Language Learning & Technology
ISSN 1094-3501 CC BY-NC-ND

June 2024, Volume 28, Issue 2
pp. 210–228

ARTICLE

Automated versus peer assessment: Effects on learners' English public speaking

Chunping Zheng, Beijing University of Posts and Telecommunications
Xu Chen, Beijing University of Posts and Telecommunications
Huayang Zhang, Shenzhen Polytechnic University
Ching Sing Chai, The Chinese University of Hong Kong

Abstract

This quasi-experimental research investigates the employment of a formative assessment platform aided by artificial intelligence in an English public speaking course. The platform integrates deep learning, automatic speech recognition, and automatic writing evaluation. It provides automated assessment and immediate feedback on speakers' public speaking anxiety and their speaking and writing competence. Fifty-two English public speaking learners were randomly assigned to two groups. The control group (G1) undertook self-, peer, and teacher assessment via the platform, while the experimental group (G2) experienced self-, automated, and teacher assessment. The ANCOVA results revealed that students in G1 perceived significantly higher social engagement than those in G2, which indicates that social interaction between learners during peer assessment cannot be substituted by automated assessment. The chi-square analysis showed students' different concerns regarding online formative assessment. While students in G1 showed concerns for peers' qualifications and willingness to provide feedback, students in G2 suggested generating more detailed automated feedback to improve self-learning. No significant differences were found in learners' English public speaking self-efficacy, engagement, or competence. This indicates that automated assessment can serve as an effective strategy for formative assessment and that AI tools may supplement peers as reliable learning companions in the foreseeable future.

Keywords: *Automated Assessment, Peer Assessment, English Public Speaking, Learner Engagement, Self-efficacy*

Language(s) Learned in This Study: *English*

APA Citation: Zheng, C., Chen, X., Zhang, H., & Chai, C. S (2024). Automated versus peer assessment: Effects on learners' English public speaking. *Language Learning & Technology*, *28*(2), 210–228. https://hdl.handle.net/10125/73577

Introduction

Artificial intelligence (AI) has evolved at an unprecedented pace, and the landscape of computer-assisted language learning has undergone changes into intelligent computer-assisted language learning. A range of state-of-the-art AI technologies has been integrated into language education to improve its efficiency and effectiveness, such as automatic speech recognition (ASR), automated writing evaluation (AWE), machine translation, and chatbots (Chen et al., 2021; Lin & Chang, 2020; Yang et al., 2023). They offer learners more timely and personalized feedback, as well as interactive and authentic learning contexts for language learners (Chen et al., 2022). However, there are still considerable gaps in our understanding of the affordances and constraints of AI for language education. It remains unclear whether AI-based automated assessment tools could achieve a similar effect on facilitating learners' development of comprehensive language skills as surrogate teachers or peers. This is an area that needs to be constantly updated, and new pedagogical innovations are needed to incorporate AI affordances into language classrooms.

Formative assessment is a "socially situated interpretive act" and refers to a process of providing learners with immediate feedback and supplementary support during their learning (Boud et al., 2018, p. 1109). Its integration to language curricula enables learners to continuously adjust their learning processes and improve learning outcomes (Bulut et al., 2022). It is thus commonly utilized in English as a foreign language (EFL) classrooms (e.g., Chien et al., 2020). Formative assessment requires teachers and learners to have continual dialogues and engagement in feedback loops (Stobart, 2008). Feedback, ideally along with suggestions, constitutes an essential part of effective formative assessment (Fu et al., 2022; Gu, 2021). However, the short durations of courses, large class sizes, and limited numbers of human raters pose challenges for formative assessment in everyday classroom practice (Chen, 2022). While assessing learners' performance, language teachers may also suffer various rater effects such as drift, fatigue, or bias (Zechner & Evanini, 2019).

Online formative assessment systems can support users to conduct formative assessment, including clarification of goals, elicitation of evidence, interpretation of evidence, provision of feedback, and student/teacher take-up and action (Godwin-Jones, 2022; Gu, 2021). They overcome time and space constraints for conducting assessment activities, ensure confidentiality, and increase learners' willingness to evaluate their own and peers' work (Hoang et al., 2022). More recently, AI-programmed automated assessment tools have become an additional mode for formative assessment with automated and instant feedback (Godwin-Jones, 2022), offering a possible alternative to large-scale, classroom-based formative assessment. However, to what extent EFL learners would engage with and benefit from automated assessment supported by AI technologies remains under-investigated. It is unclear whether AI-aided automated assessment could serve as an "intelligent" and trustworthy partner for improving learners' English language skills or supporting their learning engagement.

Effective English public speaking (EPS) is a strategic communication skill and demonstrates a learner's comprehensive communicative language ability (Zhang et al., 2019b). EPS forms a nexus of English writing, speaking, and communication (Zhang et al., 2020) and requires learners to effectively manage a multimodality of language skills (Lee & Azman, 2020). In classroom teaching, facilitating self-assessment and involving peers in formative assessment are two ways to enhance learners' public speaking skills (van Ginkel et al., 2015). This research reports on the design and development of an AI-supported formative assessment system for achieving both human and automated assessment of learners' EPS. To evaluate the effect of automated assessment on learners' EPS self-efficacy, engagement, and competence, we compared two groups of students with two different interventional conditions. The control group (G1) undertook self-, peer, and teacher assessment while the experimental group (G2) experienced self-, automated, and teacher assessment. These two procedures of online formative assessment distinguished G1 and G2.

Literature Review

Peer Assessment and Automated Assessment

Peer assessment is one of the three major strategies for formative assessment, involving iterative processes of identifying peers' learning practices in relation to the learning goals and expected outcomes, triggering tailored formative feedback that supports learning (Gikandi et al., 2011). Online peer assessment has been increasingly adopted in EFL learning contexts, and its potential effect on learning outcome has been reported (Zheng et al., 2023). However, the objectiveness and positive effects of online peer assessment have also been questioned. Students were sometimes unwilling to disapprove of their peers or usually provide friendly comments to avoid face-threatening act or embarrassment (Hoang et al., 2022; Ma, 2020). They may not be fully engaged in evaluating others' work, which results in biased, unfair, or unreliable evaluation (Dai & Wu, 2023; Zechner & Evanini, 2019).

Automated assessment systems can reduce the costs and turn-around time of human assessment (Zechner & Evanini, 2019). Their potential in language learning has been acknowledged (Hoang et al., 2020; Shang, 2022; Tan et al., 2022). Automatic scoring engines are now comparable or even superior to human raters

(Shermis, 2014; Zhai & Ma, 2022). Developed to help English learners prepare for the TOEFL iBT Speaking assessment, the SpeechRater system can evaluate learners' spontaneous speaking performance automatically regarding delivery competence, fluency and pronunciation, rhythm, vocabulary, grammar, content, and discourse. Partnering with AI-supported language learning platforms allows learners to receive automated feedback even when no human support is available (Liao, 2016; Zhai & Ma, 2022). Technological tools can therefore assist teachers in overcoming the difficulties of offering real-time and continuous evaluation both inside and outside of the classroom (Calvo & Ellis, 2010; Fu et al., 2022). This may better address the learners' expectations of waiting for teachers' feedback, particularly for large-size classes (Yang et al., 2023). Although previous studies have been conducted to investigate the effects of online feedback on language learning, the nuanced effect of online peer and automated assessment on EFL learners' self-efficacy, engagement, and EPS competence have not been fully elaborated.

Learners' Self-Perceived Confidence, Engagement, and Actual Competence in EPS

Language learners' academic achievement is a complex phenomenon that requires nuanced attention to multiple factors based on diverse theoretical frameworks. Biggs (1993) revised the presage-process-product model for a refined understanding of student learning and depicted the integrated system of learning with various facets. Past studies have identified learners' self-efficacy, motivation and engagement as essential elements leading to success or causing failures in the process of learning (Acosta-Gonzaga, 2023; Dogan, 2015). Self-efficacy refers to learners' subjective beliefs in their capabilities to attain academic goals (Pajares, 1996). Compared with their less efficacious counterparts, self-efficacious students are more engaged, motivated, and self-regulated in language learning (Chen et al., 2022; Oh, 2022). Studies have shown that self-efficacy is associated with specific language skills such as reading (Shang, 2010), writing (Sun & Wang, 2020), listening (Rahimi & Abedini, 2009), and speaking (Zhang et al., 2020). Zhang et al. (2019b) defined EPS self-efficacy as learners' confidence in their ability to successfully deliver an English public speech. They developed and validated an instrument to measure learners' EPS self-efficacy and concluded that EPS self-efficacy is possibly one of the strongest contributors to the development of learners' EPS skills (Zhang et al., 2020).

