveness of oral peer feedback as a supplement to AI-generated feedback in EAP speaking classrooms. This approach aims to enhance feedback literacy by addressing students' reluctance to engage with feedback (Winstone et al., 2017), providing a time-efficient strategy for the use of AI tools in educational settings. Moreover, the study responds to Carless's (2020) call for a deeper exploration of "students-as-partners" in feedback processes, opening avenues for broader research into AI-enhanced speaking instruction mediated by oral peer feedback across different EAP contexts. Areas for further investigation include discourse patterns, metacognitive strategies, learner self-efficacy, and EAP practitioners' perceptions, emphasising a significant gap in current research and presenting an important opportunity for future studies.

Funding

This study was funded by TDF (2324-R28-240) and the University Research Centre for Culture, Communication and Society (CCCS) at Xi'an Jiaotong-Liverpool University.

Ethics approval statement

This study was conducted in accordance with the ethical standards set forth by the Xi'an Jiaotong-Liverpool University Ethics Committee.

CRediT authorship contribution statement

Xu Jared Liu: Writing – review & editing, Writing – original draft, Validation, Software, Resources, Project administration, Methodology, Investigation, Formal analysis, Data curation, Conceptualization. **Jingwen Wang:** Writing – review & editing, Writing – original draft, Validation, Software, Resources, Project administration, Methodology, Investigation, Formal analysis, Data curation, Conceptualization. **Bin Zou:** Writing – review & editing, Supervision, Software, Resources, Funding acquisition, Conceptualization.

Declaration of competing interest

All authors declare no conflict of interest.

Acknowledgements

The authors express their heartfelt appreciation to all participants whose engagement was crucial to this research. We are also grateful to the handling editor, Prof. Guangwei Hu, and the reviewers for their invaluable support and constructive feedback during the manuscript revision process. The first author extends special thanks to his family, teachers, and friends, including Dr. Ian Bruce from the University of Waikato in New Zealand, Mary Webb, Malcolm Elliot-Hogg, Mary Elliot-Hogg, Paula Gibson, Jiaming Chen, Jiahang Li, Ya Zhang, Fuhong Tian, Jiin Yap, and Mingming Zhou, for their unwavering support throughout his academic journey.

References

Adeoye, M. A. (2023). Review of sampling techniques for education. *ASEAN Journal for Science Education, 2*(2), 87–94. https://ejournal.bumipublikasinusantara.id/index.php/ajsed/article/view/230.

Bradley, R. A., & Terry, M. E. (1952). Rank analysis of incomplete block designs: I. The method of paired comparisons. *Biometrika, 39*(3/4), 324–345. https://doi.org/10.2307/2334029

Brooks, V. (2009). Marking as judgment. *Research Papers in Education, 27*(1), 63–80. https://doi.org/10.1080/02671520903331008

Brown, J. D., & Hudson, T. (1998). The alternatives in language assessment. *Tesol Quarterly, 32*(4), 653. https://doi.org/10.2307/3587999

Bruce, I. (2015). *Theory and concepts of English for academic Purposes*. Palgrave Macmillan UK. https://doi.org/10.1057/978-1-349-59077-3

Carless, D. (2020). Longitudinal perspectives on students' experiences of feedback: A need for teacher–student partnerships. *Higher Education Research and Development, 39*(3), 425–438. https://doi.org/10.1080/07294360.2019.1684455

Carless, D., & Boud, D. (2018). The development of student feedback literacy: Enabling uptake of feedback. *Assessment & Evaluation in Higher Education, 43*(8), 1315–1325. https://doi.org/10.1080/02602938.2018.1463354

Carless, D., & Young, S. (2023). Feedback seeking and student reflective feedback literacy: A sociocultural discourse analysis. *Higher Education, 88*(3), 857–873. https://doi.org/10.1007/s10734-023-01146-1

Chen, J., Lai, P., Chan, A., Man, V., & Chan, C.-H. (2022). AI-Assisted enhancement of student presentation skills: Challenges and opportunities. *Sustainability, 15*(196), 1–19. https://doi.org/10.3390/su15010196

Chen, X., Zou, D., Xie, H., & Cheng, G. (2021). Twenty years of personalized language learning. *Educational Technology & Society, 24*(1), 205–222. https://www.jstor.org/stable/26977868.

