SCRIPT JOURNAL
Journal of Linguistic and English Teaching
http://jurnal.fkip-uwgm.ac.id/index.php/Script
P-ISSN: 2477-1880; E-ISSN: 2502-6623
August 2026, Vol. 11 No. 2

Speaking behind Avatars: EFL Students' Experiences in AI-Assisted Lip-Sync Video Production for Monologue Speaking Practice

Nadrotin Mawaddah[1], Nuril Firdaus[2], Rosa Umami Rahman[3]
Universitas Ibrahimy, Indonesia[123]
Email Correspondence: nadrotinmawaddah@gmail.com

Abstract

Background:
Many EFL students at junior high schools continue to experience speaking anxiety and low self-confidence, including when performing basic monologue speaking. While multimodal AI has increasingly been incorporated into EFL speaking instruction, its use in monologue practice is underexplored. In this context, employing AI Avatar Lip-sync tools may provide an engaging and comfortable speaking experience. Therefore, this study investigated learners' affective, engagement, and technological experiences during the learning process.

Methodology:
This convergent mixed-methods study involved 28 eighth-grade students. They were instructed to create AI-generated avatars and record monologues, which were automatically transformed into lip-synchronized animation videos using Pippit AI. Data on their detailed experiences were collected through classroom observations and a questionnaire, then analyzed using thematic analysis and descriptive statistics. Both datasets were then integrated by examining areas of convergence, complementarity, and divergence, thereby providing a more comprehensive understanding of students' learning experiences.

Findings:
Quantitative findings revealed high levels of affective (M = 3.95), engagement (M = 4.09), and technological (M = 3.79) experiences. Qualitative findings showed that students enjoyed creating their avatars and voluntarily repeated speaking practice. Additionally, they reported that speaking through avatars made them more relaxed because their faces were not exposed. Despite several technical challenges, students remained enthusiastic throughout the activity and expressed high satisfaction with the final videos.

Conclusion:
The findings suggest that monolog speaking practice using AI avatar lip-sync has the potential to provide a positive experience across affective, engagement, and technological dimensions. Students reported high enjoyment and self-monitoring through repeated practice, while feeling more relaxed and comfortable when speaking without face exposure. Despite several technical challenges, students expressed high satisfaction with the outcome videos. These findings highlight the potential of AI avatar lip-sync automation in supporting EFL speaking practices.

Originality:
Most existing studies on AI-assisted speaking predominantly focus on real-time, interactive communication skills, while the combination of recorded monologue practice and AI lip-sync avatar production in EFL junior high school contexts was underexplored. Accordingly, this study addressed the gap by examining students' experiences with AI lip-sync avatar production for recorded monologue practice.

Keywords	:	*EFL monologue speaking, students' experiences, AI lip-sync video, avatar.*
DOI	:	10.24903/sj.v11i2.2403
Received	:	May 2026
Accepted	:	August 2026
Published	:	August 2026
How to cite this article (APA)	:	Mawaddah, N., Firdaus, N., & Rahman, R. U. (2026). Speaking behind Avatars: EFL Students' Experiences in AI-Assisted Lip-Sync Video Production for Monologue Speaking Practice. *Script Journal: Journal of Linguistics and English Teaching*, 11(2), 536-559.

		https://doi.org/10.24903/sj.v11i2.2403

1. INTRODUCTION

Since the emergence of multimodal AI, research has pursued this advancement. In the EFL learning context, AI applications have been developed to serve as speaking partners, even those specifically designed for independent speaking practice, with automated feedback. *Thus,* research in this area investigates the impact of such AI on speaking performance (Ebadi et al., 2025; Shi & Shakibaei, 2025; Xing & Saeed, 2025) and affective aspects such as anxiety, self-confidence, and willingness to speak (Aldosari, 2024; F. Huang & Zou, 2024; Zou et al., 2025). These research trends continue to develop alongside the growing innovation in multimodal AI.

Previous studies on AI for EFL speaking have mostly focused on AI for interactive speaking (dialogue), with limited research on AI for monologue practice. It is noteworthy since monologues are also necessary in speaking instruction, as they offer the opportunity to produce speech independently and practice speaking fluency (S. E. Huang & Liu, 2022). For early-stage learners, monologic speaking activities can be performed at a slower speaking pace, which is often associated with lower-proficient students (Ding & Lian, 2020). Research on task interactivity in second language (L2) learning indicates that monologue and dialogue speaking activities entail different task demands; dialogue tasks create a more interactive environment, whereas monologue tasks require more independent speech production (Bui, 2026). Therefore, since dialogues and monologues serve different pedagogical functions and competencies, EFL students need opportunities to practice their speaking skills through both activities.

