Humanities & Social Sciences Communications

ARTICLE

https://doi.org/10.1057/s41599-025-05674-2 OPEN

Cultivating proficient and efficacious L2 English speakers via VoiceThread-mediated self- and peer assessments

Min-Hsun Liao[1]✉

Despite growing interest in technology-mediated language assessment, limited research has examined the differential effects of online self- versus peer assessment on L2 speaking development. This mixed-methods study investigated how VoiceThread-mediated self-assessment (SA) and peer assessment (PA) influenced the English speaking proficiency and self-efficacy of adult EFL learners. Drawing on sociocultural theory and self-regulated learning frameworks, this research explored the impact of different online assessment approaches on language learning outcomes. Thirty-seven English majors at a Taiwanese university were randomly assigned to SA ($n = 19$) and PA ($n = 18$) conditions. Over one academic year, SA participants provided narrative self-evaluations of their recorded speeches on VoiceThread, while PA participants evaluated peers' recordings and narrated their comments. Data were collected through (a) pre-post oral proficiency evaluations using adapted TOEFL speaking rubrics (interrater reliability = 0.89), (b) the English-speaking Self-efficacy Questionnaire ($\alpha = 0.92$), and (c) post-intervention interviews and surveys. ANCOVA results revealed that the SA group demonstrated significantly greater gains in overall oral proficiency ($F(1, 34) = 15.603$, $p < 0.001$, $\eta^2 = 0.32$) and self-efficacy ($F(1, 34) = 5.07$, $p < 0.05$, $\eta^2 = 0.13$). The SA group showed significantly higher performance in accuracy and linguistic complexity compared to the PA group, although no significant between-group differences emerged in fluency or pronunciation ($p > 0.05$). Regarding self-efficacy, the SA group demonstrated significantly greater improvement in overall speaking confidence, particularly in interlocutory self-efficacy. Qualitative analyses indicated that SA participants developed stronger metacognitive awareness and self-regulatory strategies, although both groups reported technological and affective challenges. These findings suggest that SA may be particularly effective for developing productive language skills and speaking confidence in technology-mediated environments. The study extends current understanding of assessment modality effects on L2 development and offers practical implications for implementing online speaking assessment in EFL contexts.

[1] Tunghai University, Taichung, Taiwan, ROC. ✉email: minhsunl@thu.edu.tw

Introduction

English-speaking proficiency has become essential for academic success, professional advancement, and cross-cultural communication in today's interconnected global society (Crystal, 2003). Although it is imperative, English as a Foreign Language (EFL) learners worldwide face considerable challenges in developing oral proficiency, with many struggling to achieve communicative competence, even after years of formal instruction (Jamshidnejad, 2020). This global challenge is particularly evident in East Asian contexts, including Taiwan, where traditional teaching methods focus on grammar and test scores rather than on communicative competence (Chen et al., 2020). Consequently, students often have a firm grasp of grammar and vocabulary but cannot speak English confidently. This lack of confidence creates a negative cycle that impedes oral production. Recent research highlights the importance of self-efficacy in language learning, particularly for speaking proficiency. Students with high levels of self-efficacy are more likely to engage in speaking activities and perform better on oral tasks across various cultural and educational contexts (Hoesny et al., 2023; Wicaksono et al., 2023; Wijaya, 2024).

To promote L2 speaking self-efficacy, which is grounded in Bandura's (1997) social cognitive theory and its emphasis on mastery experiences and social modeling, alternative assessment methods[1] such as self-assessment (SA) and peer assessment (PA) have gained popularity in EFL classrooms. These approaches align with Vygotsky's (1978) sociocultural theory, where learning occurs through social interaction, and with Zimmerman's (2000) self-regulated learning framework, which emphasizes the cyclical nature of planning, monitoring, and reflection (Ashraf and Mahdinezhad, 2015; Wicaksono et al., 2023). While these theoretical frameworks highlight the importance of learner autonomy and self-regulation, implementing learner-centred speaking assessments remains challenging in traditional classroom settings where the teacher usually dominates the assessment practice. Technology integration has emerged as a feasible solution through platforms such as VoiceThread, a web-based multimedia tool that addresses these challenges by enabling users to create, share, and discuss audio, video, and text content asynchronously. VoiceThread is especially suited for speaking practice and evaluation in language learning contexts because it creates a low-anxiety environment for oral production while facilitating meaningful peer interaction and self-reflection. Although research findings on alternative assessments have been inconsistent, with some learners still perceiving teacher evaluations as being more reliable (Kumar et al., 2023; Ritonga et al., 2021), the shift from traditional paper-and-pencil methods (Phongsirikul, 2018) to digital platforms necessitates an investigation into how SA and PA function in digital environments, particularly given their potential to provide more flexible, authentic, and engaging speaking assessment experiences than conventional methods.

Despite the growing adoption of technology-mediated assessments in language learning, several critical gaps remain in understanding their effectiveness and implementation. First, while research has examined self- and peer assessments separately, few studies have systematically compared their relative effectiveness when mediated through online platforms. Second, despite the relationship between self-/peer assessments and speaking proficiency having been explored (Hoesny et al., 2023; Wicaksono et al., 2023; Wijaya, 2024), the mechanisms by which technology-mediated SA and PA influence self-efficacy development remain underexamined, especially in the Taiwanese EFL context, where students often exhibit high language anxiety and limited speaking confidence (Jamshidnejad, 2020). Third, there is insufficient empirical evidence regarding how social cognitive theory (Bandura, 1997) manifests in digital assessment environments (Phongsirikul, 2018), particularly in cultural contexts where traditional assessment practices are deeply entrenched (Chen et al., 2020).

To address these gaps, this study aimed to: (1) compare the effectiveness of VoiceThread-mediated self- and peer assessment on Taiwanese EFL learners' speaking proficiency and self-efficacy (Ashraf and Mahdinezhad, 2015; Wicaksono et al., 2023) and (2) explore students' perceptions of using VoiceThread for assessment activities. The findings will shed new light on how online platforms can facilitate self- and peer assessments while accounting for cultural and contextual factors specific to Taiwan's EFL environment. This research will provide practical insights for Taiwanese English teachers seeking to implement technology-mediated assessment approaches that address their students' unique needs and challenges in an increasingly interconnected global society (Crystal, 2003). In addition, the results will help bridge the gap between theoretical frameworks of self-regulated learning, and their practical application in online assessment environments, potentially informing the development of more culturally responsive and theoretically grounded assessment practices in similar EFL contexts.

Literature review

Theoretical framework. This study is grounded in two theoretical frameworks that inform the application of online self- and peer assessments to influence English speaking and self-efficacy: Bandura's (1986, 1997) self-efficacy theory and Vygotsky's (1978) sociocultural theory. These frameworks have been extensively applied in L2 speaking research (Al-khresheh and Alkursheh, 2024; Kim, 2024; Li et al., 2024), providing a comprehensive lens for understanding how different assessment modalities influence language learning outcomes and self-efficacy development.

Bandura's self-efficacy theory posits that learners' beliefs about their capabilities significantly influence their performance through four primary sources: mastery experiences, vicarious experiences, social persuasion, and emotional states. First, mastery experiences, the most crucial source of self-efficacy, involve accomplishing a task (Usher and Pajares, 2008). Mastery experiences are particularly influential in language learning contexts where repeated successes build confidence and competence (Li et al., 2024). In this study, VoiceThread-mediated SA provided learners with structured opportunities to review and evaluate their own speaking performances, which allowed them to track their progress over time, identifying both improvements and areas needing attention. The asynchronous nature of VoiceThread allowed learners to attempt multiple recordings until satisfied, potentially fostering positive mastery experiences.

Second, vicarious experiences occur through observing others' successes and failures, mainly when the observed individuals are considered similar to the observers themselves (Schunk and Usher, 2019; Wang and Li, 2019). Students gain exposure to various speaking models in PA conditions by watching their peers' recorded speeches. This exposure serves dual purposes: stronger performance provides aspirational models, while observing peers at similar proficiency levels helps normalize the learning process and demonstrates achievable progress. Third, social persuasion through verbal encouragement and feedback influences self-efficacy (Aben et al., 2022). PA operationalizes this source through structured peer feedback, while SA incorporates elements of self-persuasion through guided reflection protocols. VoiceThread's multimedia commenting features enable detailed, specific feedback that can reinforce positive self-beliefs while providing constructive suggestions for improvement. Fourth, emotional and physiological states affect how individuals judge

their capabilities (Phan and Ngu, 2016). VoiceThread's asynchronous environment potentially reduces speaking anxiety by removing immediate audience pressure. Both SA and PA conditions are designed to create low-anxiety evaluation environments where L2 learners can develop comfort with speaking English over time.

Lastly, complementing self-efficacy theory, Vygotsky's sociocultural framework emphasizes the social nature of language learning and development. This framework is particularly relevant to understanding how PA facilitates L2 speaking development through social interaction and collaborative evaluation. Furthermore, both assessment conditions incorporate structured evaluation rubrics and guiding questions that scaffold learners' developing ability to assess speaking performance. This scaffolding gradually develops students' metacognitive awareness and self-regulated learning. In short, the differential engagement of self-efficacy sources and sociocultural learning mechanisms between SA and PA conditions may explain potential variations in speaking proficiency and self-efficacy outcomes.

Self-efficacy and sociocultural theories offer a solid foundation for understanding how assessment practices impact the development of L2 speaking and self-efficacy. However, substantial pedagogical challenges remain in applying these concepts to online learning settings. It is still unclear how to optimize online assessments to activate various sources of self-efficacy while maintaining the social learning advantages highlighted by sociocultural theory. This study addresses this challenge by exploring how different assessment types (self- versus peer) delivered via VoiceThread may vary in their capacity to engage those theoretical mechanisms to facilitate L2 speaking improvement and boost self-efficacy.

Self-efficacy and L2 speaking. Learning to speak a second language well is far more complicated than just activating the language learner's cognitive processing and memorizing the grammar. Based on Bandura's social learning theory (1997), affective factors such as self-efficacy also play a crucial role in shaping EFL learners' oral proficiency. EFL teachers have frequently observed in their classrooms that students with poor proficiency in English are mostly those who do not believe they can learn English well. Recent research has consistently shown a strong positive correlation between learners' self-efficacy and their EFL achievements (Kim, 2024; Li et al., 2024).

Bandura (1986) defined "self-efficacy" as learners' self-belief in learning particular tasks or skills. Learners with high self-efficacy tend to face challenges rather than avoid them, and are able to recover quickly from setbacks. Therefore, Bandura proposed that "the stronger the perceived self-efficacy, the higher the goal challenges people set for themselves" (1986, p. 362). Research in this area continues to validate the critical implications of self-efficacy theory in academic settings, showing that learners' self-perceived efficacy affects their academic achievement (Alkhresheh and Alkursheh, 2024; Wicaksono et al., 2023). In L2 learning specifically, recent research has demonstrated that anxiety and self-efficacy influence students' speaking skills (Wijaya, 2024), while public speaking self-efficacy has been found to predict achievement (McNatt, 2019).