Chien, S.-Y., Hwang, G.-J., & Jong, M. S.-Y. (2020). Effects of peer assessment within the context of spherical video-based virtual reality on EFL students' English-Speaking performance and learning perceptions. *Computers & Education, 146*, Article 103751. https://doi.org/10.1016/j.compedu.2019.103751

Dakakni, D., & Safa, N. (2023). Artificial intelligence in the L2 classroom: Implications and challenges on ethics and equity in higher education: A 21st century pandora's box. *Computers & Education: Artificial Intelligence, 5*, Article 100179. https://doi.org/10.1016/j.caeai.2023.100179
Epley, N., Waytz, A., & Cacioppo, J. T. (2007). On seeing human: A three-factor theory of anthropomorphism. *Psychological Review, 114*(4), 864–886. https://doi.org/10.1037/0033-295X.114.4.864
Esterhazy, R. (2018). What matters for productive feedback? Disciplinary practices and their relational dynamics. *Assessment & Evaluation in Higher Education, 43*(8), 1302–1314. https://doi.org/10.1080/02602938.2018.1463353
Fathi, J., Rahimi, M., & Derakhshan, A. (2024). Improving EFL learners' speaking skills and willingness to communicate via artificial intelligence-mediated interactions. *System, 121*, Article 103254. https://doi.org/10.1016/j.system.2024.103254
Ferris, D., & Tagg, T. (1996). Academic listening/speaking tasks for ESL students: Problems, suggestions, and implications. *Tesol Quarterly, 30*(2), 297–320. https://doi.org/10.2307/3588145
Gielen, S., Peeters, E., Dochy, F., Onghena, P., & Struyven, K. (2010). Improving the effectiveness of peer feedback for learning. *Learning and Instruction, 20*(4), 304–315. https://doi.org/10.1016/j.learninstruc.2009.08.007
Han, C., & Xiao, X. (2022). A comparative judgment approach to assessing Chinese Sign Language interpreting. *Language Testing, 39*(2), 289–312. https://doi.org/10.1177/02655322211038977
Han, Y., & Xu, Y. (2020). The development of student feedback literacy: The influences of teacher feedback on peer feedback. *Assessment & Evaluation in Higher Education, 45*(5), 680–696. https://doi.org/10.1080/02602938.2019.1689545
He, H., Zou, B., & Du, Y. (2024). Bridging the gap: Linking AI technology acceptance to actual improvements in EAP learners' speaking skills. *Computer Assisted Language Learning*. https://doi.org/10.31219/osf.io/syb62
Huang, F., & Zou, B. (2024). English speaking with artificial intelligence (AI): The roles of enjoyment, willingness to communicate with AI, and innovativeness. *Computers in Human Behavior, 159*, Article 108355. https://doi.org/10.1016/j.chb.2024.108355
Huerta-Macías, A. (1995). Alternative assessment: Responses to commonly asked questions. In J. C. Richards, & W. A. Renandya (Eds.), *Methodology in language teaching: An anthology of current practice* (pp. 338–343). Cambridge University Press.
Jamieson, J., Wang, L., & Church, J. (2013). In-house or commercial speaking tests: Evaluating strengths for EAP placement. *Journal of English for Academic Purposes, 12*(4), 288–298. https://doi.org/10.1016/j.jeap.2013.09.003
Jeon, J. (2021). Chatbot-assisted dynamic assessment (CA-DA) for L2 vocabulary learning and diagnosis. *Computer Assisted Language Learning, 36*(7), 1338–1364. https://doi.org/10.1080/09588221.2021.1987272
Jeon, J. (2022). Exploring AI chatbot affordances in the EFL classroom: Young learners' experiences and perspectives. *Computer Assisted Language Learning, 37*(1–2), 1–26. https://doi.org/10.1080/09588221.2021.2021241
Ji, H., Han, I., & Ko, Y. (2022). A systematic review of conversational AI in language education: Focusing on the collaboration with human teachers. *Journal of Research on Technology in Education, 55*(1), 48–63. https://doi.org/10.1080/15391523.2022.2142873
Jones, I., & Davies, B. (2024). Comparative judgement in education research. *International Journal of Research and Method in Education, 47*(2), 170–181. https://doi.org/10.1080/1743727X.2023.2242273
Jones, I., & Inglis, M. (2015). The problem of assessing problem solving: Can comparative judgement help? *Educational Studies in Mathematics, 89*, 337–355. https://doi.org/10.1007/s10649-015-9607-1
Khasawneh, M. (2023). Integrating AI-based virtual conversation partners in enhancing speaking skills in foreign languages: Insights from university students. *Journal of Southwest Jiaotong University, 58*(5). https://doi.org/10.35741/issn.0258-2724.58.5.43
Khosravi, H., Shum, S. B., Chen, G., Conati, C., Tsai, Y.-S., Kay, J., Knight, S., Martinez-Maldonado, R., Sadiq, S., & Gašević, D. (2022). Explainable artificial intelligence in education. *Computers & Education: Artificial Intelligence, 3*, Article 100074. https://doi.org/10.1016/j.caeai.2022.100074