Learner engagement may account for variances in the learning process. It refers to "the extent of a student's active and productive involvement in a learning activity" (Reeve et al., 2020, p. 2). Philp and Duchesne (2016) define learner engagement as "a state of heightened attention and involvement, in which participation is reflected not only in the cognitive dimension, but in social, behavioral, and affective dimensions as well" (p. 3). One well-articulated framework of language learner engagement comprises four interrelated dimensions: cognitive, behavioral, emotional, and social engagement (e.g., Fredricks et al., 2016; Philp & Duchesne, 2016). Liu and Yu (2022) claimed that feedback explicitness affects L2 learners' engagement and called for timely teacher and peer feedback to supplement automated feedback.

Learners' effective EPS is the product of English language learning. It is not merely a matter of self-perceived confidence or engagement, but also a display of complex communicative competence with physiological, sociolinguistic aspects as well as cognitive understandings (Morita, 2000; Zheng et al., 2023). Several rating scales or evaluation forms have been proposed to assess learners' EPS competence in speech composition and public speaking delivery (e.g., Lucas & Yin, 2011; Schreiber et al., 2012; Zhang et al., 2020). Specific instruments have also been designed and validated to evaluate learners' self-confidence or anxiety in public speaking (Coskun, 2017; McCroskey, 1970; Mörtberg et al., 2018; Paul, 1966). As for the automated assessment of learners' EPS competence, it mainly consists of the following three aspects: how learners manage their public speaking anxiety (Chen et al., 2015; Gregersen, 2020; Zheng et al., 2024), their speaking competence in speech delivery (Wu et al., 2009) and their writing competence in composing speech drafts (Gao, 2021; Yao, 2021). The extent to which learners' EPS competence is affected by their EPS self-efficacy (the students' sense of their own capabilities in EPS) and student engagement (students' involvement in EPS learning activities), still needs investigation, particularly when automated assessment is involved.

Online Formative Assessment for Learners' Self-efficacy and Engagement

Jao et al. (2022) used Mofunshow, a video dubbing app, as a tool to conduct formative assessment and found that learners' accuracy and fluency in English speaking improved after the intervention. Chen (2022) explored the effects of formative assessment provided by Orai, a communication coaching application, and found that AI-generated feedback alleviated learners' speaking anxiety and improved their EPS performance. Ochoa and Dominguez (2020) evaluated the effect of an automated feedback system to develop learners' oral presentation skills and suggested that automated feedback has a positive effect on improving the quality of oral presentations, although the effect size is smaller compared with human-generated feedback.

Beyond speaking, Zhai and Ma (2022) explored the impact of automated writing evaluation tools and revealed that the automated cognitive feedback positively influenced learners' computer self-efficacy. Sherafati and Mahmoudi Largani (2023) compared the effect of teacher feedback and automated feedback on EFL learners' writing self-efficacy and found that automated feedback could better improve learners' self-efficacy in writing. More importantly, automated feedback on learners' writing can improve learners' writing performances (e.g., Ranalli, 2021; Zhang & Hyland, 2018). These studies prepared the way for this study to compare the effects of peer and automated feedback in a quasi-experimental setting. The objective of the present study is to empirically investigate the effects of peer and automated formative assessment on learners' EPS self-efficacy, engagement, and competence. A mixed-methods research was designed to address three research questions:

1. What are the effects of peer and automated assessment on learners' EPS self-efficacy and engagement?
2. What are the effects of peer and automated assessment on learners' EPS competence?
3. How do learners perceive the two modes of online formative assessment for their EPS learning?

Research Method

Research Context and Participants

Research Context and the E-platform with AI Technology

The study was conducted in a compulsory course for EFL learners designed to improve their EPS skills at a comprehensive university in northern China. The course lasted for 16 weeks with a two-hour class period each week. As illustrated in Figure 1, the classrooms were equipped with high-resolution cameras. Two trained teaching assistants, acting as cameramen, recorded videos of individual learners' speeches in class. In order to help learners get used to being videotaped while delivering a speech, which may cause inherent anxiety, we informed the participants at the beginning of the research and videotaped two to three times in the classroom before their first formal speech. Therefore, learners had sufficient opportunity to become accustomed to being videotaped while speaking in public. After the recording, the students were asked to submit their videos (which were converted into audio format immediately after uploading), their supplementary speech drafts in Word file format, and their visual aids in PowerPoint files via an online formative assessment platform empowered by deep learning technologies (the E-platform for short, as shown in Figure 2).

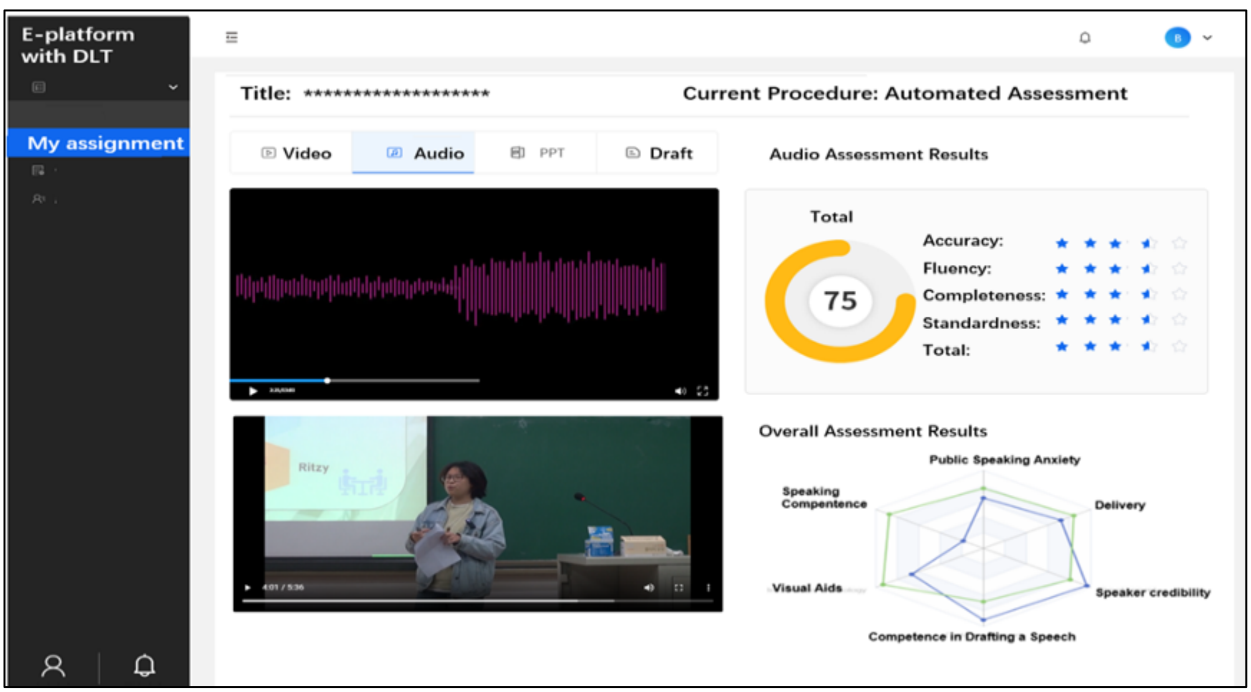
The E-platform was developed by a research team led by the first author. It can conduct automated assessment, self-, peer- and teacher assessment of learners' EPS competence based on their speech videos, audios, and speech drafts (Zheng et al., 2024). As indicated by Figure 2, a radar plot was generated based on the scores of different dimensions for human and machine assessment of learners' EPS competence, in which the green line shows the average rating of the whole class while the blue line indicates the learner's competence. Specific scores were provided referring to different sub-dimensions of learners' EPS competence.