Although monologues do not require real-time interaction, this does not mean that speaking affective barriers are reduced. Instead, according to Szyszka et al. (2026), monologues impose a complex cognitive and affective burden because students are required to plan the flow, grammar, and articulation independently within a limited timeframe, without the pauses provided by a conversation partner. This concern aligns with a preliminary investigation conducted by the researchers one month before the study at the participating schools. The researchers investigated through classroom observations and discussion with the English teacher in the target class. The results indicated that the majority of students were reluctant to

speak monologic English in front of the class, with the teacher attributing this reluctance to students' fear of making mistakes and negative evaluation. According to Krashen (1985), this condition is associated with emotional barriers that can block the processing of language input. Furthermore, it leads to speaking anxiety (Lee & Chiu, 2023). To address these constraints, EFL speaking instruction should be designed to be less pressuring and psychologically comfortable.

However, providing a sense of less anxiety is inadequate to ensure active participation in the speaking class (Philp & Duchesne, 2016). Students can feel psychologically safe but remain passive due to a lack of interest, low motivation, or simply because they perceive the task as uninteresting. Consequently, speaking activities should be designed to be comfortable and engaging (Almukhaild & King, 2024; Zhong et al., 2025), and capable of fostering immersive participation (Heutte et al., 2021). Therefore, alongside affective factors, engagement is a crucial aspect that deserves more deliberate attention.

Accordingly, one AI application that can potentially realize those aspects in monologue speaking practice is AI lip-sync or "talking head" applications. This AI can automatically transform static images into animated videos by synchronizing lip and facial movements with a speech audio track in a near-realistic manner (Ji et al., 2024; L. Liu et al., 2024). In this current research, students were assigned to create their digital animated character images, also defined as "Avatar" (Szmigiel & Lee, 2016). They also recorded their own voices in speaking monologues. Both avatar images and recorded audio are uploaded to an AI lip-sync website (Pippit AI), which automatically generates animated lip-sync videos.

The instructional strategy promotes a relatively novel pedagogical practice. Based on previous studies, representing oneself through a digital avatar may provide a more comfortable context for speaking. In public-speaking contexts, AR face filters have been reported to help speakers feel less personally exposed (Gong & Kanai, 2025; Leong et al., 2023). Similarly, it has been associated with self-disclosure when the avatar differs from the user's physical appearance and original human facial cues (Ichino et al., 2022; Yoneyama et al., 2024). In EFL contexts, Hirata (2023) also reported that this way could support speaking confidence among learners with low speaking confidence. Overall, the findings of these previous studies may be analogously understood as resembling a "masking" effect, whereby an alternative visual representation, such as an avatar, may create a sense of being less personally exposed during speaking.

Importantly, the existing studies discussed above have primarily examined AR face filters in real-time communication contexts. Differently, the present study explores the potential for a similar experience in a recorded monologue context, where an Avatar serves as a visual representation of learners while they present their recorded monologues. A key difference between the previous study and the current study lies in the use of an AI avatar lip-sync website to synchronize learners' recorded voices with the avatars they generated themselves. This method may also be relevant to young learners, as previous studies have reported their interest in cartoon characters (Gupta et al., 2022) and positive perceptions of AI-animated characters (Xu et al., 2024). *Thus*, this study combines the potential masking function of an avatar with the engaging experience of creating their own Avatars and "bringing them to life" to deliver their monologues using their own voices.

In this study, students utilized AI lip-sync applications that were not originally designed for education. Applying such media in an instructional context is challenging and requires deep evaluation to avoid implementation failure (Hidayatullah et al., 2025; Kabudi et al., 2021). Consequently, it demands an investigation of students' acceptance of the applications, including usability and satisfaction (Kim & Moon, 2025; G. Liu & Ma, 2024). Therefore, technological experiences are essential in this current study.

Although previous studies have highlighted the potential of AI for EFL speaking and avatar-based technologies, their application has largely focused on real-time interactive speaking (Hirata, 2023; Ichino et al., 2022; Yoneyama et al., 2024). Recent research has further examined the potential of AI in relation to learners' speaking anxiety and performance (Susoy, 2026), while studies on AI avatars have investigated aspects such as academic motivation, knowledge acquisition, and cognitive load (Bakhtiari et al., 2026). These studies provide evidence concerning affective, performance, and learning outcomes associated with AI-assisted speaking and avatar-based learning. However, limited attention has been given to learners' short-term experiences during the production of recorded lip-sync avatar videos, particularly for monologue speaking practice at the early secondary school level. Therefore, this study focuses on learners' experiences during the creation of AI-assisted avatar lip-sync videos for monologue speaking practice, with specific attention to affective, engagement, and technological dimensions.