Kumar, T., Soozandehfar, S. M. A., Hashemifardnia, A., & Mombeini, R. (2023). Self vs. peer assessment activities in EFL-speaking classes: Impacts on students' self-regulated learning, critical thinking, and problem-solving skills. *Language Testing in Asia, 13*(1). https://doi.org/10.1186/s40468-023-00251-3
Lantolf, J. P., & Pavlenko, A. (1995). Sociocultural theory and second language acquisition. *Annual Review of Applied Linguistics, 15*, 108–124. https://doi.org/10.1017/S0267190500002646
Lee, S., & Jeon, J. (2024). Visualizing a disembodied agent: Young EFL learners' perceptions of voice-controlled conversational agents as language partners. *Computer Assisted Language Learning, 37*(5–6), 1048–1073. https://doi.org/10.1080/09588221.2022.2067182
Lee, J. H., Shin, D., & Noh, W. (2023). Artificial intelligence-based content generator technology for Young English-as-a-Foreign-Language learners' reading enjoyment. *RELC Journal, 54*(2), 508–516. https://doi.org/10.1177/00336882231165060
Liang, J. C., Hwang, G. J., Chen, M. R. A., & Darmawansah, D. (2021). Roles and research foci of artificial intelligence in language education: An integrated bibliographic analysis and systematic review approach. *Interactive Learning Environments, 31*(7), 4270–4296. https://doi.org/10.1080/10494820.2021.1958348
Lin, C.-J., & Mubarok, H. (2021). Learning analytics for investigating the mind map-guided AI chatbot approach in an EFL flipped speaking classroom. *Educational Technology & Society, 24*(4), 16–35. https://www.jstor.org/stable/48629242.
Liu, N. F., & Carless, D. (2006). Peer feedback: The learning element of peer assessment. *Teaching in Higher Education, 11*(3), 279–290. https://doi.org/10.1080/13562510600680582
Luce, R. D. (1959). On the possible psychophysical laws. *Psychological Review, 66*(2), 81–95. https://doi.org/10.1037/h0043178
Luckin, R. (2017). Towards artificial intelligence-based assessment systems. *Nature Human Behaviour, 1*, 28. https://doi.org/10.1038/s41562-016-0028
Lyster, R., Saito, K., & Sato, M. (2013). Oral corrective feedback in second language classrooms. *Language Teaching, 46*(1), 1–40. https://doi.org/10.1017/S0261444812000365
Madhavi, E., Sivapurapu, L., Koppula, V., Sreehari, V., & Rani, P. (2023). Developing learners' English-speaking skills using ICT and AI tools. *Journal of Advanced Research in Applied Sciences and Engineering Technology, 32*(2), 142–153. https://doi.org/10.37934/araset.32.2.142153
Molloy, E., Boud, D., & Henderson, M. (2020). Developing a learning-centred framework for feedback literacy. *Assessment & Evaluation in Higher Education, 45*(4), 527–540. https://doi.org/10.1080/02602938.2019.1667955
Moussalli, S., & Cardoso, W. (2020). Intelligent personal assistants: Can they understand and be understood by accented L2 learners? *Computer Assisted Language Learning, 33*(8), 865–890. https://doi.org/10.1080/09588221.2019.1595664
Newhouse, C. P., & Cooper, M. (2013). Computer-based oral exams in Italian language studies. *ReCALL, 25*(3), 321–339. https://doi.org/10.1017/S0958344013000141
Ngo, T. N., & Hastie, D. (2025). Artificial intelligence for academic Purposes (AIAP): Integrating AI literacy into an EAP module. *English for Specific Purposes, 77*, 20–38. https://doi.org/10.1016/j.esp.2024.09.001
Paquot, M., Rubin, R., & Vandeweerd, N. (2022). Crowdsourced adaptive comparative judgment: A community-based solution for proficiency rating. *Language Learning, 72*(3), 853–885. https://doi.org/10.1111/lang.12498
Pavlenko, A. (2007). Autobiographic narratives as data in applied linguistics. *Applied Linguistics, 28*(2), 163–188. https://doi.org/10.1093/applin/amm008
Pokrivcakova, S. (2019). Preparing teachers for the application of AI-powered technologies in foreign language education. *Journal of Language and Cultural Education, 7*(3), 135–153. https://doi.org/10.2478/jolace-2019-0025
Pollitt, A. (2012a). Comparative judgement for assessment. *International Journal of Technology and Design Education, 22*(2), 157–170. https://doi.org/10.1007/s10798-011-9189-x
Pollitt, A. (2012b). The method of adaptive comparative judgement. *Assessment in Education: Principles, Policy & Practice, 19*(3), 281–300. https://doi.org/10.1080/0969594X.2012.665354
Qiao, H., & Zhao, A. (2023). Artificial intelligence-based language learning: Illuminating the impact on speaking skills and self-regulation in Chinese EFL context. *Frontiers in Psychology, 14*. https://doi.org/10.3389/fpsyg.2023.1255594
Rad, H. (2024). Revolutionizing L2 speaking proficiency, willingness to communicate, and perceptions through artificial intelligence: Acase of Speeko application. *Innovation in Language Learning and Teaching, 18*(4), 364–379. https://doi.org/10.1080/17501229.2024.2309539