Figure 1

Collecting Learners' Speech Videos in Classroom Settings

Figure 2

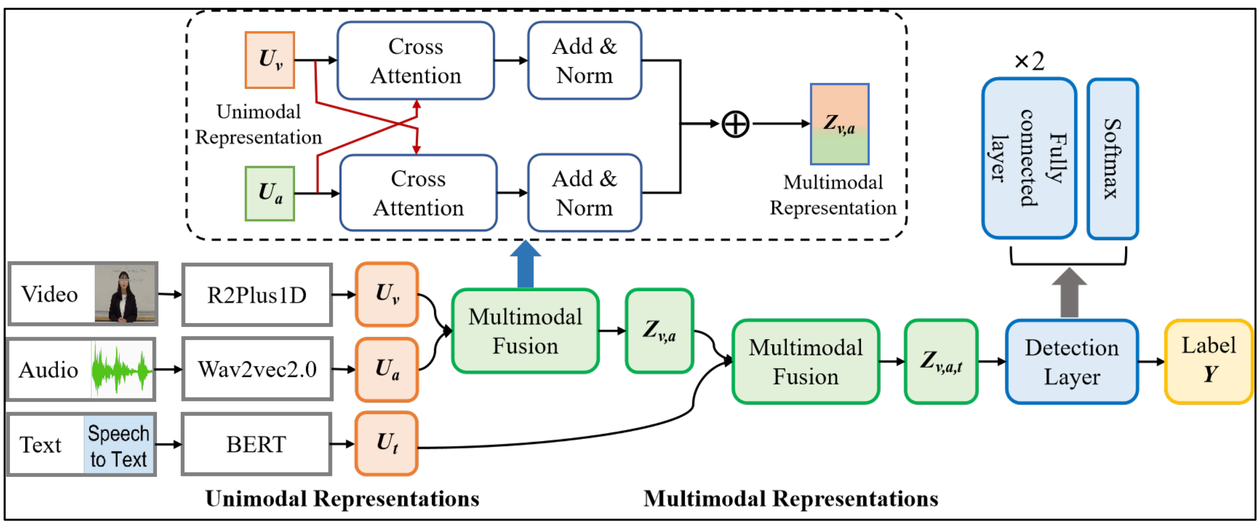
A Snapshot of the E-platform with DLT



Appendix A (available via Open Science Framework) shows the specific evaluation dimensions for the three constructs to indicate learners' EPS competence. Three different categories of AI technologies were used to achieve automated assessment and generate related feedback. First, deep learning technologies (as shown in Figure 3) were utilized to evaluate speakers' public speaking anxiety based on their facial expressions, gazing behavior, body movement, gesture and posture, and other verbal clues (as indicated in the first line of Appendix A; Chen et al., 2015; Gregersen, 2020). Second, ASR technologies from iFLYTEK (a representative AI enterprise in China, Chen et al., 2021) were employed to evaluate learners' speaking competence based on their speaking accuracy, fluency, clarity, and completeness. Third, AWE technologies from the Pigai system (one of the most explored AWE platforms within Chinese learner contexts, Barrot, 2023) were used to evaluate the quality of speakers' speech drafts according to the vocabulary, sentences, structure, and relevance. The validity and reliability of the above-mentioned technologies in the automated assessment of the three major components of EPS competence have been reported in previous studies (Bai & Hu, 2017; Gao, 2021; Wang et al., 2023; Yao, 2021; Zheng et al., 2024). The maintenance and operations of the platform were funded by the National Natural Science Foundation in China and the Big Data Research and Service Center at the first author's institution, which provides learners and teachers both on and off campus with open access to the online platform.

Figure 3

An Illustration of the Deep Learning Model for Evaluating Learners' PSA (Zheng et al., 2024)



Being slightly different from automated assessment, human assessment consists of three modes: self-, peer, and teacher assessment. Self- and peer assessment follow the same rating criteria as automated assessment, while teacher assessment takes learners' expressiveness in speaking and their idiomaticity, logic, and reasoning in writing as additional criteria for rating their speaking and writing competence. Both groups conducted self-assessment and received teacher assessment on the E-platform. In the first three weeks of classroom teaching, peers were trained to provide constructive feedback through the E-platform by illustrating how teachers offer detailed and personalized feedback on learners' EPS competence online following the subdimensions indicated in Appendix A.

Participants

The participants were recruited during the 2022-2023 academic year, including 44 female and eight male students (average age 19.25 years). They were sophomore students in the English department of a university in the northern part of China. They were selected for convenience because all the participants attended the same course and consented to participate in the research. Due to the long-existing liberal arts nature of English departments in China (Liu & Zhang, 2023), more female students apply and are recruited than male students at higher education settings in China, thus explaining the gender imbalance in the sample. The participants were then randomly assigned into two groups, with 26 students in the control group (G1), undertaking self-, peer, and teacher assessment, and another 26 students in the experimental group (G2), experiencing self-, automated, and teacher assessment.

Research Design

The present study adopted an explanatory, sequential, mixed-methods approach (Creswell, 2014) to answer the three research questions. A quasi-experiment was designed to explore the effects of two different modes of online formative assessment on learners' EPS self-efficacy, engagement, and competence. A follow-up interview was conducted, and both quantitative and qualitative data were collected and analyzed.

All the participants were asked to deliver three formal public speaking tasks according to their assigned conditions as indicated in Appendix B (available via Open Science Framework). Following the instructional design of EPS teaching (Zheng et al., 2023), the first task was a 3-minute introductory speech, the second task was a 4-to-5-minute informative speech, and the third was a 5-to-6-minute persuasive speech. As shown in Appendix B, the three formal speeches were conducted in Week 6, Week 11, and Week 16 during the instruction weeks (from Week 1 to Week 16) in the classroom. The first 3 weeks (Week 1 to Week 3) were pre-intervention practice periods, when learners were given sufficient instruction on how to conduct self- and peer-assessment, and how to get used to the video-taped teaching process. Learners' in-class public speaking was video-recorded as indicated in the research context part for online formative assessment practice after class.

As indicated in Appendix B, for the control group (G1), the learners were asked to complete the online self-, peer, and teacher assessments on their public speaking competence for their public speaking tasks. For the experimental group (G2), the students were required to complete the online self-, automated, and teacher assessments for the three public speaking tasks. The two different procedures of formative assessment constitute the independent variable in this study. The dependent variables were learners' EPS self-efficacy, engagement, and competence after receiving the two differential modes of assessment practice.

Data Collection

The quasi-experimental research was carried out over 16 successive weeks between early September and late December 2022, with two hours of teaching time each week. At the beginning of the course, informed of the research procedure and research purposes, all the participants signed consent forms and agreed to voluntarily participate in the research. We collected data from multiple sources to allow data triangulation, including students' self-reported survey responses, video-recordings of their public speaking, their speech drafts, scorings of learners' EPS competence, and their interview responses.

Measurements for EPS Self-efficacy, Engagement, and Competence

This study used two questionnaires for pre- and post-intervention surveys on learners' EPS self-efficacy and learning engagement. The first is the EPS self-efficacy scale, which measures learners' EPS self-efficacy in four dimensions with 12 items, each on a five-point Likert scale ranging from 1 = strongly disagree to 5 = strongly agree. The four dimensions are EPS learners' self-efficacy in language competence, organization competence, topic competence, and delivery competence. The structural validity of the measure was tested through exploratory and confirmatory factor analyses (Zhang et al., 2019b). The Cronbach's alpha of the instrument was 0.86 (Zhang et al., 2019b), indicating its reliability for evaluating learners' EPS self-efficacy.

The second questionnaire evaluated learners' engagement in EPS learning and was adapted from the measurements of Wang et al. (2016) and Luan et al. (2023). The survey is also on a five-point Likert scale ranging from 1 = strongly disagree to 5 = strongly agree. Twenty-one items with four dimensions, namely cognitive, behavioral, emotional, and social engagement, were retained, and the validity of the measure was also tested (Luan et al., 2023). The Cronbach's alpha values of the four constructs in this scale ranged from 0.80-0.94, while the overall alpha was 0.93, displaying the adequate internal consistency of the measurement items (Luan et al., 2023).