Based on these considerations, the study is formulated around the following research question: *"How do beginner EFL learners experience AI-assisted avatar lip-sync video production for monologue speaking across affective, engagement, and technological*

dimensions?" This study contributes to the growing body of research on AI-assisted language learning by providing empirical insights into learners' experiences with AI-assisted avatar lip-sync video production in EFL monologue speaking.

2. METHODOLOGY

2.1 Research Design

This study employed a convergent mixed-methods design that combined qualitative and quantitative approaches (Creswell et al., 2003). The qualitative approach examined students' experiences during the class, while quantitative data from questionnaires provided evidence of students' perceptions of these experiences. The two data sets were captured simultaneously and then analyzed separately. Afterwards, the qualitative themes and quantitative results were compared using an integration matrix to identify areas of convergence, complementarity, or divergence across the two strands. This integration aims to provide a comprehensive understanding of students' experiences with AI lip-sync video production in EFL monologue speaking practice.

2.2 Participants

The participants in this study were 28 eighth-grade female students (aged 13–14) from a junior high school in an Islamic boarding school in Indonesia. A total sampling approach was employed, with all students in the selected class participating in the study. The class was selected from several available eighth-grade classes with the school's permission because it was considered to include students with varying levels of English proficiency. As a female-only Islamic boarding-school context, the setting was considered relevant to the students' experiences of speaking through avatars, particularly because the activity replaced their visible facial representation with an avatar while retaining their voices. This context may have shaped how students perceived personal exposure, classroom interaction, and technology use during the activity.

2.3 Instructional Procedure

The instructional procedure was conducted over two classroom meetings during regular English lessons. Each meeting was 80 minutes (2×40 minutes). *In the first meeting,* the teacher began the lesson by explaining the speaking material and the guidelines of the video project. After that, the students rehearsed their pronunciation using Google Translate or other pronunciation tools, while the teacher provided additional support. Afterwards, they created

the avatar image using AI image generators (ChatGPT or Gemini) and recorded their voices in speaking monologues. *In the second meeting,* they uploaded the Avatar images and the recorded audio to Pippit AI to generate AI lip-sync animation videos. After the videos were generated, the students watched and reviewed their videos together.