Ranta, L., & Lyster, R. (2007). A cognitive approach to improving immersion students' oral language abilities: The Awareness–Practice–Feedback sequence. In R. DeKeyser (Ed.), *Practice in a second language: Perspectives from applied linguistics and cognitive psychology* (pp. 141–160). Cambridge University Press. https://doi.org/10.1017/CBO9780511667275.009.

Rodríguez-González, E., & Castañeda, M. E. (2016). The effects and perceptions of trained peer feedback in L2 speaking: Impact on revision and speaking quality. *Innovation in Language Learning and Teaching, 12*(2), 120–136. https://doi.org/10.1080/17501229.2015.1108978

Ruegg, R. (2015). The relative effects of peer and teacher feedback on improvement in EFL students' writing ability. *Linguistics and Education, 29*, 73–82. https://doi.org/10.1016/j.linged.2014.12.001

Schünemann, N., Spörer, N., Völlinger, V. A., & Brunstein, J. C. (2017). Peer feedback mediates the impact of self-regulation procedures on strategy use and reading comprehension in reciprocal teaching groups. *Instructional Science, 45*(4), 395–415. https://doi.org/10.1007/s11251-017-9409-1

Sharadgah, T. A., & Sa'di, R. A. (2022). A systematic review of research on the use of artificial intelligence in English language teaching and learning (2015-2021): What are the current effects? *Journal of Information Technology Education: Research, 21*, 337–377. https://doi.org/10.28945/4999

Shazly, R. (2021). Effects of artificial intelligence on English speaking anxiety and speaking performance: A case study. *Expert Systems, 38*(3). https://doi.org/10.1111/exsy.12667

Sherman, D., Mentzer, N., Bartholomew, S., Chesley, A., Baniya, S., & Laux, D. (2022). Across the disciplines: Our gained knowledge in assessing a first-year integrated experience. *International Journal of Technology and Design Education, 32*, 1369–1391. https://doi.org/10.1007/s10798-020-09650-6

Skehan, P. (2009). Modelling second language performance: Integrating complexity, accuracy, fluency, and Lexis. *Applied Linguistics, 30*(4), 510–532. https://doi.org/10.1093/applin/amp047

Storch, N. (2017). Peer corrective feedback in computer-mediated collaborative writing. In H. Nassaji, & E. Kartchava (Eds.), *Corrective feedback in second Language Teaching and learning* (pp. 66–79). Routledge. https://doi.org/10.4324/9781315621432-6.