In addition, literature on the relationship between self-efficacy beliefs and performance implies that self-efficacy is susceptible to change via positive experiences. Studies investigating self-efficacy belief change have consistently argued that self-efficacy beliefs are flexible and subject to development, rather than being fixed (Alkhresheh and Alkursheh, 2024; Bandura and Schunk, 1981; Kim, 2024). The critical issue is developing instructional activities and novel applications to enable students to experience success and to reconstruct their self-efficacy perceptions accordingly. Goal setting and online technology use have been identified as effective mechanisms for cultivating self-efficacy beliefs (Basileo et al., 2024; Wang and Sun, 2024).

In brief, prior studies suggested that the relationship between mastery experience and self-efficacy is cyclical and two-way: greater efficacy leads to more significant effort and persistence, resulting in better performance and enhanced efficacy. While much research has focused on learners' perceived self-efficacy in foreign language learning (Li and Zhang, 2023; Sengul and Buyukkarci, 2023), there remains a gap in understanding how specific instructional strategies, such as VoiceThread-mediated self- and peer assessment, might affect L2 speaking self-efficacy. In EFL classrooms, adopting alternative assessment methods such as self- and peer evaluation has proven valuable for promoting goal setting and enhancing learner autonomy (Ashraf and Mahdinezhad, 2015; Liao, 2023; Wicaksono et al., 2023). These assessment practices may play a crucial role in developing students' self-efficacy. More empirical studies are needed to understand how these assessment practices play a crucial role in developing students' self-efficacy by activating multiple sources of efficacy information within a sociocultural framework of learning and development.

Theoretical and empirical studies on self-efficacy in L2 speaking have pointed out a significant pedagogical challenge. Although self-efficacy beliefs are malleable and crucial for speaking development, less is known about how specific instructional approaches can effectively cultivate these beliefs in online environments. This study's second research question intended to fill the gap by examining how different VoiceThread-mediated assessment types might affect L2 speaking self-efficacy. Understanding these relationships could help educators design more effective online speaking activities that simultaneously develop linguistic competence and strengthen self-efficacy beliefs.

Self- and peer assessment in L2 learning. Recent studies on self- and peer assessment in language learning have consistently demonstrated their positive impact on learners' language proficiency, self-efficacy, and autonomy. For instance, research by Aldosari and Alsager (2023) highlighted how SA improves language skills and fosters resilience, creativity, and learner autonomy among EFL learners. These findings align with earlier studies that emphasized the role of SA in promoting metacognitive skills and self-directed learning (Alibakhshi and Sarani, 2014; Chen, 2010). Research suggests that these assessment practices can significantly enhance oral language skills. Alibakhshi and Sarani (2014) studied how SA affects speaking fluency and accuracy. They worked with 60 EFL students at pre- and upper-intermediate levels, split between a control and an experimental group. Their pre- and post-test results confirmed that SA boosted speaking accuracy and fluency, suggesting its value for students and teachers. SA has proven particularly beneficial in speaking classes.

Besides SA, PA has also been shown to contribute significantly to language development. A study by Esfandiari and Tavassoli (2019) demonstrated that PA enhances learners' oral proficiency and boosts their confidence and social skills by fostering a sense of accountability. Joo (2016) examined students' ability to evaluate both their oral and peers' work. The findings highlighted that speaking improvement stems from the assessment and students' active involvement in the evaluation process. Similarly, research by Ashraf and Mahdinezhad (2015) found that PA activities positively impacted learner autonomy in speaking tasks. The collaborative nature of PA allows students to reflect on their peers' performances while improving their own. Moreover, Zarei

and Usefli (2019) investigated how different assessment types affected Iranian EFL learners' general and academic self-efficacy. After taking the Primary English Test (PET), 94 students were divided into three groups, each experiencing self-, peer, or teacher assessment. Students completed general and academic self-efficacy questionnaires before and after the intervention. ANCOVA analysis revealed that while the assessment types had similar impacts on general self-efficacy, self- and peer assessment proved more effective than teacher assessment for academic self-efficacy. Interestingly, self- and peer assessment showed comparable effectiveness in boosting academic self-efficacy.

Studies comparing the effects of self- and peer assessment on second language speaking have revealed that both assessment methods can positively impact learners' speaking skills, although their effects may differ across implementation contexts. Imani (2021) found that self- and peer assessment had comparable positive effects on both impulsive and reflective EFL learners' speaking ability, with no significant interaction between assessment type and cognitive style. However, Maleki et al. (2024) observed that SA was more effective than PA in terms of improving ESP students' speaking ability and reducing their second language speaking anxiety. This finding was echoed by Zheng et al. (2024), who pinpointed the optimal sequence for integrating SA first for high-anxiety learners versus introducing PA for low-anxiety learners. Zheng et al. (2024) studied AI-supported formative assessment in English public speaking with three key findings: peer assessment led to higher social engagement than automated assessment, students had mixed views on different assessment types, and both self- and peer assessment improved speaking skills, with varying impacts on anxiety, engagement, and specific language abilities. The study highlights the limitations of automated assessment in replacing peer interaction for language learning.

However, challenges related to the reliability of self- and peer assessment persist. Studies have identified issues such as assessor bias or evaluation inconsistency (Singh, 2015; Yinjaroen and Chiramanee, 2011). For example, some learners overestimate or underestimate their abilities compared to teacher assessments (Patri, 2002). In light of these challenges, implementing structured guidelines is key to mitigating such issues. Falchikov (2007) suggested that providing clear criteria and training can improve the accuracy of both self- and peer assessments. For example, Chen's (2010) study showed that Chinese university students' ratings aligned more closely with teachers' assessments after two rounds of practice and guidance.

To explicitly compare SA and PA, Table 1 summarizes the language-learning-related benefits and the limitations of SA and PA.

In short, the literature indicates that both SA and PA offer distinct yet complementary benefits for L2 speaking development. SA appears particularly effective for improving speaking accuracy/fluency and reducing anxiety (Maleki et al., 2024), whereas PA contributes to social engagement and collaborative learning (Zheng et al., 2024). Both assessment types show promising effects on learners' academic self-efficacy (Zarei and Usefli, 2019); however, their relative effectiveness may vary depending on learner characteristics, such as anxiety levels (Zheng et al., 2023). Most existing studies have examined these assessment types in traditional classroom settings, leaving their comparative effects largely underexplored in online environments, which leads to the first research question. Moreover, considering the implementation challenges identified by the previous research, including assessment reliability and peer feedback quality (Singh, 2015; Zheng et al., 2024), online platforms have emerged as a feasible medium to mitigate the drawbacks of alternative assessments with their affordance of providing clear guidance, convenient training, and constant feedback for optimal self- and peer assessment. This potential of online platforms for enhancing SA and PA practice necessitates investigating how tools such as VoiceThread might mediate these assessment types to optimize L2 speaking development.

Online assessment in L2 speaking. Online assessment, or e-assessment, refers to using digital tools to deliver, store, and grade student performance in various forms, including text, audio, video, and images. It can be conducted synchronously or asynchronously for individuals or groups (Crisp, 2011). Recent studies have documented the positive impact of online assessment on L2 speaking skills. One effective tool for online assessment is VoiceThread, which harnesses unique affordances to enhance oral proficiency through self- and peer assessment. The platform's asynchronous feedback system enables assessors to provide targeted commentary at specific moments in speakers' oral productions. At the same time, its multimodal interface combines audio, visual, and text-based interactions to support comprehensive assessment.

For example, in Pagkalinawan's (2021) study, students who recorded and assessed their presentations via VoiceThread improved their vocabulary, grammar, fluency, and pronunciation. The platform's multimodal features allowed students to receive feedback in multiple formats. The asynchronous nature of the tool boosted their confidence by enabling them to review and respond to feedback multiple times, leading to enhanced preparation and delivery skills. Similarly, Salehi and Rowan (2015) found that VoiceThread's multimodal commenting features facilitated detailed peer assessment, allowing students to critically evaluate each other's work through oral and written feedback, improving their speaking performance. Moreover, Liao's (2023) study on online self-and peer assessment demonstrated how VoiceThread's combination of visual, audio, and text-based interaction supported the development of learner autonomy and improved L2 English speaking skills. Participants in the experimental group who engaged with these multimodal assessment features outperformed the control group in speaking performance and demonstrated a stronger sense of learner autonomy.

Studies examining students' perceptions of online speaking assessments have revealed both platform-specific benefits and challenges. Students applauded VoiceThread for its multimodal feedback options and the flexibility of working on assessments at their convenience (Amalia, 2018; Komang, 2021; Wibowo and Novitasari, 2021). Similarly, Çetin Köroğlu's (2021) study with 52 English-major university students in Turkey revealed that participants using digital formative assessment tools outperformed those taking traditional speaking tests, attributing this success to the opportunity for learner self-reflection and peer collaboration through task completion.

However, it is important to note that VoiceThread, like any tool, has its challenges. For instance, the delayed feedback from the asynchronous nature of the platform could disrupt the assessment flow. Giving specific feedback on speaking skills in a digital format can be challenging for L2 speakers, and so may require modeling and training. While some challenges stem from technical aspects of oral recording and playback quality, others involve pedagogical considerations, such as providing adequate training, keeping assessment consistent across different review sessions, and ensuring effective use of multimodal feedback tools. In short, while online speaking assessment platforms offer enhanced opportunities for detailed feedback and engagement in second language acquisition, successful implementation requires careful attention to platform-specific affordances and limitations

Table 1 Comparing self-assessment and peer assessment.

	Self-Assessment	Peer Assessment
Language-learning-related benefits	• Improves language skills and fosters resilience, creativity, and learner autonomy (Aldosari and Alsager, 2023)	• Enhances oral proficiency while boosting confidence, social skills and a sense of accountability among learners (Esfandiari and Tavassoli, 2019)
	• Enhances speaking accuracy and fluency while promoting metacognitive skills and self-directed learning (Alibakhshi and Sarani, 2014; Chen, 2010)	• Positively impacts learner autonomy in speaking tasks (Ashraf and Mahdinezhad, 2015)
	• More effective than peer assessment in improving ESP students' speaking ability and reducing second language speaking anxiety (Maleki et al., 2024)	• Promotes social interaction and engagement among learners (Zheng et al., 2024)
	• Particularly beneficial as an initial assessment method for high-anxiety learners (Zheng et al., 2023)	• Enables students to reflect on peers' performances while improving their own
	• Comparable to peer assessment in boosting academic self-efficacy (Zarei and Usefli, 2019)	• Particularly beneficial as an initial assessment method for low-anxiety learners (Zheng et al., 2023)
		• More effective than teacher assessment for academic self-efficacy (Zarei and Usefli, 2019)
Limitations	• Issues with reliability due to assessor bias (Singh, 2015; Yinjaroen and Chiramanee, 2011)	• Concerns about peers' qualifications and their varying levels of willingness to provide feedback (Zheng et al., 2024)
	• Students may overestimate or underestimate their abilities compared to teacher assessments (Patri, 2002)	• Issues with evaluation inconsistency (Singh, 2015; Yinjaroen and Chiramanee, 2011)
	• May not provide the social engagement benefits that come with peer interaction (Zheng et al., 2024)	• May be less effective than self-assessment for anxiety reduction in certain contexts (Maleki et al., 2024)

The literature on online speaking assessment highlights both opportunities and challenges in implementing digital tools for L2 speaking development. While platforms such as VoiceThread offer unique affordances for multimodal feedback and flexible assessment opportunities, successful implementation requires careful consideration of both technical and pedagogical factors. This connects directly to all three research questions in this study, as how VoiceThread-mediated assessment types affect both learning outcomes and learner perceptions is examined.