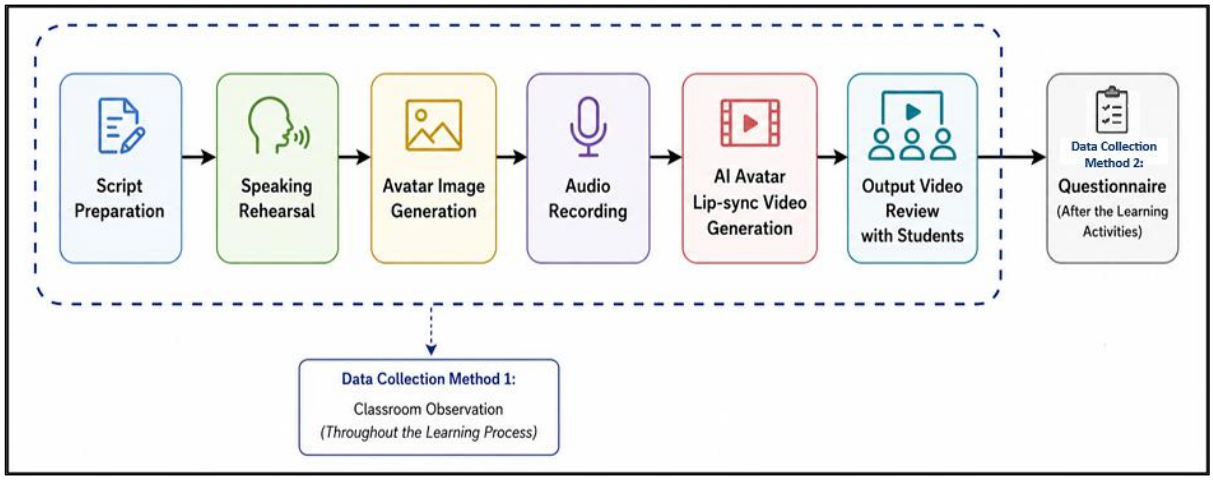

Figure 1. Learning Process and Data Collection Procedures

2.4 Ethical Consideration

As the research participants were minors (13-14 years old), ethical and privacy considerations were prioritized throughout the activities. *First,* the study was conducted with permission from the participating school, with the research proposal provided for their review. The three screenshots of video outputs created by the students using Pippit AI were displayed in this article (Figure 3) with permission from the participating school and the students' homeroom teacher. *Second,* the students did not use their personal email accounts; the AI platform was accessed using email accounts provided by the English teacher and the researchers for the learning activities. *Third,* to minimize the disclosure of personal identities, the audio recordings uploaded to the AI were assigned participant codes and used solely to generate lip-sync videos. No student names, photographs, or identifying information were mentioned in the audio recordings, videos, or the research report and publications.

2.5 Data Collection Techniques and Instrument Development

Two instruments were applied in this study.

1) Observation

Observation data were collected using structured field notes. The observer was the second researcher, who was specifically assigned to observe the learning activities and was not the students' regular classroom teacher. The first researcher served as the teacher/facilitator, while the third researcher assisted with documenting the learning process and provided

technical assistance during several stages of video production. Two structured observation records were produced, corresponding to the two instructional sessions. To manage potential observer subjectivity, the observer used predetermined observation domains and observable indicators and maintained a clear distinction between descriptive notes and researcher interpretations. Following the observations, the records were reviewed by the research team, and interpretations were discussed and clarified by referring to the original descriptive notes and the predefined observation indicators.

The instrument was designed as a table by bifurcating the records into several parts: 1) *time/stages* (monologue speaking rehearsal, making avatar, recording audio, generating video), *2) descriptive notes* to record students' real experiences during learning, and 3) *researcher interpretations* to record the researcher's analytical interpretation. To guide the observations, three domains and their observable indicators were specified in Table 1:

Table 1. Indicators of Observation

Domain	Observable indicators
Affective-related behaviors	Verbal and non-verbal responses during speaking and task performance
Engagement	Participation, involvement, and initiative during learning activities
Technology-related interactions	Actions when operating the tools / AI and responses to technical issues

2) Questionnaires

A questionnaire was used to obtain students' perceptions of the learning experiences. It was developed based on several aspects of learning experiences in these studies: students' affective perceptions related to anxiety and confidence (Krashen, 1985), engagement (Sundqvist & Uztosun, 2024), and technological experiences (usability and satisfaction) (G. Liu & Ma, 2024). It consists of 12 Likert-scale items and 2 open-ended questions. The Likert-scale items were measured on a 5-point scale: 1 (strongly disagree), 2 (disagree), 3 (neutral), 4 (agree), and 5 (strongly agree).

Table 2. Constructs of Questionnaire

Dimension	Item	Questions
Affective	1.	"I feel more relaxed speaking English using this method than speaking directly in front of the class."
	2.	"I am not afraid of making mistakes because my real face is not exposed."
	3.	"I feel confident when recording my voice."
	4.	"I feel confident with my animated video."
Engagement	5.	"I paid full attention while making the avatar video project."
	6.	"I paid attention to my pronunciation while recording."
	7.	"I really enjoyed creating my avatar video"
	8.	"I tried my best to make my video as good as possible."
Technological	9.	"The application used in this learning activity was easy to use."
	10.	"The application did not confuse me."
	11.	"I enjoyed using the application."
	12.	"The outcome videos are impressive."

Additionally, the two open-ended questions were: *1) "What part of today's learning activities did you enjoy the most? Why did you like it?"* and *2) "What part of today's learning activities did you find the most difficult?"* These questions were included to capture qualitative insights into students' favorite parts of the learning activity and the challenges they faced.

The questionnaire was administered in Indonesian to ensure that the A1-level participants could understand the items accurately. This instrument has been validated by two experts, comprising a Computer-Assisted Language Learning (CALL) expert and an EFL speaking expert. The experts evaluated the questionnaire for content relevance, construct representation, language clarity, and item appropriateness. Based on their recommendations, several items were revised to simplify wording, ensure the language was appropriate for junior high school learners, and align several questions with the AI-Assisted speaking activities. After that, to assess reliability, the revised Likert-scale questionnaire was pilot-tested with students who had characteristics similar to those of the target participants. The pilot test results demonstrated excellent internal consistency displayed in the following table:

Table 3. Reliability of Questionnaire

Subscale	Items	Cronbach's Alpha
Affective	4	0.86
Engagement	4	0.91
Technological	4	0.88
Overall	12	0.947

2.6 Data Analysis

The qualitative data from classroom observations and the two open-ended questionnaire responses were analyzed thematically (Braun & Clarke, 2006). The two qualitative sources were first coded separately to preserve the distinct nature of the data. The thematic analysis comprised several stages. *First*, one researcher (Researcher 2) conducted the initial coding by identifying meaningful data segments and assigning codes related to students' affective experiences, engagement, and technological experiences. *Second*, the other two researchers then reviewed the codes and compared them with the original data. Differences were discussed and resolved by referring back to the original data. *Third*, the codes were then grouped into broader themes based on recurring patterns. Finally, the themes from the two data sources were compared to identify convergent, complementary, or divergent findings.