As for learners' EPS competence, we referred to three constructs, namely speakers' management public speaking anxiety (the lower the PSA score, the better self-management of PSA), speaking and writing competence (the higher the scores, the better). As indicated by Appendix A, based on a five-point Likert scale (1 = very anxious/poor, 2 = anxious/fair, 3 = average, 4 = calm/good, 5 = very calm/excellent), both machine assessment and human assessment were able to score learners' public speaking competence based on the three constructs. Our previous studies used the three constructs as a measurement (Zheng et al., 2024), and the reliability is acceptable, with Cronbach's alpha for the three sub-scales ranging from 0.79 to 0.87. Three teachers were invited to rate learners' public speaking competence, and their average scores were taken as learners' final scores. All teachers had PhD degrees in applied linguistics and had rich experience in teaching EPS at the university level. The teachers participated in a calibration session, in which they practice-rated multiple speech samples (six students in each group, comprising three males and three females) and discussed disagreements. Then, learners' speech videos, audios, and drafts in two groups were randomized, anonymized, and rated independently by the teachers. The inter-rater reliability of teachers' scoring on the three constructs was measured by Pearson's correlation (r ranges from 0.66 to 0.70, $p < 0.01$). Teachers' average scores of learners' first formal public speeches were used as the pre-test scores to indicate their initial EPS competence, and their scoring of learners' last formal public speeches were used as the post-test scores to indicate the group differences after the invention in EPS competence.

Interview

Semi-structured interviews were conducted among all the participants to gain a deeper understanding of their perceptions of the two different modes of online formative practices for EPS learning. The interviews allowed the researcher to deviate from the interview protocol to ask follow-up questions and explore unexpected subjects raised by the students. Each interview lasted for about 30 minutes. The interviews were audiotaped and transcribed.

Data Analysis

The quantitative data included learners' self-reported data concerning their EPS self-efficacy and engagement in the pre- and post-tests surveys and teachers' average scores of the learners' public speaking competence in the first and final formal speeches. Analysis of covariance (ANCOVA) was conducted to explore the possible effects of online formative assessment on learners' EPS self-efficacy, engagement, and competence. ANCOVA was employed to analyze the post-test scores by "using the pre-test scores as the covariate to control for pre-existing differences on the dependent variable" (Newsom, 2021, p. 1).

The qualitative data were learners' interview responses. We conducted thematic analyses on their interview data following the previous analytical frameworks (Dai & Wu, 2023; Zheng et al., 2023). The prominent themes were put into different categories and sub-categories. Then, we conducted the chi-square test based on the frequency of different categories to explain the effects of the two modes of online formative assessment on learners' public speaking self-efficacy, engagement, and competence in the two different groups. All the quantitative data were analyzed using SPSS 25.0, while the qualitative data were analyzed using NVivo 11.0.

Results

Online Peer and Automated Assessment for EPS Self-efficacy

The pre-test data on EPS self-efficacy were derived from the questionnaire collected before the first speech. For G1, the means and standard deviations of the pre-test data were 3.06 (SD = 0.72) in language competence (LC), 3.05 (SD = 0.66) in delivery competence (DC), 3.81 (SD = 0.51) in topic competence (TC), and 3.42 (SD = 0.45) in organization competence (OC). For G2, the means and standard deviations of the pre-test data were 3.10 (SD = 0.81) in LC, 3.13 (SD = 0.86) in DC, 3.44 (SD = 0.68) in TC, and 3.37 (SD =0.68) in OC. No significant differences in LC, DC, or OC (t = -0.18, *n.s.* for LC; t = -0.36, *n.s.* for

DC; $t = 0.31$, *n.s.* for OC) were found between the two conditions. There was a significant difference in TC ($t = 2.23$). Thus, students in both conditions shared roughly similar prior EPS self-efficacy when they gave their first English speeches.

ANCOVA was conducted to determine the possible effects of different conditions on EPS self-efficacy. Table 1 indicated no significant difference for the effects of different conditions on LC (F (1, 50) = 0.00, $p > 0.05$, $\eta^2 = 0.00$), DC (F(1, 50) = 2.19, $p > 0.05$, $\eta^2 = 0.04$), TC (F(1, 50) = 0.18, $p > 0.05$, $\eta^2 = 0.00$), or OC (F(1, 50) = 0.43, $p > 0.05$, $\eta^2 = 0.01$).

Table 1

ANCOVA Results of Learners' English Public Speaking Self-Efficacy (n = 52)

	G1 (control condition) (*n* = 26)				G2 (experimental condition) (*n* = 26)						
	Pretest		Posttest		Pretest		Posttest		ANCOVA		
Type	*M*	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*	*F*(1,50)	η^2	Effect size
LC	3.06	0.72	3.28	0.12	3.10	0.81	3.30	0.12	0.00	0.00	-
DC	3.05	0.66	3.18	0.10	3.13	0.86	3.39	0.10	2.19	0.04	-
TC	3.81	0.51	3.86	0.10	3.44	0.68	3.81	0.10	0.18	0.00	-
OC	3.42	0.45	3.79	0.09	3.37	0.68	3.71	0.09	0.43	0.01	-

Online Peer and Automated Assessment for Learner Engagement

The pre-test data were derived from the questionnaire collected after the first speech (as shown in Table 2). For G1, the means and standard deviations of the pre-test data were 3.97 (SD = 0.46) in cognitive engagement (CE), 4.15 (SD = 0.56) in behavioral engagement (BE), 3.46 (SD = 0.81) in emotional engagement (EE), and 4.12 (SD =0.46) in social engagement (SE). For G2, the means and standard deviations of the pre-test data were 3.87 (SD = 0.47) in CE, 4.02 (SD = 0.46) in BE, 3.36 (SD = 0.70) in EE, and 4.12 (SD = 0.43) in SE. No significant differences for CE, BE, EE, or SE ($t = 0.79$, *n.s.* for CE; $t = 0.97$, *n.s.* for BE; $t = 0.51$, *n.s.* for EE; $t = 0.00$, *n.s.* for SE) were found. Thus, students in both conditions shared similar prior EPS learning engagement when they gave their first English speeches.

ANCOVA was conducted to determine the possible effects of different conditions on EPS learning engagement. As shown in Table 2, the outcome indicates a significant difference for the effects of different conditions on SE (F (1, 50) = 4.65, $p < 0.05$, $\eta^2 = 0.09$). However, no significant differences were found in the dimensions of CE (F (1, 50) = 0.48, $p > 0.05$, $\eta^2 = 0.01$), BE (F (1, 50) = 1.01, $p > 0.05$, $\eta^2 = 0.02$), or EE (F (1, 50) = 0.70, $p > 0.05$, $\eta^2 = 0.01$).