The present study. Despite growing interest in online speaking assessment for second language learners, significant gaps exist in understanding how theoretical frameworks, assessment practices, and digital platforms intersect to support L2 speaking and self-efficacy development. First, while self-efficacy theory and sociocultural theory suggest different mechanisms through which assessment practices might influence learning outcomes, empirical evidence about how these mechanisms operate in online environments remains scarce. Second, although research demonstrates the potential of both self- and peer assessment for L2 speaking development, educators face continued challenges in implementing these practices effectively in online contexts. Third, while digital platforms such as VoiceThread offer promising features for speaking assessment, a better understanding of how to leverage these tools to optimize speaking development and self-efficacy growth is needed. This study intended to fill these gaps by addressing the following three research questions:

1. What are the effects of VoiceThread-mediated self-assessment compared to peer assessment on the oral proficiency development of L2 English speakers?
2. What are the effects of VoiceThread-mediated self-assessment compared to peer assessment on the speaking self-efficacy of L2 English speakers?
3. What are L2 English speakers' perceptions of VoiceThread-mediated self- and peer assessments after participating in them for 1 year?

Methods

Research design. This mixed-methods study investigated the differential effects of VoiceThread-mediated SA versus PA on English-speaking proficiency and self-efficacy, grounded in self-regulated learning and sociocultural theory. The design aimed to address critical challenges in EFL contexts, such as low sense of confidence among Taiwanese students, limited opportunities for authentic speaking practice, and lack of formative assessment. VoiceThread-mediated assessment was chosen to create a structured environment that promotes self-regulated learning through its asynchronous and multimodal features. VoiceThread's asynchronous and multimodal nature helps reduce performance anxiety by allowing multiple recording attempts, enables deeper reflection through unlimited replay capabilities, facilitates formative assessment through standardized protocols, and creates a permanent record of progress for metacognitive monitoring. Participants were 37 adult English as a foreign language (EFL) students (power analysis indicated minimum required $n = 35$, $\alpha = 0.05$, power = 0.80, Faul et al., 2009) enrolled in a required English Oral Training One course at a 4-year university in central Taiwan. Two intact classes, previously formed based on standard university enrollment procedures, were randomly assigned to SA ($n = 19$) and PA ($n = 18$) conditions.

Drawing from Vygotsky's sociocultural theory, the design leverages technology-mediated interaction to create zones of proximal development, where learners develop oral proficiency through structured self-reflection (SA group) or peer scaffolding (PA group). The assessment protocols were designed to facilitate the internalization of speaking strategies through systematic engagement with cultural tools (VoiceThread features) and metacognitive processes. The implementation protocol operationalized key theoretical principles while addressing practical challenges in EFL contexts. For the SA group, the three-step evaluation process exemplifies self-regulated learning theory through its emphasis on planning (pre-speech preparation), monitoring (performance review), and reflection (video-based commentary). Similarly, the PA group's systematic review process creates opportunities for collaborative learning and peer

scaffolding, consistent with sociocultural theory's emphasis on social interaction in learning.

The quasi-experimental design was used to examine assessment types as the independent variable. English oral proficiency (measured using the validated in-house speaking assessment, inter-rater reliability $\alpha = 0.89$) and English-speaking self-efficacy (measured using the ESSEQ scale, Cronbach's $\alpha = 0.92$) were dependent variables. The random assignment of intact classes, rather than individual students, was chosen to maintain ecological validity and prevent treatment diffusion while still providing reasonable control over potential confounds. Baseline equivalences in dependent variables between groups were established via independent samples *t*-tests. Two independent samples *t*-tests were conducted to establish baseline equivalence between the SA and PA groups based on their pre-test scores. For English oral proficiency, there was no significant difference between the SA group ($M = 13.84$, SD = 1.97) and the PA group ($M = 14.56$, SD = 1.55), $t(35) = 1.221$, $p = 0.230$. Similarly, speaking self-efficacy scores did not differ significantly between the SA group ($M = 3.46$, SD = 0.48) and the PA group ($M = 3.54$, SD = 0.62), $t(35) = -0.416$, $p = 0.680$. These results indicate that both groups began the study with comparable levels of English oral proficiency and speaking self-efficacy, thereby supporting the assumption of baseline equivalence between the groups. Both quantitative and qualitative data were collected over the academic year. Quantitative data included pre- and post-intervention measures of oral proficiency and self-efficacy. Qualitative data comprised a post-project questionnaire, open-ended survey, and semi-structured group interview.

VoiceThread-mediated assessments were conducted monthly following a standardized protocol. ANCOVA was employed to analyse quantitative data after confirming assumptions of normality and homogeneity of variance. The selection of ANCOVA over repeated measures ANOVA was guided by both statistical and methodological considerations. Given the quasi-experimental design with two intact classes (SA = 19, PA = 18), ANCOVA offered superior statistical control by treating pre-test scores as covariates, thus accounting for initial between-group differences in English proficiency and speaking self-efficacy. This approach enhanced statistical power through reduced error variance, and provided more precise estimates of the intervention effects while controlling for pre-existing differences that could influence the outcomes. Additionally, the primary research focus was comparing the effectiveness of two assessment approaches rather than the temporal pattern of change; ANCOVA aligned more closely with the analytical objectives. Qualitative data underwent thematic analysis following a systematic approach. Integration of quantitative and qualitative findings followed a convergent parallel design.

Participants and pedagogical context. Using convenience sampling, 37 first-year, English-major students from the two intact Oral Training One classes taught by the researcher were recruited with their informed consent. Convenience sampling provided practical and easy access to the participants. However, the researcher acknowledges its potential limitations, such as reduced generalizability, potential bias, and insufficient representation of the broader population. Participants' English proficiency levels, based on the institution's placement examination scores calibrated to the Common European Framework of Reference for Languages, ranged from B1 to B2. Two intact classes were randomly assigned to the SA or PA condition. Prior to data collection, institutional review board approval was obtained, and all participants provided written informed consent.

The English Oral Training One course, a year-long required component of the undergraduate English curriculum at a comprehensive university in Taiwan, was designed to develop students' public speaking competence. The course curriculum comprised eight systematically sequenced speech tasks of increasing complexity, ranging from a 3-minute descriptive speech about a meaningful object to a 7-min persuasive presentation. Speech assignments were spaced approximately 2–3 weeks apart to allow adequate preparation time. Both SA and PA groups received identical instructional content, speech assignments, and evaluation criteria to maintain treatment fidelity. During non-presentation weeks, instruction focused on genre-specific rhetorical strategies, organizational patterns, and delivery techniques. Model speeches demonstrating varying levels of performance were analysed to establish clear quality benchmarks. Prior to each speech task, detailed scoring rubrics were provided and explained.

All speeches were video-recorded and uploaded to VoiceThread for subsequent assessment activities. The instructor provided standardized 2–3 min oral feedback to each student immediately following their in-class presentations, using consistent evaluation criteria across both groups.

Implementation of VoiceThread-mediated self- and peer assessment. Following the in-class speech delivery, students in both groups were required to complete their retrospective, VoiceThread-mediated assessment by the end of the following week. VoiceThread, a web-based multimodal collaboration platform, was selected for its multimedia capabilities and accessibility features. The platform enables asynchronous video-based interaction, and supports both individual reflection and peer feedback through integrated multimedia commenting tools.

The assessment protocol was implemented consistently across all eight speech tasks throughout the academic year. The VoiceThread platform facilitated the systematic scaffolding of assessment skills through several structured supports. Initially, the instructor provided students with detailed assessment guidelines and exemplar videos demonstrating effective evaluation techniques. The platform's multimedia interface allowed the instructor to embed time-stamped comments at specific points in sample speeches, modeling how to provide focused feedback on various aspects of oral performance (e.g., pronunciation, organization, and delivery). The following forms of scaffolding were carried out to prepare participants for actual assessments.

For the SA group, scaffolding was implemented through a three-tiered process. First, students used guided SA forms that prompted reflection on specific speech components. Second, they analysed instructor-annotated model speeches that highlighted key assessment criteria. Third, they practiced self-evaluation with shorter speech segments before posting a finalized one. The platform's playback features enabled students to revisit specific portions of their speeches repeatedly, supporting detailed analysis and reflection.

The PA group followed a similar form of scaffolding but emphasized peer interaction features. Students first practiced peer assessment using pre-selected speech samples with instructor guidance. The platform's threaded commenting feature allowed collaborative dialogue about assessment criteria and standards. Students could build upon each other's observations while receiving peer feedback. The asynchronous nature of VoiceThread allowed students to carefully construct and revise their feedback before posting it on VoiceThread.

To facilitate peer-assessment, all participants were placed in groups of three or four on VoiceThread. Participants in the SA group completed a three-step evaluation process: reviewing their recorded speech performance on VoiceThread, completing a structured SA form, and recording video-based reflective

commentary using specified assessment criteria (see Appendix A). For the PA group, participants followed a systematic review process that involved viewing their group members' recorded speeches, analysing performance using standardized assessment criteria, and recording video-based evaluative feedback.

To ensure meaningful engagement with the feedback, students were required to submit their 3-min-long video-recorded comments on VoiceThread, guided by structured assessment sheets. This standardized length was chosen to ensure consistency in feedback depth across all assessments. The students were encouraged to address all the evaluation criteria on the assessment sheets rather than just meeting the minimal length of the video-recorded assessment. The teaching assistant monitored both submission of assessments and the quality of the submission. All students had a week after their speech to complete their assessments (self- or peer), with grade penalties for late submissions. In groups of three to four students, each PA participant evaluated two group members' speeches to ensure a balanced distribution of peer feedback within the group, and to avoid always receiving feedback from the same group member. English was used for all recordings as it is the primary language of instruction in this English department. While this might have affected some students' ability to express complex thoughts, it aligned with the course's language learning objectives and provided additional speaking practice.