The quantitative data from the Likert-scale questionnaire were analyzed using mean scores and standard deviations (SD) (Creswell, 2015). Mean scores were interpreted using five

equal-width categories: very low (1.00–1.80), low (1.81–2.60), moderate (2.61–3.40), high (3.41–4.20), and very high (4.21–5.00). The categories follow the statistical principle for determining class width (Larson & Farber, 2019), calculated by dividing the range of the five-point scale by the number of categories: (5 − 1)/5 = 0.80. In addition, SD values were interpreted using three equal-width categories to indicate response variability: low (0.00–0.67), moderate (0.68–1.33), and high (1.34–2.00). The SD interval width was calculated by dividing the theoretical SD range of 0–2 for a five-point scale by three categories, yielding an interval width of approximately 0.67.

The qualitative and quantitative findings were then compared to identify areas of convergence, complementarity, or divergence between the two datasets. The integrated findings were interpreted to provide a more comprehensive understanding of learners' affective, engagement, and technological experiences with AI-Assisted avatar lip-sync video production.

3. FINDINGS

3.1 Students' Experiences in Creating EFL Monologue Speaking Videos Using Lip Sync AI Technology

The learning activity began with an explanation of the video production process. The teacher demonstrated the workflow and provided sample video outputs as references. Afterward, the students rehearsed the speaking scripts they had prepared as homework. The topics were Islamic tourism sites around the world, such as the Taj Mahal, Hagia Sophia, Tarim, and other destinations, including those in Indonesia. To support pronunciation accuracy, students were encouraged to use Google Translate's audio feature for unfamiliar lexical items. It aims to foster efficiency through self-regulated learning.

3.1.1 Making The AI-Generated Avatar

In this stage, the students created their avatars in ChatGPT or Gemini by *typing a descriptive prompt* for avatar appearances, and the background setting must match the site mentioned in the speaking scripts. For example, *"Please make an anime-style image of a beautiful young girl wearing a pink hijab, the background is Hagia Sophia. Portrait resolution 9:16.*" (Figure 1). They were allowed to choose their avatars' preferred animation styles, such as anime, cartoon, 3D animated, or semi-realistic, and to describe details like fashion,

appearance, pose, and expressions. This stage almost overlapped in time, since they were enthusiastically engaged with the detailed creations. The teacher even had to limit their time during this session and continue to the next stage.

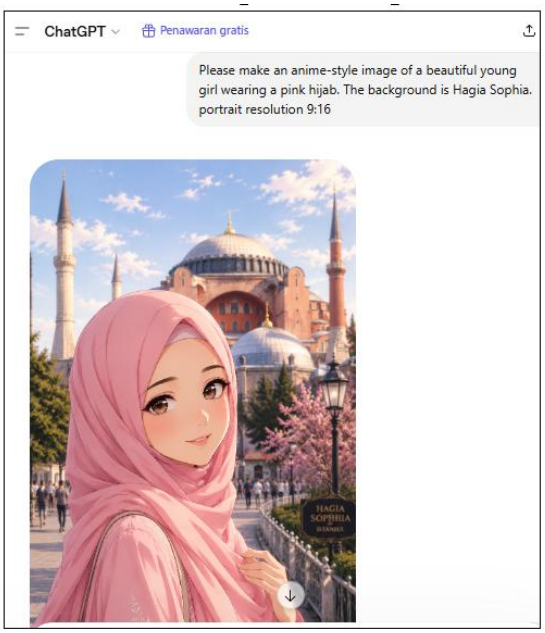

Figure 2. Example of a student-produced avatar created using ChatGPT.
Source: The image was reproduced using ChatGPT based on a student-written descriptive prompt.