Table 2

ANCOVA Results of Learners' English Public Speaking Learning Engagement (n = 52)

	G1 (control condition) (*n* = 26)				G2 (experimental condition) (*n* = 26)						
	Pretest		Posttest		Pretest		Posttest		ANCOVA		
Type	*M*	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*	*F*(1,50)	η^2	Effect size
CE	3.97	0.46	4.09	0.08	3.87	0.47	4.02	0.08	0.48	0.01	-
BE	4.15	0.56	4.11	0.08	4.02	0.46	4.00	0.08	1.01	0.02	-
EE	3.46	0.81	3.78	0.10	3.36	0.70	3.66	0.10	0.70	0.01	-
SE	4.12	0.46	4.29	0.08	4.12	0.43	4.05	0.08	4.65*	0.09	Medium

Online Peer and Automated Assessment for EPS Competence

The pre-test data were derived from the teachers' assessments after the first speech (as shown in Table 3). For G1, the means and standard deviations of the pre-test data were 1.80 (SD = 0.30) in public speaking anxiety (PSA), 4.07 (SD = 0.27) in speaking competence (SC), and 4.57 (SD = 0.29) in writing competence (WC). For G2, the means and standard deviations of the pre-test data were 1.81 (SD = 0.44) in public speaking anxiety (PSA), 4.05 (SD = 0.32) in SC, 4.54 (SD = 0.30) in WC, 3.98 (SD = 0.39) in SD, and 4.17 (SD =0.26) in SC. No significant differences (t = 0.12, *n.s.* for PSA; t = 0.33, *n.s.* for SC; t = 0.34, *n.s.* for WC) were found. Thus, students in both conditions shared similar prior EPS competence when they gave their first speech. ANCOVA was conducted to determine the possible effects of the different conditions on EPS competence. Table 3 indicated no significant difference for the effects of different conditions on SC ($F(1, 50) = 0.36$, $p > 0.05$, $\eta^2 = 0.01$), WC ($F(1, 50) = 1.01$, $p > 0.05$, $\eta^2 = 0.02$), SD ($F(1, 50) = 1.64$, $p > 0.05$, $\eta^2 = 0.03$), or SC ($F(1, 50) = 0.04$, $p > 0.05$, $\eta^2 = 0.00$).

Table 3

ANCOVA Results of Learners' English Public Speaking Competence (n = 52)

	G1 (control condition) (n = 26)				G2 (experimental condition) (n = 26)						
	Pretest		Posttest		Pretest		Posttest		ANCOVA		
Type	M	SD	M	SD	M	SD	M	SD	F(1,50)	η^2	Effect size
PSA	1.80	0.30	1.56	0.39	1.81	0.44	1.68	0.29	1.77	0.03	-
SC	4.07	0.27	4.11	0.05	4.05	0.32	4.06	0.05	0.36	0.01	-
WC	4.57	0.29	4.35	0.05	4.54	0.30	4.28	0.05	1.01	0.02	-

Learners' Perceptions of Peer and Automated Assessment in EPS Learning

Our study analyzed students' attitudes and perceptions toward the two modes of formative assessment based on learners' responses to semi-structured interviews. Qualitative data analyses were applied to clarify our quantitative findings. Based on the analytical framework proposed by previous studies (Dai & Wu, 2023 Zheng et al., 2023), the present research identified two major categories of students' perceptions toward different modes of formative assessment: positive perceptions (including educational, social, psychological, and technical affordances) and negative perceptions (including educational, social, and technical concerns). A chi-square test was conducted to probe into the differences between students' perceptions of the peer and automated assessments.

As shown in Table 4, there were no statistically significant differences in learners' positive perceptions for the two modes of formative assessment (X^2 =8.18, df = 7, p = .317 > .05). However, statistically significant differences in learners' negative perceptions were found for the two different modes of assessment (X^2 =17.44, df = 6, p = .008 < .05).

Table 4

Emerged Categories and Sub-Categories of Students' Perceptions of Peer and Automated Assessment

		Frequency	
Category	**Subcategory**	**G1**	**G2**
educational affordances	providing effective and targeted scaffolding	4 (6.8, -0.5)	5 (9.1, 0.5)
	facilitating a comprehensive self-understanding	10 (16.9, 0.4)	8 (14.5, -0.4)
	offering opportunities for self-reflection	6 (10.2, 0.2)	5 (9.1, -0.2)
	developing speech and writing skills	9 (15.3, -0.7)	11 (20, 0.7)

	enhancing speech delivery skills	13 (22.0, -0.9)	16 (29.1, 0.9)
social and psychological affordances	promoting interaction	3 (5.1, 1.7)	0 (0, -1.7)
	increasing self-efficacy in EPS	13 (22.0, -0.9)	6 (10.9, -1.6)
technical affordances	improving convenience and objectiveness	1 (1.7, -1.5)	4 (7.3, 1.5)
positive perceptions		**Chi-square test X^2 =8.18, *df* = 7, *p* = .317 > .05**	
educational concerns	peers' qualification for assessment	2 (7.7, 1.5)	0 (0, -1.5)
	time-consuming for the practice	2 (7.7, 1.5)	1 (3.6, -0.7)
	subjectiveness of assessment	5 (19.2, 2.4)	0 (0, -2.4)
social concerns	unwillingness to assess	1 (3.8, 1)	0 (0, -1.0)
technical concerns	technical breakdowns	4 (15.4, 0.1)	4 (14.3, -0.1)
	robustness of automated assessment	0 (0, -3.2)	9 (32.1, 3.2)
	deficiency of the assessment platform	12 (46.2, -0.3)	14 (50, 0.3)
negative perceptions		**Chi-square test X^2 =17.44, *df* = 6, *p* = .008 < .05**	

Note. Numbers in parentheses indicate column percentages and adjusted standardized residuals.

Learners' Positive Perceptions of Peer and Automated Assessment Online

As shown in Table 4, participants from both groups highlighted the educational affordances of formative assessment practice for providing effective scaffolding for their EPS learning and improved their self-reflection in speech delivery and speaking and writing skills. One student in G1 said, "peer feedback helped me realize the errors in my speech draft which I might not be aware of previously." After reviewing peer feedback, one student in G1 said that "I [tried] to organize my speech structure more appropriately and chose rhetoric to make my speech more impressive and interesting." A student in G2 stated that automated assessment helped him develop their writing structure, saying, "the automated feedback on the coherence and cohesion makes my speech more logical and convincing." Students from both groups reported that feedback on the platform enhanced their speech delivery skills, such as body language, eye contact, and articulation in expressions. Students in G1 believed that assessment from peers helped them "avoid unnecessary body movements." One student in G2 said that he needed to overcome his "fast speech rate, monotonous intonation, and stiff facial expression" to achieve a high score in automated assessments.

As for social and psychological affordances, students from G1 perceived peer assessment very useful for enhancing their interaction and rapport with others. Students in both groups reiterated that formative assessment practice improved their self-efficacy in delivering EPS. Students also acknowledged the convenience of using the platform. Four students in G2 highlighted the objectiveness of automated assessment: "Self-assessment may be subjective, but teacher and automated assessment are more objective. The AI-generated comments and suggestions are more constructive and tailored to my own performance."

Learners' Concerns for Peer and Automated Assessment Online

Our chi-square test showed a significant difference in learners' concerns with the two modes of formative assessment. Students in G1 doubted the reliability of their peers' scoring and feedback, complaining that peer assessment is too "subjective" and "inefficient (no immediate feedback)." In G2, students showed their concerns about automated assessment, criticizing its robustness: "For some tasks, automated scoring is rather rigorous and the scores are relatively low. I was wondering how the automated scores were generated on our speech tasks. Moreover, automated feedback was roughly the same for similar speech tasks." One student in G1 showed his disappointment toward peer feedback: "Sometimes I commented my peers' works carefully but my peers were not serious about providing constructive comments or feedback, which frustrated me." They also recommended that the process of peer feedback be anonymous because their fellow students "only posted positive comments [to] avoid face-threatening act." Technical issues existed for both groups and students in both groups offered constructive suggestions for improving the effectiveness of the E-platform, including adding more personalized automated feedback with AI technologies.

Discussion

Comparison between Automated Assessment and Peer Assessment

Our first finding shows no significant difference between the two assessment modes in terms of their effects on learners' overall EPS self-efficacy, engagement, and competence. Although prior research emphasizes the distinctive features of human or machine feedback independently (e.g., Shang, 2022), the current research indicates that automated assessment can achieve similar effects as peer assessment on EFL learners' EPS learning. The research provides empirical evidence of comparable effects of AI technologies in language testing and assessment. Given the importance of feedback in language learning, this study contributes to educators' confidence in employing artificial intelligence to facilitate learners' language development (Hoang et al., 2020; Shang, 2022; Tan et al., 2022). More importantly, the findings indicate that automated and peer assessment could play a complimentary role in language education.