In sum, assessment guidelines and rubrics were standardized across both conditions, with only the focus of evaluation (self- vs. peer) differing between groups. Regular monitoring of assessment completion and feedback quality was conducted to maintain implementation fidelity. The structured nature of both assessment protocols ensured consistency in feedback provision while accommodating the distinct characteristics of the self- and peer evaluation processes.

Data collection. Quantitative data were collected from the English proficiency evaluation, English-speaking self-efficacy questionnaire (ESSEQ), and VoiceThread-mediated assessment survey, while qualitative data were collected from a post-project open-ended question survey and group interview. English proficiency evaluation and ESSEQ were administered to the participants before and after the year-long engagement in VoiceThread-mediated assessments. The perception questionnaire survey, open-ended survey, and group interview were conducted toward the end of the project. Four data collection instruments were utilized to answer the three research questions.

- **English speaking proficiency test:** 37 students recorded one 3-min speech before and after the intervention. The two topics, "My most memorable vacation" and "My most unforgettable achievement", were counterbalanced between tests through random assignment within each condition (SA and PA). Within each condition, approximately half the students were randomly assigned to speak about their vacation in the pre-test and about their achievement in the post-test. At the same time, the other half did the reverse. This between-test counterbalancing aimed to reduce potential differences in topic difficulty or complexity. Both topics were chosen to be similar, encouraging comparable narrative skills and complexity, thus alleviating the effect of topic familiarity on performance outcomes.
  Three trained raters, blind to participant condition and assessment timing, evaluated the recordings using an adapted TOEFL Independent Speaking Rubric (interrater reliability = 0.89, Appendix B). The rubric assessed delivery (fluency and pronunciation) and language use (accuracy and complexity) on a five-point scale (1 = very poor to 5 = very good). Three experienced lecturers from the English language centre, each with at least 5 years of teaching experience, served as raters. The raters underwent a 3-h training session conducted by the researcher. It involved carefully listening to five sample speech videos, studying the scoring rubrics, and comparing scores among the three raters until they established 89% inter-rater reliability. Every recorded speech was evaluated independently by two raters, with the average of their scores serving as the pre-test or post-test English oral proficiency scores. Each rater scored 24 to 25 recorded speeches for the pre-test at the beginning and another 24 or 25 for the post-test at the end of the academic year.
- **The English-speaking self-efficacy questionnaire (ESSEQ, Appendix C):** This questionnaire was adapted from Alavi et al. (2004) validated questionnaire through a systematic process. The adaptation involved reviewing the items to align with the study context. Two experts in computer-assisted language learning and educational psychology reviewed the adapted items using a standardized evaluation rubric focusing on content relevance, linguistic clarity, and cultural appropriateness. Their feedback led to the finalized version of the 24-item ESSEQ, which was pilot-tested with 15 students matching the target population's characteristics. The final instrument measured three dimensions:
  Accuracy (8 items, $\alpha = 0.92$): measures participants' confidence in grammatical and phonological precision (e.g., "I can use accurate grammar while speaking English" and "I am able to speak with correct pronunciation").
  Fluency (9 items, $\alpha = 0.91$): assesses confidence in maintaining natural, smooth speech flow (e.g., "I am capable of speaking for a few minutes in such a way as to clearly express my ideas" and "I am able to give a well-organized speech").
  Interlocutory self-efficacy (7 items, $\alpha = 0.93$): Evaluates confidence in various interactive speaking contexts (e.g., "I am comfortable talking with foreigners" and "I am able to discuss subjects of general interest with my classmates in English"). Respondents used a five-point Likert scale (1 = strongly disagree to 5 = strongly agree).
- **The VoiceThread-mediated self-/peer assessment questionnaire (VTSQ):** This questionnaire was administered post-intervention to assess participants' perceptions of the assessment process. The 19-item instrument was adapted from Chiang's (2018) validated questionnaire to address the specific context of VoiceThread-mediated assessment. The questionnaire encompassed four key aspects: (a) the usefulness of VoiceThread-mediated assessment for enhancing speaking fluency, (b) the usefulness of VoiceThread-mediated assessment for enhancing accuracy, (c) the usefulness of VoiceThread-mediated assessment for cultivating speaking self-efficacy, and (d) the features and functions of the VoiceThread platform for assessment purposes. Two parallel versions were developed: one for the SA group and one for the PA group, with items appropriately worded to reflect the respective assessment type (e.g., "Reviewing my own speech…" vs. "Reviewing my peers' speech…"). The equivalence between versions was established through expert review. Participants rated items on a five-point Likert scale (1 = strongly disagree to 5 = strongly agree). The overall instrument demonstrated strong internal consistency ($\alpha = 0.94$) and content validity through expert review.
- **Qualitative data:** The study employed a sequential qualitative data collection approach, beginning with open-ended surveys followed by semi-structured group

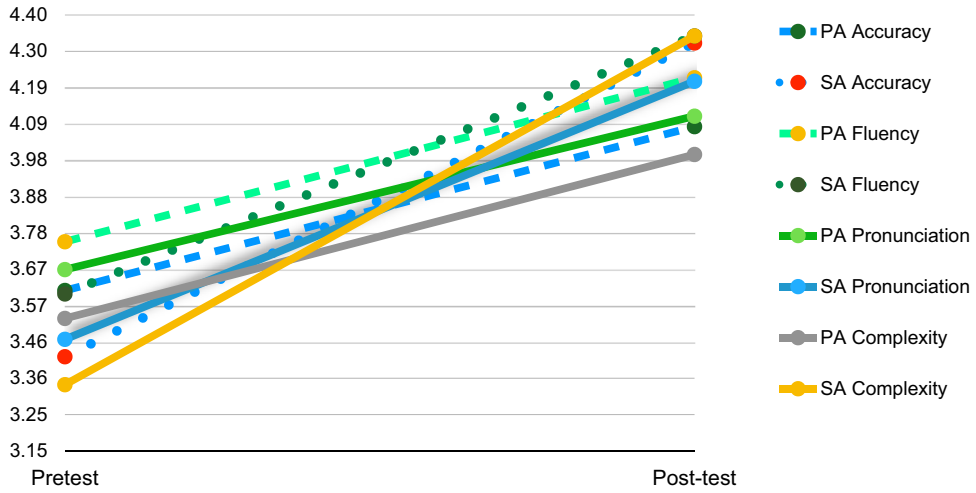

Fig. 1 Comparing the English oral proficiency sub-skills of the PA and SA groups.

interviews. Both the surveys and interviews were conducted in English, the primary language of instruction in the program, to maintain consistency with the learning environment. This sequential design was deliberately chosen to allow participants to first reflect on and document their experiences thoughtfully through written responses, without the immediate pressure of face-to-face interaction. The open-ended survey explored participants' experiences with VoiceThread-mediated assessment, including perceived benefits, challenges, linguistic gains, and future implementation suggestions. The SA group was asked about reviewing their recorded speech, whereas the PA group was asked about giving comments to and receiving comments from their peers. The written format provided participants with time to process their year-long learning journey and to articulate their thoughts in detail. Building on these initial reflections, follow-up semi-structured group interviews (45–60 min) were conducted to probe more deeply into themes emerging from the survey responses. The sample interview questions included inquiries about changes in video-recording comments with VoiceThread (for both), suggestions for platform improvements, self-reflection on recording review processes and personal growth in self-observation (for SA), and perspectives on balancing supportive and constructive peer feedback (for PA). Twelve participants (six from the SA and six from the PA group) were purposively selected for the interviews based on the following criteria: participants provided detailed examples of their VoiceThread experiences that aligned with the study's focus on assessment practices; participants raised unique challenges or benefits that needed elaboration; and participants whose initial responses suggested distinctive patterns in their use of the platform. This sequential approach enabled participants to engage in both individual written reflection and collaborative dialogue, where they could elaborate on their written responses and react to peers' experiences.

Data analysis. Quantitative analyses were conducted using IBM SPSS 27.0. Analysis of covariance (ANCOVA) was employed to examine the differential effects of VoiceThread-mediated SA versus peer assessment on participants' English-speaking proficiency and self-efficacy. Pre-test scores served as covariates to control for initial group differences. Prior to conducting ANCOVAs, assumptions of normality (Shapiro-Wilk test), homogeneity of variance (Levene's test), and homogeneity of regression slopes were tested. Effect sizes were reported using partial eta squared (η^2p). Descriptive statistics and reliability coefficients were calculated for the VoiceThread-mediated Self-/Peer Assessment Questionnaire (VTSQ) to examine participants' perceptions of the assessment process.

A systematic approach was employed for qualitative data analysis to analyse responses from open-ended surveys and semi-structured group interviews (2 groups, 45–60 min each). Initial thematic analysis of survey responses informed the subsequent interview protocols. Using an iterative coding process, two researchers independently coded interview transcripts to triangulate with the tentative themes identified from the open-ended survey. When coding disagreements occurred, the researchers discussed until a consensus was reached. Emerging themes were categorized into main themes and sub-themes using constant comparative analysis. Chi-square tests were conducted to compare the occurrence of themes between the SA and PA groups. Thus, the potential group differences in their perceived effects of self- versus peer assessment were identified. Following mixed-methods best practices, qualitative findings were integrated with quantitative results through methodological triangulation to enhance the validity and comprehensiveness of interpretations. To ensure trustworthiness of the qualitative analysis, member checking was conducted with voluntary participants ($n = 12$) by sharing the identified themes and interpretations for their feedback.

Results

Effects of VoiceThread-mediated self-assessment versus peer assessment on English oral proficiency.

- Independent oral assessments were administered to the SA ($n = 19$) and PA ($n = 18$) groups at the beginning and toward the end of the school year. The PA had slightly higher pre-test scores ($M = 14.56$, SD = 1.55) compared to the SA ($M = 13.84$, SD = 1.96) in terms of overall oral proficiency.

Fig. 1 shows how English oral sub-skills developed differently between the groups. After the interventions, both groups demonstrated improvement in their post-test scores, with the SA showing a considerably higher overall mean score ($M = 17.50$, SD = 0.50) compared to the PA ($M = 16.42$, SD = 1.05). This pattern of improvement was consistent across most subcategories, including accuracy (PA: $M = 4.08$, SD = 0.31; SA: $M = 4.32$,

Table 2 ANCOVA results of English oral proficiency ($n = 37$).

Oral Proficiency Overall	Peer Assessment Group (PA) ($n = 18$)				Self-Assessment Group (SA) ($n = 19$)				ANCOVA		
	Pre-test		Post-test		Pre-test		Post-test				
	M	SD	*M*	SD	*M*	SD	*M*	SD	*F* (1,34)	η^2	Effect size
	14.56	1.55	16.42	1.05	13.84	1.96	17.50	0.50	15.60***	0.32	Large
Accuracy	3.61	0.40	4.08	0.31	3.42	0.51	4.32	0.25	8.51**	0.20	Large
Fluency	3.75	0.52	4.22	0.43	3.60	0.54	4.34	0.29	1.43	0.04	-
Pronunciation	3.67	0.49	4.11	0.27	3.47	0.59	4.21	0.25	2.52	0.09	-
Complexity	3.53	0.40	4.00	0.34	3.34	0.53	4.34	0.24	12.70**	0.27	Large

**$p < 0.01$, ***$p < 0.001$.

SD = 0.25), fluency (PA: $M = 4.22$, SD = 0.43; SA: $M = 4.34$, SD = 0.29), pronunciation (PA: $M = 4.11$, SD = 0.27; SA: $M = 4.21$, SD = 0.25), and complexity (PA: $M = 4.00$, SD = 0.34; SA: $M = 4.34$, SD = 0.24).