3.1.2 Audio Recording

During recording, without being commanded, they repeatedly re-recorded their voices until they achieved a satisfactory audio output. When the teacher asked why they kept recording the same script, some students explained that they had mispronounced certain words or were not yet fluent enough. Others responded more generally, saying, *"Wait, Ma'am, it is not good enough yet."* These behaviors indicate self-monitoring and persistence, as students independently noticed aspects of their pronunciation or fluency that they wanted to improve. However, the specific standards students used to determine whether a recording was *"good enough"* could not be fully established through classroom observation. Some students also asked the teacher for confirmation, such as, *"Ma'am, is this correct?"* In these cases, the teacher provided feedback to help them evaluate whether the recording was sufficiently accurate. Students who did not seek confirmation continued to evaluate and revise their recordings independently. Thus, the observation is interpreted as evidence of self-monitoring and persistence, rather than confirmed self-regulated learning.

3.1.3 Generating Videos Using Lip Sync AI Website (Pippit AI)

Before this stage, several students asked, *"What is next?"* They were eager for the next stage. Some students with sufficient technological proficiency completed the task earlier by

following the guidelines. In contrast, some students lagged and lost track of time, so that at the end, they were concerned about completing the task. Teachers then helped the students to speed up their progress. Although the teacher did not instruct immediate completion, the students continued working on the activity and repeatedly asked about the next stage. As the activity progressed, the students appeared increasingly eager and enthusiastic to see the final result.

Even though task completion did not occur simultaneously across students, the overall process did not show boredom or pressure during task execution. On the contrary, the students appeared highly immersed in the activity, as indicated by their reduced awareness of the remaining time. Some students became so engaged in completing the video production process that they did not appear to notice the passage of time until the teacher reminded them of the remaining class time. This observation suggests a flow-like experience characterized by sustained involvement and reduced awareness of time during the activity.

3.1.4 Students' Reactions to the Outcome Videos

Finally, when the videos were finished (Figure 2), they reacted with excitement and amazement. They also laughed at errors such as unusual gestures and unexpected details generated by AI. Some comments recorded during the classroom observation included, *"Wow, it is so cute!", "The pose is so funny!",* and *"The finger looks weird".* A student even expressed, *"Can we make another one tomorrow?"*

Figure 3. Screenshots of original student-produced AI-generated avatar videos created using Pippit AI.
Source: Original student-produced videos generated using Pippit AI during the research activity

3.1.5 Challenges during the process

During the process, several challenges were encountered. *First*, classroom noise was impossible to eliminate since the activity was in a classroom. *Second*, their microphone quality

varied, resulting in some videos with barely audible sound. *Third*, technical issues in internet connection and email verification occurred. *Fourth*, the activity applied several applications: Google Translate, Gemini/ChatGPT, voice recorder, and Pippit AI. Consequently, this learning activity required well-structured classroom management. Despite these challenges, the overall implementation proceeded smoothly, with the AI lip-sync animation videos as the outcomes.

3.2 Students' Perceptions Toward Their Experiences

3.2.1. Results of Likert-scale Questionnaire

The above observation results were then compared and integrated with the quantitative data from the Likert-scale questionnaire, separated into three parts based on the aspects. *The first* part is affective factors, as displayed in Table 4.

Table 4 Results of Questionnaire Related to Affective Factors (Anxiety and Confidence).

No	Questions	Means	Category	SD	Category
1.	"I feel more relaxed speaking English using this method than speaking directly in front of the class."	3.82	High	0.82	Moderate variability
2.	"I am not afraid of making mistakes because my real face is not exposed."	3.79	High	1.03	Moderate variability
3.	"I feel confident when recording my voice."	4.14	High	0.80	Moderate variability
4.	"I feel confident with my animated video."	4.04	High	0.96	Moderate variability
Affective factors:					
•	Anxiety (item 1-2)	3.80	High	0.86	Moderate variability
•	Confidence (item 3-4)	4.09	High	0.81	Moderate variability
Overall Affective factors		**3.95**	**High**	**0.78**	**Moderate variability**

The overall affective aspect achieved a high mean score (M = 3.95) with moderate variability (SD = 0.78). Specifically, the anxiety dimension received a high mean score (M = 3.80, SD = 0.86), as did confidence (M = 4.09, SD = 0.81). It indicates that the students perceived speaking comfort and confidence when delivering their monologues during the lesson.

Table 5. Results of Questionnaire Related to Engagement.

No	Questions	Means	Category	SD	Category
B.	**Engagement**				
5.	"I paid full attention while making the avatar video project."	4.39	Very High	0.69	Moderate variability
6.	"I paid attention to my pronunciation while recording."	4.11	High	0.83	Moderate variability
7.	"While creating the video, I felt engaged and lost track of time."	4.25	Very High	0.93	Moderate variability
8.	"I tried my best to make my video as good as possible."	3.61	High	0.96	Moderate variability
Overall Engagement		**4.09**	**High**	**0.73**	**Moderate variability**

The students' engagement also showed a high mean score (M = 4.09) with moderate variability (SD = 0.73). This finding indicates that students reported being highly engaged and actively involved during the speaking activities involving AI lip-sync avatars.