Automated assessment possesses the affordances of timeliness, efficiency, and objectiveness (Zhai & Ma, 2022; Zhang et al., 2019a). This study indicates that the employment of it does not comprise students' efficacy, engagement, or competence in replacement of peer assessment. Hence, in context where peer assessment is not feasible because of sociocultural or physical conditions, teachers can employ automated assessment. It could be an affordable and feasible alternative solution (Godwin-Jones, 2022; Gu, 2021). However, automated assessment has been noted to be limited in providing learners with feedback on their higher-order thinking (e.g., Lee & Choi, 2017). Its accuracy in evaluating more complicated constructs, such as logical strength and the quality of specific contents in learners' public speaking, still needs improvement (e.g., Zhang et al., 2019a). This study expands our understanding of the effects of automated and peer assessment on learners' EPS self-efficacy, engagement, and competence. It demonstrates that automated assessment may serve as an alternative to peer assessment once AI technologies are mature enough to ensure accuracy in automated scoring and provide personalized feedback on learner performance. As suggested by previous scholars, different types of automated feedback on "learners' learning progress (e.g., learning improvement, learning behaviors and affective reaction)" should be generated, being more specific and dynamic rather than being "generic and formulaic" (Fu et al., 2022, p. 1).

The Role of Peer Assessment in Promoting Social Engagement

Our second finding shows that students with peer assessment reported higher social engagement than those with automated assessment. Learners' experience with peer assessment helped them to feel more socially engaged in EPS learning and to adopt more positive social interactions with peers. The AI-supported platform was able to provide learners with immediate feedback and create a stress-free and less anxiety-provoking learning environment (Wang et al., 2023). However, machine feedback is usually analogous to human feedback and lacks positive social interaction (Wang et al., 2022). Language learning is a social practice and requires learners to communicate with a real audience such as peers or teachers rather than merely with machines or computers, and the effectiveness of learners' development of linguistic skills is more related to teacher or peer-led interaction rather than machine-led interaction (Fu et al.; 2022; Lai, 2010). Therefore, the role of peers in promoting learners' social engagement in language learning is indispensable. While well-learned machines can act as surrogate interlocuters, students still regard peers as a real audience who may communicate and make more specific comments on their performance. Therefore, as reported by learners with peer assessment, we witnessed their enhanced engagement with peer feedback by revising their speeches and making them more impressive and interesting to the audience.

Learners' Perceptions of Peer and Automated Assessment in EPS learning

The present study identified similar affordances and concerns regarding peer and automated assessment for learners' EPS learning. The results are consistent with those of Weaver and Esposto (2012), showing that online formative assessment promotes learners' speaking performance and facilitates their self-reflection.

Our qualitative results echo previous research findings that human and automated feedback positively affect learners' self-efficacy (Zhai & Ma, 2022). In addition, automated assessments were regarded as timely, effective, and objective tools by our participants, like comments in previous research (e.g., Ochoa & Dominguez, 2020).

Our research also reveals learners' concerns for the quality of peer and automated feedback of the AI-supported assessment platform. Learners showed their critical and skeptical attitudes toward peer and automated assessment through the intervention. They wondered about the uncertainty and robustness of automated assessment results and in some cases questioned the reliability of their peers' evaluations, which has also appeared in previous studies (Shi et al., 2022). The findings provide useful implications for the development of online formative assessment systems. The systems could provide more meaningful and personalized feedback by summarizing human raters' comments and suggestions and aligning those comments with learners' specific EPS performance and competence. Reward modules could be embedded in the E-platform, like offering online badges for those dedicated peers in peer scoring and feedback, to improve the participation of peer assessment. We may also incorporate Large Language Model techniques to improve the naturalness of the feedback. On the other hand, specific training or scaffolding could be provided to peer students to improve the quality of their feedback.

Conclusion

This paper reports our quasi-experimental and pilot study on an EPS formative assessment platform empowered by AI technologies. Our quantitative data show that there were generally no significant differences in learners' EPS self-efficacy, engagement, or competence among students experiencing different procedures of online formative assessment practice. This indicates that automated assessment can serve as an effective strategy for formative assessment, and that AI tools may serve as a reliable learning companion as peer students to provide commentary feedback in the foreseeable future. However, our nuanced findings show that students involved in peer assessment perceive higher levels of social engagement. This suggests that the current AI tools may best serve to supplement our existing strategies to enhance learners' EPS competence. It also demonstrates students' perceptions of the indispensable role of peer assessment for engaging learners in social interactions in language learning.

Peers' qualifications and willingness to provide feedback, and the quality of automated feedback are two major concerns for AI-supported online formative assessment. While automated assessment can provide immediate feedback, peers' feedback is perceived as being more comprehensive and nuanced. We therefore suggest incorporating automated and peer feedback by leveraging the strengths of different feedback sources. Researchers, course teachers and technological experts should collaborate to facilitate coordination between automated and peer feedback. Teachers could provide more guidance and support for the peer students to improve the quality of their feedback. With learning analytics, the AI program can synthesize comments given by human raters and customize the feedback according to learners' varying performance. Machines can therefore learn from human feedback and deliver more natural and personalized recommendations for learners to improve their EPS competence.

The research is limited in its generalizability to other contexts due to the uneven distribution of male and female students as our participants. The research team is collaborating with multiple institutions to improve the AI-supported assessment of learners' EPS competence in China, and we are positive that upcoming research would include more representative samples with an evenly distributed ratio of male and female students. Our findings revealed learners' varied engagement with machine and peer feedback. Future studies may explore the content of peer and automated feedback to look for categories of comments that significantly enhance students' engagement and development in EPS competence. Such efforts may help create reciprocal improvements in the quality of both machine and human feedback.

Acknowledgements

This research is funded by the National Natural Science Foundation of China (62107005), and the BUPT Research Project for Postgraduate Education (2024Y015). We would like to extend our special thanks to Professor Chin-Chung Tsai and Professor Jyh-Chong Liang for their insightful comments and anonymous reviewers for their professional suggestions.

References

Acosta-Gonzaga, E. (2023). The effects of self-esteem and academic engagement on university students' performance. *Behavioral Sciences*, *13*(4), 348. https://doi.org/10.3390/bs13040348

Bai, L., & Hu, G. (2017). In the face of fallible AWE feedback: How do students respond? *Educational Psychology*, *37*(1), 67–81. https://doi.org/10.1080/01443410.2016.1223275

Barrot, J. S. (2023). Using automated written corrective feedback in the writing classrooms: Effects on L2 writing accuracy. *Computer Assisted Language Learning*, *36*(4), 584–607. https://doi.org/10.1080/09588221.2021.1936071

Biggs, J. (1993). What do inventories of students' learning processes really measure? A theoretical review and clarification. *British Journal of Educational Psychology*, *63*(1), 3–19. https://doi.org/10.1111/j.2044-8279.1993.tb01038.x

Boud, D., Dawson, P., Bearman, M., Bennett, S., Joughin, G., & Molloy, E. (2018). Reframing assessment research: Through a practice perspective. *Studies in Higher Education*, *43*(7), 1107–1118. https://doi.org/10.1080/03075079.2016.1202913

Bulut, O., Gorgun, G., Yildirim-Erbasli, S. N., Wongvorachan, T., Daniels, L. M., Gao, Y., Lai, K. W., & Shin, J. (2022). Standing on the shoulders of giants: Online formative assessments as the foundation for predictive learning analytics models. *British Journal of Educational Technology*, *52*, 19–39. https://doi.org/10.1111/bjet.13276

Calvo, R. A., & Ellis, R. A. (2010). Students' conceptions of tutor and automated feedback in professional writing. *Journal of Engineering Education*, *99*(4), 427–438. https://doi.org/10.1002/j.2168-9830.2010.tb01072.x

Chen, N. T. M., Thomas, L. M., Clarke, P. J. F., Hickie, I. B., & Guastella, A. J. (2015). Hyperscanning and avoidance in social anxiety disorder: The visual scanpath during public speaking. *Psychiatry research*, *225*(3), 667–672. https://doi.org/10.1016/j.psychres.2014.11.025

Chen, X., Lake, J., & Padilla, A. M. (2021). Grit and motivation for learning English among Japanese university students. *System*, *96*, 102411. https://doi.org/10.1016/j.system.2020.102411

Chen, X., Meurers, D., & Rebuschat, P. (2022). ICALL offering individually adaptive input: Effects of complex input on L2 development. *Language Learning & Technology*, *26*(1), 1–21. https://hdl.handle.net/10125/73496

Chen, Y. (2022). Effects of technology-enhanced language learning on reducing EFL learners' public speaking anxiety. *Computer Assisted Language Learning,* 1–25. https://doi.org/10.1080/09588221.2022.2055083

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*, 103751. https://doi.org/10.1016/j.compedu.2019.103751

Coskun, A. (2017). The effect of pecha kucha presentations on students' English public speaking anxiety. *Profile: Issues in Teachers' Professional Development*, *19*, 11–22. https://doi.org/10.15446/profile.v19n_sup1.68495

Creswell, J. W. (2014). *Research design: Qualitative, quantitative, and mixed methods approaches* (4th eds.). SAGE.