- ANCOVAs were conducted with assessment type (peer vs. self-assessment) as the independent variable on four dimensions of English oral proficiency, using post-test scores as dependent variables and pre-test scores as covariates. The assumptions of normality and homogeneity of variances were met for all analyses (Levene's tests: oral proficiency $F(1, 35) = 3.757$, $p = 0.061$; accuracy $F(1, 35) = 0.142$, $p = 0.708$; complexity $F(1, 35) = 0.059$, $p = 0.809$; fluency $F(1, 35) = 0.404$, $p = 0.529$; pronunciation $F(1, 35) = 0.032$, $p = 0.860$).

Table 2 shows significant effects of assessment type on oral proficiency ($F(1, 34) = 15.603$, $p < 0.001$, partial $\eta^2 = 0.32$), accuracy ($F(1, 34) = 8.511$, $p = 0.006$, partial $\eta^2 = 0.200$), and complexity ($F(1, 34) = 12.699$, $p = 0.001$, partial $\eta^2 = 0.272$). The SA group consistently outperformed the PA group across these dimensions. Specifically, the SA group showed higher scores in oral proficiency ($M = 17.50$, SD = 0.50 vs. $M = 16.42$, SD = 1.05), accuracy ($M = 4.32$, SD = 0.25 vs. $M = 4.08$, SD = 0.31), and complexity ($M = 4.34$, SD = 0.24 vs. $M = 4.00$, SD = 0.34).

No significant differences were found between groups for fluency or pronunciation ($\eta^2 = 0.04$ and 0.09, respectively). Pre-test scores as covariates were non-significant for oral proficiency ($F(1, 34) = 0.033$, $p = 0.856$, partial $\eta^2 = 0.001$) and complexity ($F(1, 34) = 0.406$, $p = 0.528$), while accuracy showed a marginal relationship ($F(1, 34) = 3.136$, $p = 0.086$, partial $\eta^2 = 0.084$).

- The findings revealed differential effects of assessment type on various aspects of L2 oral performance. SA yielded significantly better outcomes than PA for oral proficiency, accuracy, and complexity measures. For overall oral proficiency, the SA group ($M = 17.50$, SD = 0.50) outperformed the PA group ($M = 16.42$, SD = 1.05). Similarly, significant advantages were found for the self-assessment group in both accuracy ($F(1, 34) = 8.511$, $p = 0.006$, partial $\eta^2 = 0.200$) and complexity ($F(1, 34) = 12.699$, $p = 0.001$, partial $\eta^2 = 0.272$). The impact of assessment type was not uniform across all oral language dimensions. Fluency and pronunciation showed insignificant between-group differences. The type of assessment had a minimal impact on fluency and pronunciation. Pre-test performance did not significantly influence post-test outcomes, except for a marginal effect on accuracy. The effectiveness of SA was independent of learners' initial proficiency levels, particularly for overall oral proficiency and complexity measures.

Effects of VoiceThread-mediated self-assessment versus peer assessment on English speaking self-efficacy.

- The English-speaking self-efficacy questionnaire (ESSEQ) was administered to both the SA ($n = 19$) and PA ($n = 18$) groups before and after the year-long intervention. Fig. 2 compares how English-speaking self-efficacy developed differently between the groups. The SA group demonstrated more substantial growth in overall self-efficacy from the pre-test ($M = 3.46$, SD = 0.48) to the post-test ($M = 3.72$, SD = 0.41) compared to the PA group's modest improvement from the pre-test ($M = 3.54$, SD = 0.62) to the post-test ($M = 3.58$, SD = 0.61). The PA group's accuracy increased from the pre-test ($M = 3.44$, SD = 0.54) to the post-test ($M = 3.64$, SD = 0.60), while the SA group demonstrated similar growth from the pre-test ($M = 3.21$, SD = 0.54) to the post-test ($M = 3.71$, SD = 0.25). For fluency, the PA group improved from the pre-test ($M = 3.50$, SD = 0.66) to the post-test ($M = 3.67$, SD = 0.71), and the SA group showing slight enhancement from the pre-test ($M = 3.50$, SD = 0.65) to the post-test ($M = 3.64$, SD = 0.67). For interlocutory self-efficacy, the PA group exhibited a decline from the pre-test ($M = 3.67$, SD = 0.84) to the post-test ($M = 3.44$, SD = 0.75), while the SA group demonstrated continued improvement from the pre-test ($M = 3.68$, SD = 0.47) to the post-test ($M = 3.81$, SD = 0.62), suggesting differential impacts of assessment approaches on students' perceived interactive speaking capabilities.

- After confirming normal distribution of residuals and homogeneity of variances (Levene's tests: all $ps > 0.14$), ANCOVAs were conducted with assessment type (self- vs. peer) as the independent variable, using post-test scores of four dimensions of English-speaking self-efficacy as dependent variables, controlling for the pre-test scores. SA was found to be more effective than PA in terms of enhancing students' overall English-speaking self-efficacy, particularly in their confidence during interactive speaking tasks. Table 3 shows significant differences in overall English-speaking self-efficacy between the groups, $F(1, 34) = 5.07$, $p < 0.05$, $\eta^2 = 0.13$. For interlocution-related self-efficacy, a significant effect was found, $F(1, 34) = 10.56$, $p < 0.01$, $\eta^2 = 0.237$. The SA group ($M = 3.81$, SD = 0.62) significantly outperformed the PA group ($M = 3.44$, SD = 0.75), demonstrating the considerable effect of SA on students' interlocution-related self-efficacy. Nevertheless, no significant differences were found between the groups in terms of accuracy ($F(1, 34) = 1.18$, $p > 0.05$,

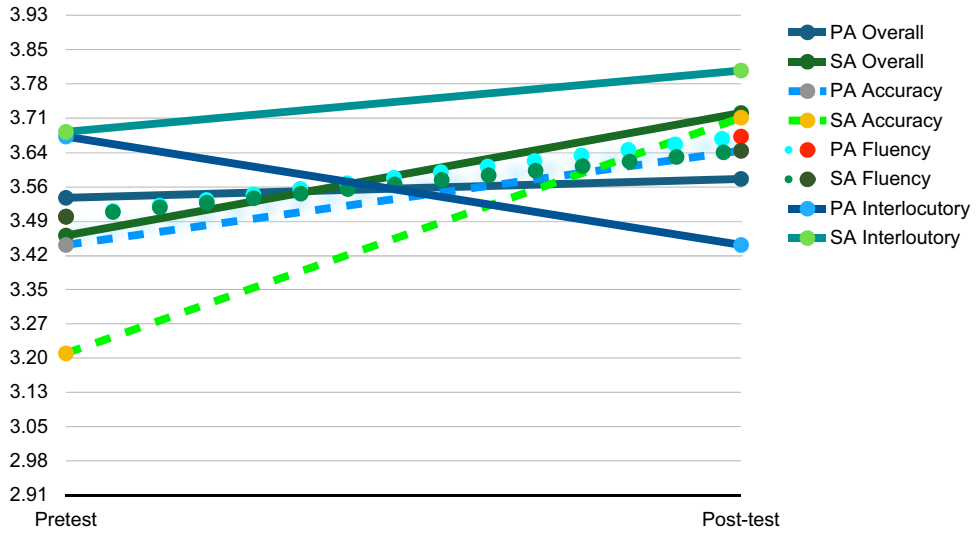

Fig. 2 Comparing the English-speaking self-efficacy of the PA and SA groups.

Table 3 ANCOVA results of English-speaking self-efficacy ($n = 37$).

English-speaking self-efficacy Overall	Peer Assessment Group (PA) ($n = 18$)				Self-Assessment Group (SA) ($n = 19$)						
	Pre-test		Post-test		Pre-test		Post-test		ANCOVA		
	M	SD	*M*	SD	*M*	SD	*M*	SD	*F* (1,34)	η^2	Effect size
	3.54	0.62	3.58	0.61	3.46	0.48	3.72	0.41	5.07*	0.13	-
Accuracy	3.44	0.54	3.64	0.60	3.21	0.54	3.71	0.25	1.18	0.03	-
Fluency	3.50	0.66	3.67	0.71	3.50	0.65	3.64	0.67	0.04	0.001	-
Interlocutory	3.67	0.84	3.44	0.75	3.68	0.47	3.81	0.62	10.56**	0.24	-

*$p < 0.05$, **$p < 0.01$.

$\eta^2 = 0.03$) or fluency-related self-efficacy ($F(1, 34) = 0.04$, $p > 0.05$, $\eta^2 = 0.001$), suggesting that the assessment method primarily influenced students' confidence in their interactive speaking abilities rather than their perceived technical language skills.

In short, the assessment type had a significant impact on students' English-speaking self-efficacy, with SA being more beneficial than PA. The assessment type did not affect students' accuracy-related or fluency-related speaking self-efficacy. VoiceThread-mediated SA had a greater positive effect on L2 learners' interlocutory self-efficacy than PA, highlighting its potential benefits and encouraging further exploration.

The SA and PA groups' perceptions of VoiceThread-mediated assessment.

- The questionnaire survey, open-ended survey, and group interview were carried out at the end of the project to gain an emic view of VoiceThread-mediated self- and peer assessment from the participants. The descriptive statistics of the questionnaire survey will be presented first, followed by the chi-square tests of different perceptions deducted from the qualitative analysis of the themes which emerged from the open-ended survey and group interview. The overall results showed that the SA and PA groups perceived VoiceThread-mediated assessment differently, despite some common retrospective thoughts after a year-long participation.
- The comparative analysis of the 19-item questionnaire survey (Appendix D) from the SA and PA groups highlights significant disparities in their perceptions of VoiceThread-mediated assessment. The SA group consistently demonstrated more positive perceptions across questionnaire items, with mean scores ranging from 3.28 to 4.33, compared to the PA group's lower range of 2.39–4.11. This overarching trend suggests that students who engaged in SA generally found their experience to be more beneficial. A notable point of agreement between both groups was their highest ranking of in-class speech practices (Item 1), with means of 4.33 and 4.11 for the SA and PA groups, respectively. This shared emphasis indicates that, regardless of the assessment approach, students valued the confidence-building aspects of in-class speaking activities. However, the groups diverged significantly in their other top-ranked items. The SA group particularly valued personal growth through self-review (mean = 4.28) and instructor feedback (mean = 4.06), while the PA group prioritized instructor feedback (mean = 3.78) and benefits from reviewing peer speeches (mean = 3.72).