Table 6. Results of Questionnaire Related to Technological Experiences.

No	Questions	Means	Category	SD	Category
9.	"The application used in this learning activity was easy to use."	3.26	Moderate	1.38	High variability
10.	"The application did not confuse me during use."	3.39	Moderate	1.23	Moderate variability
11.	"I enjoyed using the application."	4.26	Very High	0.76	Moderate variability
12.	"The outcome videos are impressive."	4.48	Very High	0.70	Moderate variability
•	Usability	3.30	Moderate	1.23	Moderate variability
•	Satisfaction	4.37	Very high	0.66	Low variability
Overall Technological Experiences		**3.79**	**High**	**0.86**	**Moderate variability**

In technological experiences, the overall aspect yielded a high mean score (M = 3.79, SD = 0.86). Interestingly, the usability received a moderate mean score (M = 3.30) with moderate variability (SD = 1.23). Conversely, the students reported very high satisfaction with the technology (M = 4.37, SD = 0.66), with low variability. This contrast suggests that, despite several technical constraints, they were highly satisfied with the learning experience and the outcome videos.

Overall, the Likert-scale questionnaires yielded high and consistent responses across all aspects (affective, engagement, and technological experiences). It indicates that the implementation of AI-assisted avatar lip-syncs video production for EFL monologue speaking practice was perceived positively by the students. This explanation is more comprehensively displayed in the following chart.

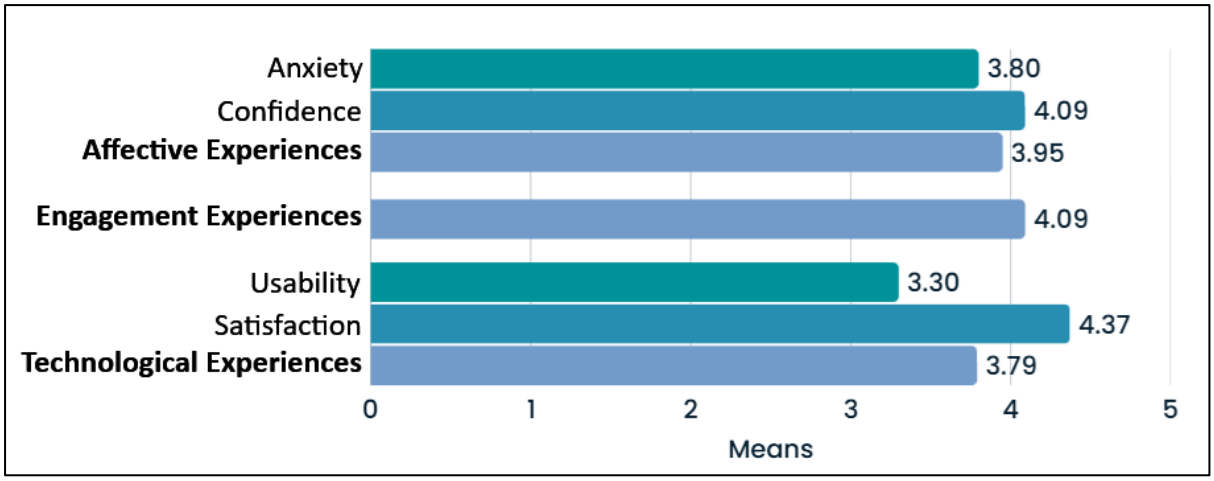

Figure 4. Chart of Mean Scores

3.2.2. Results of Likert-scale Questionnaire

The open-ended questions were: *1) "What part of today's learning activities did you enjoy the most? Why did you like it*?" (students' favorite segments) and *2) "What part of today's learning activities did you find the most difficult?"* (Obstacles). The results of thematic analysis are in the following table.