Dai, Y., & Wu, Z. (2023). Mobile-assisted pronunciation learning with feedback from peers and/or automatic speech recognition: A mixed-methods study. *Computer Assisted Language Learning, 36*(5-6), 861–884. https://doi.org/10.1080/09588221.2021.1952272

Dogan, U. (2015). Student engagement, academic self-efficacy, and academic motivation as predictors of academic performance. *Anthropologist, 20*(3), 553–561. https://doi.org/10.1080/09720073.2015.11891759

Fredricks, J. A., Filsecker, M., & Lawson, M. A. (2016). Student engagement, context, and adjustment: Addressing definitional, measurement, and methodological issues. *Learning and Instruction*, *43*, 1–4. https://doi.org/10.1016/j.learninstruc.2016.02.002

Fu, Q. K., Zou, D., Xie, H., & Cheng, G. (2022). A review of AWE feedback: Types, learning outcomes, and implications. *Computer Assisted Language Learning*, 3*7*(1-2), 179–221. https://doi.org/10.1080/09588221.2022.2033787

Gao, J. (2021). Exploring the feedback quality of an automated writing evaluation system Pigai. *International Journal of Emerging Technologies in Learning, 16*(11), 322–330. https://www.learntechlib.org/p/219987/

Gikandi, J. W., Morrow, D., & Davis, N. E. (2011). Online formative assessment in higher education: A review of the literature. *Computers & Education*, *57*(4), 2333–2351. https://doi.org/10.1016/j.compedu.2011.06.004

Godwin-Jones, R. (2022). Partnering with AI: Intelligent writing assistance and instructed language learning. *Language Learning & Technology*, *26*(2), 5–24. http://doi.org/10125/73474

Gregersen, T. (2020). Dynamic properties of language anxiety. *Studies in Second Language Learning and Teaching, 10*(1), 67–87. http://dx.doi.org/10.14746/ssllt.2020.10.1.4

Gu, P. (2021). *Classroom-based formative assessment.* Foreign Language Teaching and Research Press.

Hoang, L. P., Le, H. T., Van Tran, H., Phan, T. C., Vo, D. M., Le, P. A., Nguyen, D. T., Pong-inwong, C. (2022). Does evaluating peer assessment accuracy and taking it into account in calculating assessor's final score enhance online peer assessment quality? *Education and Information Technologies*, *27*(3), 4007–4035. https://doi.org/10.1007/s10639-021-10763-1

Hoang, L. P., Phuong, L. A., Arch-int, S., & Arch-Int, N. (2020). Multidimensional assessment of open-ended questions for enhancing the quality of peer assessment in e-learning environments. In P. Hershey (Eds.), *Learning and Performance Assessment: Concepts, Methodologies, Tools, and Applications* (pp. 147–173). IGI Global. https://doi.org/10.4018/978-1-7998-0420-8.ch008

Jao, C. Y., Yeh, H. C., Huang, W. R., & Chen, N. S. (2022). Using video dubbing to foster college students' English-speaking ability. *Computer Assisted Language Learning*. 1–23. https://doi.org/10.1080/09588221.2022.2049824

Lai, Y.-H. (2010). Which do students prefer to evaluate their essays: Peers or computer program. *British Journal of Educational Technology*, *41*(3), 432–454. https://doi.org/10.1111/j.1467-8535.2009.00959.x

Lee, J., & Choi, H. (2017). What affects learner's higher-order thinking in technology-enhanced learning environments? The effects of learner factors. *Computers & Education*, *115*, 143–152. https://doi.org/https://doi.org/10.1016/j.compedu.2017.06.015

Lee, S., & Azman, H. (2020). Introducing a responsive multimodal oral presentation pedagogy: Integrating ted videos with web 2.0, collaborative learning and teacher feedback. *RELC Journal*, *53*(3), 703–711. https://doi.org/10.1177/0033688220945426

Liao, H. C. (2016). Using automated writing evaluation to reduce grammar errors in writing. *ELT Journal*, *70*(3), 308–319. https://doi.org/10.1093/elt/ccv058

Lin, M. P. C., & Chang, D. (2020). Enhancing post-secondary writers' writing skills with a chatbot. *Journal of Educational Technology & Society*, *23*(1), 78–92. https://www.jstor.org/stable/26915408

Liu, S., & Yu, G. (2022). L2 learners' engagement with automated feedback: An eye-tracking study. *Language Learning & Technology*, *26*(2), 78–105. https://doi.org/10125/73480

Liu, Y., & Zhang, X. (2023). Understanding academic transition and self-regulation: A case study of English majors in China. *Humanities and Social Sciences Communications*, *10*(1), 98. https://doi.org/10.1057/s41599-023-01596-z

Luan, L., Hong, J. C., Cao, M., Dong, Y., & Hou, X. (2023). Exploring the role of online EFL learners' perceived social support in their learning engagement: A structural equation model. *Interactive Learning Environments, 31*(3), 1703–1714. https://doi.org/10.1080/10494820.2020.1855211

Lucas, S. E., & Yin, S. (2011). *The art of public speaking (teacher's book)* (10th ed.). Foreign Language Teaching and Research Press.

Ma, Q. (2020). Examining the role of inter-group peer online feedback on wiki writing in an EAP context. *Computer Assisted Language Learning*, *33*(3), 197–216. https://doi.org/10.1080/09588221.2018.1556703

McCroskey, J. C. (1970). Measures of communication-bound anxiety. *Speech Monographs*, *37*(4), 269–288. https://doi.org/10.1080/03637757009375677

Morita, N. (2000). Discourse socialization through oral classroom activities in a TESL graduate program. *TESOL Quarterly*, *34*(2). 279–310. https://doi.org/10.2307/3587953

Mörtberg, E., Jansson-Fröjmark, M., Pettersson, A., & Hennlid-Oredsson, T. (2018). Psychometric properties of the personal report of public speaking anxiety (PRPSA) in a sample of university students in Sweden. *International Journal of Cognitive Therapy*, *11*, 421–433. https://doi.org/10.1007/s41811-018-0022-0

Newsom, J. T. (2021). *Analysis of covariance (ANCOVA)* [Unpublished course materials]. Portland State University. Retrieved from http://web.pdx.edu/~newsomj/mvclass/ho_ancova.pdf

Ochoa, X., & Dominguez, F. (2020). Controlled evaluation of a multimodal system to improve oral presentation skills in a real learning setting. *British Journal of Educational Technology*, *51*(5), 1615–1630. https://doi.org/10.1111/bjet.12987

Oh, Y, K. (2022). The growth trajectories of L2 self-efficacy and its effects on L2 learning: Using a curve-of-factors model. *Applied Linguistics*, *43*(1), 147–167. https://doi.org/10.1093/applin/amab005

Pajares, F. (1996). Self-efficacy beliefs in academic settings. *Review of Educational Research*, *66*(4), 543–578. https://doi.org/10.3102/00346543066004543

Paul, G. (1966). *Insight vs. desensitization in psychotherapy*. Stanford University Press.