  The SA and PA groups perceived learning outcomes differently. SA participants reported substantially higher benefits in oral fluency improvement (SA mean = 3.83 vs. PA mean = 2.78), grammatical accuracy and complexity (SA mean = 3.83 vs. PA mean = 2.78), and lexical diversity

(SA mean = 3.51 vs. PA mean = 2.72). This pattern suggests that self-reflection might be more effective than peer feedback in promoting perceived language development. The qualitative data revealed that SA participants engaged in systematic self-correction and targeted improvement, which might help explain the differently perceived language growth between the groups.

Motivation and engagement levels also differed significantly between the groups, particularly in related measures of self-efficacy. The SA group demonstrated higher motivation for creating video comments (mean = 3.44 vs. 2.89), greater willingness to continue with online tasks (mean = 3.42 vs. 2.83), and more comfort with online video posting (mean = 3.78 vs. 3.44). The qualitative data help explain this growth in self-efficacy. For instance, SA participants described a progression from discomfort to confidence through their self-review process. As one SA student noted, "I became less shy of oral tests," while another emphasized their growth journey: "I have grown more confident." In contrast, PA participants' comments revealed ongoing struggles with self-efficacy: "I'm not brave enough to make mistakes in public even though I'm still in the learning process. But I think I made great progress with it". These differences in affective aspiration help explain why SA participants reported higher comfort levels with video tasks.

- A chi-square test of independence was conducted to examine potential differences in the SA and PA groups' perceived benefits. The analysis shows a statistically significant association between the two groups and the perceived benefits ($\chi^2 = 9.847$, df = 3, $p = 0.020$). Table 4 presents the two most significant differences of the four analysed.

The most notable difference emerged in the advantage of performance review, where the SA group exhibited a substantially higher frequency (52.0%, adjusted residual = 3.1) compared to the PA group (25.8%, adjusted residual = −3.1). For instance, one female student in the SA group commented, "If I don't see my video, I won't find out that my pronunciation is weird. My gestures are also weird in the video, which helped me see I have a lot to improve while giving a speech". Another SA participant said, "I often write down that I failed to express my idea so that I can fix it and say it right next time". These specific, detailed approaches to self-review help explain why SA participants reported higher motivation scores for video tasks (mean = 3.44 vs. 2.89).

While both groups maintained relatively balanced distributions across observation opportunity (self: 28.0%, peer: 32.3%), their observation patterns revealed different learning approaches. The PA group's comments showed more general observations, as exemplified by one student: "I saw some problems in other people's speeches, and I try not to make the same mistakes in my speech. That's a benefit to me. Also, I can learn from others' public speaking skills while watching their speeches". The PA group valued emotional support through peer feedback, as evidenced by one male student's comment: "I always feel too nervous to speak at a normal speed, but one time I saw a comment from my teammate; she said I was so brave to challenge myself to do an impromptu speech. I was inspired a lot". This focus on general observation and emotional support, rather than specific language improvements, helps explain why PA participants reported lower scores in specific language skills such as grammatical accuracy (mean = 2.78) and lexical diversity (mean = 2.72).

- A chi-square test of independence revealed significant differences in the types of difficulties experienced by the SA and PA groups ($\chi^2 = 19.7847$, $df = 4$, $p < 0.001$). Table 5 displays the two most significant differences out of the five analysed. The SA group reported more UI/Technical problems (35.7%, SR = 1.7) than the PA group did (11.5%, SR = −1.7), with SA students specifically noting issues such as, "I find the machine-generated subtitles inaccurate. It is not helpful for self-correction", and interface challenges where "adding the self-correcting comments in my recorded speech is time-consuming". Despite these interface challenges, the SA group was more willing to continue with online tasks (mean = 3.42 vs. 2.83), suggesting these issues did not significantly impact their engagement.

Table 4 Categories and sub-categories of students' perceived benefits of VoiceThread-mediated self- and peer assessment.

Category	Subcategory	Frequency	
		Self-Assessment	Peer Assessment
Performance Review	Ability to review/Self-review	26 (52.0, 3.1)*	8 (25.8, −3.1)*
Observation Opportunity	Self-observation/Receiving feedback	14 (28.0, 0.7)	10 (32.3, −0.7)

Note: Numbers in parentheses show (percentage, adjusted residual). *$p < 0.05$.

Table 5 Categories and sub-categories of students' perceived difficulties of VoiceThread-mediated self- and peer assessment.

Category	Subcategory	Frequency	
		Self-Assessment	Peer Assessment
UI/Technical Problems	Interface and technical issues	10 (35.7, 1.7)	3 (11.5, −1.7)
Technical & Connection Issues	Technical difficulties and connectivity	0 (0, −2.8)*	11 (42.3, 2.8)*

Note: Numbers in parentheses show (percentage, adjusted residual). *$p < 0.05$.

However, technical and connection issues were reported substantially more frequently in the PA group (42.3%, SR = 2.8) compared to the SA group (0%, SR = −2.8), with one PA student reporting that "The slow speed of the internet prevented VoiceThread from showing the voice and the picture simultaneously." These synchronization problems help explain the PA group's lower comfort with online video posting (mean = 3.44 vs. SA's 3.78) and reduced motivation for creating video comments (mean = 2.89 vs. SA's 3.44). Despite these shared challenges, the SA group's higher engagement scores suggest that their more structured self-review helped them overcome these technological hindrances more effectively than the PA group.

These findings indicate that while both assessment modes faced some common challenges, they each encountered distinct technical and user-interface challenges that may need to be addressed differently for successful implementation. The participants' comments highlighted the need for flexible, user-friendly solutions tailored to each assessment mode.

Discussion

This study investigated how VoiceThread-mediated SA and PA influence L2 learners' oral proficiency and English-speaking self-efficacy. The findings reveal that while both assessment modalities enhanced speaking skills, SA demonstrated superior effectiveness in terms of developing oral accuracy, complexity, and interlocutory self-efficacy. These results extend our understanding of technology-mediated assessment in second language acquisition by illuminating the differential impacts of assessment approaches on specific components of language development and learner self-perception. The following discussion examines three major findings in relation to previous research: the effects of online assessment types on L2 oral proficiency, their distinct impacts on English-speaking self-efficacy, and the participants' perceptions of their year-long engagement in VoiceThread-mediated SA and PA.

Effects of online assessment types on L2 oral proficiency. The study's findings support the consensus in the literature that both self- and peer assessment can positively impact learners' speaking skills. The results also suggest that SA may be more effective in certain aspects of oral proficiency, particularly accuracy and complexity. Nevertheless, these findings should be interpreted within their specific context. The effectiveness of assessment types may vary substantially based on contextual factors, such as learners' language proficiency, cultural background, digital literacy, and classroom dynamics.

The study's results indicate that the SA and PA groups improved overall oral proficiency after a year-long engagement with VoiceThread-mediated assessments. The SA group showed significantly higher post-test scores ($M = 17.50$, SD = 0.50) than the PA group ($M = 16.42$, SD = 1.05). This finding aligns with recent research by Maleki et al. (2024), which found SA to be more effective than PA in terms of improving ESP students' speaking ability. The more substantial impact of SA on oral proficiency outcomes also resonates with earlier studies by Chen (2010) and Alibakhshi and Sarani (2014), which emphasized the role of SA in promoting metacognitive skills and self-directed learning. However, this difference should be considered alongside important contextual factors: the participants were predominantly intermediate-level learners from a cultural background that traditionally emphasizes individual achievement, and the study was conducted in a specific technological environment that may have differentially impacted SA and PA implementations. For example, the predominantly intermediate-level learners in this study may have had sufficient baseline proficiency to engage in meaningful self-evaluation, aligning with Alibakhshi and Sarani's (2014) findings on SA's effectiveness for accuracy and fluency development. In contrast, peer assessment might prove more beneficial in cultural contexts emphasizing collective learning and social interaction, as Esfandiari and Tavassoli's (2019) research showed PA's positive impact on speaking skills development.

In addition, the superior performance of the SA group in accuracy and complexity suggests that online SA may enhance Zimmerman's (2000) cyclical phases of self-regulated learning (planning, performance, and self-reflection). The asynchronous nature of VoiceThread strengthened the self-reflection phase by allowing students to review their performance multiple times and make detailed observations about their speaking progress. This finding extends Zimmerman's framework by demonstrating how digital platforms can create enhanced opportunities for metacognitive monitoring and control in language learning contexts.

Similar to previous studies (Hammad Al-Rashidi et al., 2023; Imani, 2021), the differential effects on specific language skills (stronger impact on accuracy and complexity but not on fluency or pronunciation) suggest that technology-mediated SA may operate through distinct metacognitive pathways. While traditional self-regulated learning theory emphasizes the importance of goal-setting and strategic planning, the current findings indicate that digital platforms may enhance the 'performance control' phase of self-regulation through their capacity for precise, repeatable self-observation. This theoretical extension helps explain why accuracy and complexity, which benefit from careful analysis and conscious monitoring, showed more remarkable improvement than fluency and pronunciation, which rely more on automaticity and implicit learning.

Effects of online assessment types on English-speaking self-efficacy. The second research question addressed the differential effects of VoiceThread-mediated SA versus PA on English-speaking self-efficacy. The ANCOVA findings revealed that SA had a more positive impact on overall speaking self-efficacy than peer assessment, particularly in developing interlocutory self-efficacy. These results align with Bandura's (1997) self-efficacy theory and echo previous studies (Aldosari and Alsager, 2023; Alibakhshi and Sarani, 2014), emphasizing the crucial role of mastery experiences in developing self-efficacy beliefs. VoiceThread-mediated SA created a reinforcing cycle between mastery experiences and self-efficacy. Students who evaluated their recorded speeches with narrated self-critiques gained concrete evidence of their progress in making, enhancing their self-efficacy. This heightened efficacy propelled greater effort and persistence, improving performance. The SA group showed higher overall self-efficacy scores ($M = 3.72$) than the PA group ($M = 3.58$), with a notably strong effect on interlocutory self-efficacy ($\eta^2 = 0.24$), suggesting that SA provided more authentic mastery experiences than peer feedback.

The discrepancy between SA and PA might be attributed to the online medium used in the current study, as VoiceThread provides a structured environment for self-reflection that may amplify the benefits of SA. The most unexpected finding was the significant positive effect of SA on interlocutory self-efficacy ($\eta^2 = 0.24$), while PA saw a decline in this dimension. The findings challenge traditional interpretations of Vygotsky's sociocultural theory, which emphasizes the primacy of social interaction in learning. While PA theoretically provides more opportunities for social learning, the results suggest that technology-mediated SA may create a new form of "internal scaffolding" that effectively supports language development. The VoiceThread platform appeared to transform SA from a solitary activity into a form of mediated social interaction, where learners engage in dialogue with their recorded selves. This finding extends sociocultural theory by suggesting that digital tools can create innovative forms of scaffolding that complement traditional social interaction. This outcome suggests that SA may be particularly beneficial for developing confidence in interactive speaking situations. The decline in interlocutory self-efficacy in the PA group might reflect what Singh (2015) and Yinjaroen and Chiramanee (2011) identified as challenges in PA, such as assessor bias or evaluation inconsistency.