Sun, H. (2021). *A review of research on automatic scoring of spoken English at home and abroad*. Research: Frontiers of Foreign Language Education.

Sun, H., & Zhang, M. (2020). A comparative study of machine scoring and manual scoring of spoken English. *Foreign Language Research, 37*(4), 57–62.

Sutton, P. (2012). Conceptualizing feedback literacy: Knowing, being, and acting. *Innovations in Education & Teaching International, 49*(1), 31–40. https://doi.org/10.1080/14703297.2012.647781

van de Poel, K., & Gasiorek, J. (2024). Using AI to expand the "Toolbox" for EAP writing instruction: Student experiences and perceptions of ChatGPT's instructional potential. *AILA Review*. https://doi.org/10.1075/aila.24029.van

van Popta, E., Kral, M., Camp, G., Martens, R. L., & Simons, P. R.-J. (2017). Exploring the value of peer feedback in online learning for the provider. *Educational Research Review, 20*, 24–34. https://doi.org/10.1016/j.edurev.2016.10.003

Verhavert, S., Bouwer, R., Donche, V., & De Maeyer, S. (2019). A meta-analysis on the reliability of comparative judgement. *Assessment in Education: Principles, Policy & Practice, 26*(5), 541–562. https://doi.org/10.1080/0969594X.2019.1602027

Vygotsky, L. S., & Cole, M. (1978). *Mind in society: Development of higher psychological processes*. Harvard university press.

Winstone, N. E., Nash, R. A., Parker, M., & Rowntree, J. (2017). Supporting learners' agentic engagement with feedback: A systematic review and a taxonomy of recipience processes. *Educational Psychologist, 52*(1), 17–37. https://doi.org/10.1080/00461520.2016.1207538

Wu, J. G., & Miller, L. (2020). Improving English learners' speaking through mobile-assisted peer feedback. *RELC Journal, 51*(1), 168–178. https://doi.org/10.1177/0033688219895335

Yang, H., Kim, H., Lee, J. H., & Shin, D. (2022). Implementation of an AI chatbot as an English conversation partner in EFL speaking classes. *ReCALL, 34*(3), 327–343. https://doi.org/10.1017/s0958344022000039

Yeh, H. C., Tseng, S. S., & Chen, Y. S. (2019). Using online peer feedback through blogs to promote speaking performance. *Journal of Educational Technology & Society, 22*(1), 1–14. https://www.jstor.org/stable/26558824.

Yu, S., & Lee, I. (2016). Peer feedback in second language writing (2005–2014). *Language Teaching, 49*(4), 461–493. https://doi.org/10.1017/S0261444816000161

Zhang, C., Meng, Y., & Ma, X. (2024). Artificial intelligence in EFL speaking: Impact on enjoyment, anxiety, and willingness to communicate. *System, 121*, Article 103259. https://doi.org/10.1016/j.system.2024.103259

Zou, B., Guan, X., Shao, Y., & Chen, P. (2023). Supporting speaking practice by social network-based interaction in Artificial Intelligence (AI)-assisted language learning. *Sustainability, 15*(4), 2872. https://doi.org/10.3390/su15042872

Zou, B., Liviero, S., Ma, Q., Zhang, W., Du, Y., & Xing, P. (2024). Exploring EFL learners' perceived promise and limitations of using an Artificial Intelligence speech evaluation system for speaking practice. *System, 126*, Article 103497. https://doi.org/10.1016/j.system.2024.103497

Xu (Jared) Liu is an Associate Language Lecturer at the English Language Centre of the School of Languages at Xi'an Jiaotong-Liverpool University. His research interests encompass English for Academic Purposes, technology-assisted language learning, writing and genre pedagogy, agency and academic literacy.

Jingwen Wang is a final-year Ph.D. candidate at the University of Southampton. Her research interests are diverse and include L2 language testing and assessment, with a specific focus on Adaptive Comparative Judgment, pronunciation assessment, machine scoring, automated speaking assessment, and the application of statistical methods in applied linguistics.

Dr. Bin Zou is a Senior Associate Professor in the Department of Applied Linguistics at Xi'an Jiaotong-Liverpool University, China. He received his Ph.D. degree in TESOL from the University of Bristol, UK. His research interests include English Language Teaching, English for Academic Purposes and Artificial Intelligence technology in education. He has published many papers in respected journals and contributed chapters to several books.