Philp, J., & Duchesne, S. (2016). Exploring engagement in tasks in the language classroom. *Annual Review of Applied Linguistics, 36*, 50–72. https://doi.org/10.1017/S0267190515000094

Rahimi, A., & Abedini, A. (2009). The interface between EFL learners' self-efficacy concerning listening comprehension and listening proficiency. *Novitas Royal Research on Youth and Language, 3*(1), 14–28. http://www.novitasroyal.org/Vol_3_1/rahimi_abedini.pdf

Ranalli, J. (2021). L2 student engagement with automated feedback on writing: Potential for learning and issues of trust. *Journal of Second Language Writing, 52*, 1–16. https://doi.org/10.1016/j.jslw.2021.100816

Reeve, J., Cheon, S. H., & Jang, H. (2020). How and why students make academic progress: Reconceptualizing the student engagement construct to increase its explanatory power. *Contemporary Educational Psychology*, *62*, 1–12. https://doi.org/10.1016/j.cedpsych.2020.101899

Schreiber, L. M., Paul, G. D., & Shibley, L. R. (2012). The development and test of the public speaking competence rubric. *Communication Education*, *61*(3), 205–233. https://doi.org/10.1080/03634523.2012.670709

Shang, H. F. (2010). Reading strategy use, self-efficacy and EFL reading comprehension. *The Asian EFL Journal Quarterly*, *12*(2), 18–42.

Shang, H. F. (2022). Exploring online peer feedback and automated corrective feedback on EFL writing performance. *Interactive Learning Environments*, *30*(1), 4–16. https://doi.org/10.1080/10494820.2019.1629601

Sherafati, N., & Mahmoudi Largani, F. (2023). The potentiality of computer-based feedback in fostering EFL learners' writing performance, self-regulation ability, and self-efficacy beliefs. *Journal of Computers in Education, 10*(1), 27–55. https://doi.org/10.1007/s40692-022-00221-3

Shermis, M. D. (2014). State-of-the-art automated essay scoring: Competition, results, and future directions from a United States demonstration. *Assessing Writing*, *20*, 53–76. https://doi.org/10.1016/j.asw.2013.04.001

Shi, Z., Liu, F., Lai, C., & Jin, T. (2022). Enhancing the use of evidence in argumentative writing through collaborative processing of content-based automated writing evaluation feedback. *Language Learning & Technology*, *26*(2), 106–128. https://doi.org/10125/73481

Stobart, G. (2008). *Testing times: The uses and abuses of assessment*. Routledge.

Sun, T., & Wang, C. (2020). College students' writing self-efficacy and writing self-regulated learning strategies in learning English as a foreign language. *System*, *90*, 1–17. https://doi.org/10.1016/j.system.2020.102221

Tan, S., Cho, Y. W., & Xu, W. (2022). Exploring the effects of automated written corrective feedback, computer-mediated peer feedback and their combination mode on EFL learner's writing performance. *Interactive Learning Environments*, *31*(10), 7276–7286. https://doi.org/10.1080/10494820.2022.2066137

van Ginkel, S., Gulikers, J., Biemans, H., & Mulder, M. (2015). Towards a set of design principles for developing oral presentation competence: A synthesis of research in higher education. *Educational Research Review*, *14*, 62–80. https://doi.org/10.1016/j.edurev.2015.02.002

Wang, M. T., Fredricks, J. A., Ye, F., Hofkens, T. L., & Linn, J. S. (2016). The math and science engagement scales: Scale development, validation, and psychometric properties. *Learning and Instruction*, *43*, 16–26. https://doi.org/10.1016/j.learninstruc.2016.01.008

Wang, X., Liu, Q., Pang, H., Tan, S. C., Lei, J., Wallace, M. P., & Li, L. (2023). What matters in AI-supported learning: A study of human-AI interactions in language learning using cluster analysis and epistemic network analysis. *Computers & Education*, *194*, 1–17. https://doi.org/10.1016/j.compedu.2022.104703

Wang, X., Pang, H., Wallace, M. P., Wang, Q., & Chen, W. (2022). Learners' perceived AI presences in AI-supported language learning: A study of AI as a humanized agent from community of inquiry. *Computer Assisted Language Learning*, 1–27. https://doi.org/10.1080/09588221.2022.2056203

Weaver, D., & Esposto, A. (2012). Peer assessment as a method of improving student engagement. *Assessment & Evaluation in Higher Education*, *37*(7), 805–816. https://doi.org/10.1080/02602938.2011.576309

Wu, G., Yanhua, L., Yijie, L., Lei, P., Eryu, W., & Lirong, D. (2009). *iFLY system for the NIST 2008 speaker recognition evaluation* [Paper presentation]. 2009 IEEE International Conference on Acoustics, Speech and Signal Processing, Taipei. https://doi.org/10.1109/ICASSP.2009.4960557

Yang, H., Gao, C., & Shen, H. z. (2023). Learner interaction with, and response to, AI-programmed automated writing evaluation feedback in EFL writing: An exploratory study. *Education and Information Technologies, 29,* 3837–3858. https://doi.org/10.1007/s10639-023-11991-3

Yao, D. (2021). Automated writing evaluation for ESL learners: A case study of Pigai system. *Journal of Asia TEFL*, *18*(3), 949–958. https://doi.org/10.18823/asiatefl.2021.18.3.14.949

Zechner, K., & Evanini, K. (2019). *Automated speaking assessment: Using language technologies to score spontaneous speech*. Routledge.

Zhai, N., & Ma, X. (2022). Automated writing evaluation (AWE) feedback: A systematic investigation of college students' acceptance. *Computer Assisted Language Learning*, *35*(9), 2817–2842. https://doi.org/10.1080/09588221.2021.1897019

Zhang, M., Yao, L., Haberman, S. J., & Dorans, N. J. (2019a). Assessing scoring accuracy and assessment accuracy for spoken responses: Using human and machine scores. In K. Zechner, & K. Evanini (Eds.), *Automated speaking assessment* (pp. 32–58). Routledge.

Zhang, X., Ardasheva, Y., & Austin, B. W. (2020). Self-efficacy and English public speaking performance: A mixed method approach. *English for Specific Purposes*, *59*, 1–16. https://doi.org/10.1016/j.esp.2020.02.001

Zhang, X., Ardasheva, Y., Egbert, J., & Ullrich-French, S. C. (2019b). Building assessments for self-efficacy in English public speaking self-efficacy in China. *The Asia-Pacific Education Researcher, 28*(5), 411–420. https://doi.org/10.1007/s40299-019-00441-9

Zhang, Z. V., & Hyland, K. (2018). Student engagement with teacher and automated feedback on L2 writing. *Assessing Writing*, *36*, 90–102. https://doi.org/10.1016/j.asw.2018.02.004

Zheng, C., Wang, L., & Chai, C. S. (2023). Self-assessment first or peer-assessment first: Effects of video-based formative practice on learners' English public speaking anxiety and performance. *Computer Assisted Language Learning*, *36*(4), 806–839. https://doi.org/10.1080/09588221.2021.1946562

Zheng, C., Zhang, T., Chen, X., Zhang, H., Wan, J., & Wu, B. (2024). Assessing learners' English public speaking anxiety with multimodal deep learning technologies. *Computer Assisted Language Learning,* 1–29. https://doi.org/10.1080/09588221.2024.2351129

About the Authors

Chunping Zheng is currently a professor and the Deputy Dean in School of Humanities at Beijing University of Posts and Telecommunications. She also serves as the Director of the Center for Research on Technology-Enhanced Language Education. Her research interests are computer-assisted language learning and computer-assisted translation. Chunping Zheng is the corresponding author.

E-mail: zhengchunping@bupt.edu.cn

ORCiD: https://orcid.org/0000-0002-0407-2381

Xu Chen is an MA graduate in School of Humanities at Beijing University of Posts and Telecommunications. His research interests include computer-assisted language learning and applied linguistics.

E-mail: 772094107@qq.con

Huayang Zhang is currently a lecturer in School of Communication at Shenzhen Polytechnic University. His research interests are multimodal learning analytics and Artificial Intelligence in education.

E-mail: zhanghuay@szpu.edu.cn

Ching Sing Chai is currently a professor in the department of Curriculum and Instruction at The Chinese University of Hong Kong. His research interests include teacher education, design thinking and technological pedagogical content knowledge.

E-mail: CSChai@cuhk.edu.hk