The more substantial overall impact of SA aligns with recent research demonstrating that self-efficacy beliefs are malleable rather than fixed traits (Al-khresheh and Alkursheh, 2024; Kim, 2024). The VoiceThread platform facilitated this development through several key mechanisms. First, its asynchronous nature allowed students to set specific speaking goals, record multiple attempts, and track their progress over time, which are the features that Basileo et al. (2024) identified as crucial for self-efficacy development. Second, the platform's multimedia

capabilities enabled students to hear and see their performances, which provided concrete evidence of their speaking abilities and progress. This aligns with Wang and Sun's (2024) finding that technology-enhanced environments can strengthen self-efficacy by offering immediate, tangible feedback and opportunities for self-reflection. Resonating this technological benefit, Komang (2021) found that online formative assessment was particularly effective in terms of promoting English language learners' motivation, self-efficacy, and self-regulated learning. The combination of goal-setting features and technological affordances may have created an ideal environment for students to recognize their improving capabilities, thus strengthening their self-efficacy.

In sum, the more substantial effect of SA on interlocutory self-efficacy ($\eta^2 = 0.237$) presents an interesting theoretical paradox: students developed greater confidence in interactive speaking through predominantly individual practice. This finding suggests a need to revisit theoretical assumptions about the relationship between social interaction and self-efficacy development in online environments. The asynchronous, multimodal nature of VoiceThread may create a unique form of "temporal scaffolding" from which learners can gradually build confidence through repeated self-observation and reflection before engaging in live interaction.

Participants' perceptions of the year-long engagement in VoiceThread-mediated SA and PA. Analysis of the qualitative data revealed three themes in the participants' perceptions of VoiceThread-mediated self- and peer assessment: perceived impact on speaking development and autonomous learning, assessment mode preferences and self-efficacy beliefs, and technical experiences affecting engagement.

Participants in the SA group predominantly perceived the assessment experience as facilitating autonomous learning and speaking improvement. Their responses frequently highlighted the value of being able to 'review and improve speech' (11 responses) and 'see oral presentations' (9 responses), indicating positive perceptions of SA's role in supporting self-directed learning. These perceived benefits aligned with their significantly higher engagement in performance review activities (52.0%, adjusted residual = 3.1) compared to PA participants (25.8%, adjusted residual = −3.1). The SA participants' positive perceptions of autonomous learning opportunities support Alibakhshi and Sarani's (2014) findings on SA benefits and extend Maleki et al. (2024) research by illuminating why learners might find SA more effective than PA for speaking development.

Learners' perceptions of assessment modes revealed distinct patterns in their self-efficacy development. SA participants reported more diverse perceived benefits (12 unique categories versus 11 for PA) and emphasized autonomous learning experiences through themes such as 'I can correct myself' and 'confidence cultivation'. In contrast, PA participants' perceptions were notably influenced by technical challenges, with a significant portion reporting connection issues (42.3%, SR = 2.8) compared to no such reports from SA participants (0%, SR = −2.8). These negative technical experiences appeared to affect their confidence in interactive speaking contexts, as reflected in their declining interlocutory self-efficacy scores (from $M = 3.67$–3.44). These perceptual differences challenge Zheng et al. (2023) recommendations regarding PA for low-anxiety learners, and suggest that the structured online environment of VoiceThread may better support self-efficacy development through SA, aligning with Wang and Sun's (2024) observations about technology-enhanced learning environments.

Learners' perceptions of technical aspects showed different challenges between assessment modes that influenced participants' overall learning experiences. While both groups reported similar video accessibility concerns (SA: 42.9%, PA: 34.6%), their experiences with other technical aspects differed markedly. SA participants reported more UI/Technical problems (35.7%, SR = 1.7) but fewer connection issues, whereas PA participants perceived connection difficulties as a major barrier (42.3%, SR = 2.8). These differential perceptions of technical challenges help explain why SA participants reported more positive experiences with accuracy and complexity development, suggesting they felt more capable of overcoming technical barriers through self-directed practice. These findings extend Wibowo and Novitasari's (2021) research on technical challenges in online assessment by highlighting how learners' perceptions of technical issues can vary based on assessment mode and influence their engagement patterns, while supporting Liao's (2023) observations about the potential of online platforms to facilitate self-directed learning when learners perceive technical obstacles as being manageable.

Conclusion

The findings support and extend both Bandura's self-efficacy theory and Zimmerman's (2000) cyclical phases of self-regulated learning in L2 contexts. The SA group's more substantial speaking proficiency and self-efficacy gains align with Bandura's emphasis on mastery experiences. The superior performance of the SA group in accuracy and complexity extends Zimmerman's framework by demonstrating how digital platforms can enhance opportunities for metacognitive monitoring and control. However, the findings diverge from Vygotsky's sociocultural theory, which emphasizes social interaction in learning. The unexpected positive effect of SA on interlocutory self-efficacy, contrasted with PA's decline, reveals how technology-mediated SA creates a novel form of "internal scaffolding" through VoiceThread's structured environment. The platform transforms SA from a solitary activity into mediated social interaction as learners engage in dialogue with their recorded selves. It extends sociocultural theory by demonstrating how digital tools can innovate traditional scaffolding processes. At the same time, the qualitative data elucidate specific mechanisms through which these experiences are facilitated by technology-mediated self-assessment. The emergence of technical barriers as a significant factor in assessment effectiveness suggests that Bandura's concept of environmental factors in self-efficacy development should be expanded to consider the role of digital learning environments. Furthermore, the SA group demonstrated greater resilience in circumventing technical barriers through self-directed practice, suggesting that SA may be more robust against technological disruption in online environments.

The study's methodological limitations point to several crucial directions for future research in technology-mediated language assessment. From a methodological perspective, the study's reliance on a single online platform (VoiceThread) may limit the generalizability to other technology-mediated assessment contexts. Specifically, while VoiceThread offers multimodal commenting options, its asynchronous nature may have influenced the patterns of peer interaction compared to what might be possible with synchronous assessment tools. The platform's specific features and interface design shaped how students chose to engage with peer feedback, and how they utilized different modes of communication. The small sample size and the specific context further constrain the study's generalizability to other language learner populations. A notable limitation is the absence of a control group that performed speaking tasks without assessment interventions, making it difficult to isolate the effects of assessment from potential improvements through task repetition alone. Technical limitations also warrant consideration. Technical and connection issues, particularly in the PA group, may have confounded PA's potential effectiveness. Considering these

limitations, future research in the following directions is recommended. First, future studies should include control groups that perform speaking tasks without assessment interventions to isolate the unique effects of assessment practices from task repetition. Second, they should investigate assessment effectiveness across different technology platforms to establish platform-independent principles for online language assessment. Third, future investigations should explore how systematic peer assessor training might reduce the observed performance gap through self-assessment. Fourth, individual learner factors warrant further examination as the relationship among anxiety levels, self-efficacy, and assessment preferences might substantially affect assessment effectiveness. Finally, developing hybrid assessment models presents a promising avenue for future research, including exploring how AI-supported assessment can complement human assessment.

The findings have significant implications for pedagogical practice and policy development in technology-mediated language learning. Course designers should choose the assessment mode based on specific language skill targets, particularly considering SA's superior effectiveness in terms of developing accuracy and complexity. While both assessment modes faced common challenges in video accessibility, they each encountered distinct technical and usability challenges that required different implementation strategies. This suggests the need for flexible, user-friendly solutions tailored to each assessment mode. Given the lack of significant differences in fluency and pronunciation outcomes, integrating multiple assessment types might be necessary to develop oral proficiency holistically. While the PA group showed mediocre language gains, PA demonstrated notable strengths in fostering critical thinking skills and authentic communication as students engaged in evaluating others' work. The social interaction inherent in PA plays an irreplaceable role in language learning, as language is fundamentally a tool for communication. This was evidenced by students' positive responses to peer feedback, with one participant noting how encouragement from group members boosted his confidence—highlighting peer assessment's unique capacity for social persuasion in building self-efficacy. These findings have important implications for Taiwan's 2030 Bilingual Nation Policy, particularly in teacher preparation programs and curriculum development. Pre-service teachers will need specific training in utilizing self- and peer assessment in bilingual contexts, emphasizing creating supportive digital learning environments that accommodate diverse learner profiles. The challenge remains to create integrated assessment approaches that can harness the demonstrated benefits of SA while preserving the valuable social learning opportunities offered by PA, all within technically robust and pedagogically sound online learning environments.

Data Availability

The datasets generated during and/or analyzed during the current study are available from the corresponding author on reasonable request.

Received: 4 January 2025; Accepted: 29 July 2025;
Published online: 07 August 2025

Note

1 Alternative assessment broadly refers to non-traditional, learner-centered methods such as portfolios, self-assessments, and peer evaluations focusing on authentic tasks and process-oriented feedback. These methods aim to foster authentic language use, reduce anxiety, and improve speaking accuracy and fluency compared to traditional assessments (Phongsirikul, 2018).

References

Aben J, Timmermans A, Dingyloudi F, Mascareño Lara M, Strijbos JW (2022) What influences students' peer-feedback uptake? Relations between error tolerance, feedback tolerance, writing self-efficacy, perceived language skills and peer-feedback processing. Learn Individ Differ 97:102175. https://doi.org/10.1016/j.lindif.2022.102175

Alavi S, Sadighi F, Samani S (2004) Developing a foreign language learning self-efficacy scale for Iranian students. Soc Sci Humanit Shiraz Univ 21(1):94–101

Al-khresheh MH, Alkursheh TO (2024) An integrated model exploring the relationship between self-efficacy, technology integration via Blackboard, English proficiency, and Saudi EFL students' academic achievement. Humanit Soc Sci Commun 11:287. https://doi.org/10.1057/s41599-024-02783-2

Aldosari MS, Alsager HN (2023) A step toward autonomy in education: probing into the effects of practicing self-assessment, resilience, and creativity in task supported language learning. BMC Psychol 11:434. https://doi.org/10.1186/s40359-023-01478-8

Alibakhshi G, Sarani A (2014) Self-assessment impact on EFL learners' speaking fluency and accuracy. TELL 8(2):119–143

Al-Rashidi A, Vadivel B, Khalil N, Basim N (2023) The comparative impacts of portfolio-based assessment, self-assessment, and scafolded peer assessment on reading comprehension, vocabulary learning, and grammatical accuracy: insights from working memory capacity. Lang Test Asia 13(1):24. https://doi.org/10.1186/s40468-023-00237-1

Amalia R (2018) Students' perception of online assessment use in Schoology in EFL classrooms. Dissertation, Sunan Ampel State Islamic University

Ashraf H, Mahdinezhad M (2015) The role of peer-assessment versus self-assessment in prompting autonomy in language use: a case of EFL learners. Iran J Lang Test 5(2):110–120

Bandura A (1986) The explanatory and predictive scope of self-efficacy theory. J Clin Soc Psychol 4:359–373

Bandura A (1997) Self-efficacy: the exercise of control. W. H. Freema, New York

Bandura A, Schunk DH (1981) Cultivating competence, self-efficacy, and intrinsic interest through proximal self-motivation. J Pers Soc Psychol 41(3):586–598. https://doi.org/10.1037/0022-3514.41.3.586

Basileo LD, Otto B, Lyons M, Vannini N, Toth MD (2024) The role of self-efficacy, motivation, and perceived support of students' basic psychological needs in academic achievement. Front Educ 9:1385442. https://doi.org/10.3389/feduc.2024.1385442

Çetin Köroğlu Z (2021) Using digital formative assessment to evaluate EFL learners' English speaking skills. GIST Educ Learn Res J 22:103–123. https://doi.org/10.26817/16925777.1001

Chen CH (2010) The implementation and evaluation of a mobile self- and peer-assessment system. Comput Educ 55(1):229–236

Chen F, Kao SM, Tsou W (2020) Toward ELF-informed bilingual education in Taiwan: addressing incongruity between policy and practice. Engl Teach Learn 44(1):175–191

Chiang MH (2018) Effects of self and peer assessment on learner autonomy and speaking proficiency among adult EFL learners. In: Paper presented at 7th International Conference on Teaching, Education & Learning (ICTEL), Singapore

Crisp G (2011) Teacher's handbook on e-assessment: a handbook to support teachers in using e-assessment to improve and evidence student learning and outcomes. Creative Commons, San Francisco

Crystal D (2003) English as a global language. Cambridge University Press

Esfandiari S, Tavassoli K (2019) The comparative effect of self-assessment vs. peer-assessment on young EFL learners' performance on selective and productive reading tasks. Iran J Appl Linguist 22(2):1–35

Falchikov N (2007) The place of peers in learning and assessment. Rethinking Assessment in Higher Education 128–144

Faul F, Erdfelder E, Buchner A, Lang AG (2009) Statistical power analyses using G*Power 3.1: tests for correlation and regression analyses. Behav Res Methods 41(4):1149–1160

Hammad Al-Rashidi A, Vadivel B, Ramadan Khalil N et al. (2023) The comparative impacts of portfolio-based assessment, self-assessment, and scaffolded peer assessment on reading comprehension, vocabulary learning, and grammatical accuracy. Lang Test Asia 13:24. https://doi.org/10.1186/s40468-023-00237-1

Hoesny MU, Setyosari P, Praherdhiono H, Suryati N (2023) The correlation of speaking self-efficacy, speaking proficiency and gender in ESP context. Pegem J Educ Instr 13(2):191–199

Imani S (2021) The comparative effect of self-assessment and peer assessment on reflective and impulsive EFL learners' speaking skill. J Foreign Lang Teach Transl Stud 6(4):99–120. https://doi.org/10.22034/efl.2022.325985.1139

Jamshidnejad A (2020) Introduction: challenges of L2 oral communication in EFL contexts. In: Jamshidnejad A (ed) Speaking English as a Second Language. Palgrave Macmillan, Cham. https://doi.org/10.1007/978-3-030-55057-8_1

Joo SH (2016) Self and peer assessment of speaking. ALTESOL 16(2):68–83

Kim H (2024) Exploring the interplay of language mindsets, self-efficacy, engagement, and perceived proficiency in L2 learning. Humanit Soc Sci Commun 11:1295. https://doi.org/10.1057/s41599-024-03783-y
Komang LK (2021) The implementation of online formative assessment in English learning. J Educ Study 1(1):76–84
Kumar T, Soozandehfar SMA, Hashemifardnia A et al. (2023) Self vs. peer assessment activities in EFL-speaking classes: impacts on students' self-regulated learning, critical thinking, and problem-solving skills. Lang Test Asia 13:36. https://doi.org/10.1186/s40468-023-00251-3
Li C, Zhang LJ (2023) The development of accuracy and fluency in second language (L2) speaking related to self-efficacy through online scaffolding: a latent growth curve modeling analysis. J Psycholinguist Res 52:1371–1395. https://doi.org/10.1007/s10936-023-09950-7
Li J, Wang C, Zhao Y, Li Y (2024) Boosting learners' confidence in learning English: can self-efficacy-based intervention make a difference? TESOL J 58:1518–1547. https://doi.org/10.1002/tesq.3292
Liao MH (2023) Enhancing L2 English speaking and learner autonomy via online self- and peer-assessment. Taiwan J TESOL 20(1):33–66. https://doi.org/10.30397/TJTESOL.202304_20(1).0002
Maleki R, Zivari Naser S, Ghaffari S (2024) The comparative effect of self-assessment and peer-assessment on speaking ability and anxiety of students of psychology. Teach Learn Engl Spec Purp 1(2):15–34. https://doi.org/10.22034/tlesp.2024.476864.1021
McNatt DB (2019) Enhancing public speaking confidence, skills, and performance: an experiment of service-learning. Int J Manag Educ 17(2):276–285
Pagkalinawan L (2021) Oral presentation on Voicethread: a collaborative assessment strategy in enhancing language proficiency and oral communication practices. Adv. Soc Sci Educ Humanit Res 533:267–274
Patri M (2002) The influence of peer feedback on self- and peer-assessment of oral skills. Lang Test 19(2):109–131
Phan HP, Ngu BH (2016) Sources of self-efficacy in academic contexts: a longitudinal perspective. Sch Psychol Q 31(4):548–564. https://doi.org/10.1037/spq0000151
Phongsirikul M (2018) Traditional and alternative assessments in ELT: students' and teachers' perceptions. rEFLections 25(1):61–84
Ritonga M, Kustati M, Budiarti M et al. (2021) Arabic as foreign language learning in pandemic COVID-19 as perceived by students and teachers. Linguist Cult Rev 5(1):75–92
Salehi N, Rowan M (2015) Using VoiceThread to enhance learning. Minnesota eLearning Summit 1–12
Schunk DH, Usher EL (2019) Social cognitive theory and motivation. In: Ryan RM (ed) The Oxford handbook of human motivation. Oxford University Press. https://doi.org/10.1093/oxfordhb/9780190666453.013.2
Sengul BG, Buyukkarci K (2023) A correlation study on EFL students' self-efficacy and attitudes towards English and foreign language speaking anxiety. Innov Res ELT 4(1):32–46. https://doi.org/10.29329/irelt.2023.558.3
Singh S (2015) Self-assessment of oral proficiency among ESL learners. ELT Voice India 5(1):1–7
Usher EL, Pajares F (2008) Self-efficacy for self-regulated learning: a validation study. Educ Psychol Meas 68(3):443–463. https://doi.org/10.1177/0013164407308475
Vygotsky LS (1978) Mind in society: development of higher psychological processes. Cole M, John-Steiner V, Scribner S, Souberman E (eds). Harvard University Press. https://doi.org/10.2307/j.ctvjf9vz4
Wang Q, Li S (2019) The relationship between task motivation and L2 motivation: An empirical study. In: Wen Z, Ahmadian M (eds) Researching L2 task performance and pedagogy: in honour of Peter Skehan. John Benjamins, Amsterdam, p 67–92
Wang Y, Sun PP (2024) Development and validation of scales for speaking self-efficacy: constructs, sources, and relations. PLoS ONE 19(1):e0297517. https://doi.org/10.1371/journal.pone.0297517
Wibowo FE, Novitasari U (2021) An analysis of online assessment in teaching English. PROJECT 4(3):521–529. https://doi.org/10.22460/project.v4i3.p521-529
Wicaksono BH, Ismail SM, Sultanova SA et al. (2023) I like language assessment: EFL learners' voices about self-assessment, self-efficacy, grit tendencies, academic resilience, and academic demotivation in online instruction. Lang Test Asia 13(37):1–18. https://doi.org/10.1186/s40468-023-00252-2
Wijaya KF (2024) The impacts of self-efficacy on EFL learners' speaking skills. JELITA J Educ Lang Innov Appl Linguist 3(2):121–133. https://doi.org/10.37058/jelita.v3i2.6878
Yinjaroen P, Chiramanee T (2011) Peer assessment of oral English proficiency. In: Paqper presented at third international conference on humanities and social sciences, Hat Yai, Thailand
Zarei AA, Usefli Z (2019) Types of assessment affecting Iranian EFL learners' general and academic self-efficacy. Indones J Learn Instr 2(2):55–66
Zheng C, Wang L, Chai CS (2023) Self-assessment first or peer-assessment first: effects of video-based formative practice on learners' English public speaking anxiety and performance. Comput Assist Lang Learn 36(4):806–839. https://doi.org/10.1080/09588221.2021.1946562
Zheng C, Chen X, Zhang H, Chai CS (2024) Automated versus peer assessment: effects on learners' English public speaking. Lang Learn Technol 28(2):210–228. https://hdl.handle.net/10125/73577
Zimmerman BJ (2000) Attaining self-regulation: a social cognitive perspective. In: Boekaerts M, Pintrich PR, Zeidner M (eds) Handbook of self-regulation. Academic Press, pp 13–39. https://doi.org/10.1016/B978-012109890-2/50031-7

Acknowledgements

This research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors.

Author contributions

The author, Min-Hsun Liao, planned and designed the study, conducted the research, analyzed the data, and composed the manuscript.

Competing interests

The authors declare no competing interests.

Ethical approval

This study was approved by the Research Ethics Committee at the National Changhua University of Education. It was approved on July 11, 2018, with an approval number NCUEREC-106-024. All research was performed in accordance with the relevant guidelines and regulations, including the Declaration of Helsinki.

Informed consent

Prior to participation, all subjects received a detailed participant information sheet outlining the study objectives, procedures, potential risks, and benefits. They were explicitly informed that participation was voluntary and that they could withdraw at any time without consequences. Participants were assured that all data collected would remain confidential and anonymized, with no identifying information disclosed in research outputs. Written informed consent was obtained during the week of September 10–14, 2018, from all participants before their inclusion in the study, and each participant received a copy of their signed consent form for personal records.

AI tool declaration

This paper used Claude 3 for manuscript proofreading and reference style conversion. All AI-generated content underwent rigorous human verification and validation. The author is solely responsible for the findings and conclusions presented.

Additional information

Supplementary information The online version contains supplementary material available at https://doi.org/10.1057/s41599-025-05674-2.

Correspondence and requests for materials should be addressed to Min-Hsun Liao.