Table 7. Result of Open-ended Questionnaires

Questions	Themes	Example Responses	Frequency
Question 1 (students' favorite segments)	Audio recording	*"Recording. I can record again if my pronunciation is wrong"* *"Recording, since my face is not shown."*	15
	Creating avatar	*"Making the avatars, because I can make my avatar as I want."*	5
	The outcome videos	*"The results. It is amazing. My avatar is beautiful, and it can talk"*	5
Question 2 (Obstacles)	Technical issues	*"Internet connection"* *"Email verification"*	8
	Noise	*"Noises when recording"*	7
	Pronunciation	*"Pronouncing the sentences."*	5
	Usability	*"How to use the application."*	1

The coding process revealed that a single response could be assigned more than one code when it contained multiple relevant ideas. Moreover, of the 28 responses, three students were found to be unclear or unrelated to Question 1 *(students' favorite segments)* by answering *"joyful learning"* and *"self-confidence"*. The responses were then excluded from the thematic categories. Accordingly, the frequencies reported in Table 6 represent coded theme occurrences rather than the number of participants, which caused the total frequency to be higher or lower than the number of participants.

The table confirmed and enriched the classroom observation by providing the student perspective, revealing several obstacles encountered during the process: technical difficulties, classroom noise, and other usability issues. A few reported difficulties with pronunciation. Despite these obstacles, students shared their thoughts on their favorite stages of learning, such as creating avatars, recording, and generating videos. Interestingly, many students reported enjoying the recording process for the characters they created. Moreover, they delivered compliments on fun, joyful activities and improvements in self-confidence. These data show students' positive perceptions of AI-Assisted avatar lip-sync video production for EFL monologue speaking.

3.3. Integration Matrix of Qualitative and Quantitative Findings

Table 4. Integration Matrix of Qualitative and Quantitative Findings

Dimension	Observation	Likert-scale questionnaire	Open-ended questionnaire	Integration
Affective	Students appeared **excitement**, particularly during avatar creation. During speaking practice, students appeared **relaxed** and **persistence** to the quality. They also appeared **eager** to watch the result and **amazed** by the generated video.	**Overall M = 3.95 (high)**. Students reported feeling relatively relaxed and less fear of making mistakes because their real faces were not exposed (M = 3.80). Confidence was also relatively high (M = 4.09).	Students expressed positive thoughts about creating avatars because they could make their avatars as they wanted and enjoyed the outcome videos because the results were amazing and their avatars could talk.	**Convergence and complementarity.** The observation and quantitative findings converge in showing students' positive affective responses, including excitement, enjoyment, relaxation, and confidence. The open-ended responses complement these findings by highlighting creative freedom, excitement about the speaking avatars, and speaking comfort when students' faces were not exposed.
Engagement	Students pay **attention** to details of their creation. During recording, they performed self-monitoring by **repeating** their recordings voluntarily. They showed **full involvement** that they lost track of time.	**Overall M = 4.09**. Full attention to the project was high (M = 4.39), attention to pronunciation was high (M = 4.11), and feeling engaged/lost track of time was high (M = 4.25).	Recording was the most favored stage (15 students), particularly because students could repeat recordings when their pronunciation was wrong.	**Convergence.** The observed voluntary repeated recording, persistence, joyful, and flow-like experiences. Those findings correspond closely with the high quantitative engagement score, while open-ended responses explain that opportunities to repeat and improve pronunciation supported students' engagement.
Technological	Students experienced several technical **obstacles**, including sign-up/login problems and complicated stages in using several tools.	**M = 3.79**. Students reported moderate ease of use (M= 3.30), but **satisfaction** with the application and the outcome videos was very high (M = 4.37).	Students reported **technical and usability difficulties**, but also **valued** recording, avatar customization, and the outcome videos.	**Convergence and divergence.** Quantitative and qualitative findings converge on the presence of technical difficulties, but diverge in overall evaluation: despite these difficulties, students reported very high satisfaction with the application and outcome videos.

4. DISCUSSION

The integrated findings in Table 8 indicate that students reported positive affective responses, including excitement, enjoyment, comfort, and confidence during the activity. These responses may be understood in relation to several features of the learning activity. *First*, avatar masking may have contributed to speaking comfort, as some students reported feeling

more comfortable when their faces were not exposed. *Second*, the recorded format allowed students to speak without performing their monologues directly in front of their classmates. *Third*, students could repeatedly record their voices when their pronunciation was not satisfactory, providing opportunities for practice and improvement. *Finally*, teacher support during pronunciation rehearsal may also have contributed to their positive experience. This complements the quantitative and observational findings by suggesting that students' positive affective responses were associated with multiple features of the learning activity rather than the avatar alone.

The findings above can be interpreted through the lens of affective filters theory by Krashen (1985) and several studies on the "avatar effect", which suggest that avatars may provide a comfortable and supportive speaking practice (Hirata, 2023; Ichino et al., 2022; Leong et al., 2023). Furthermore, their enthusiasm for creating avatars, including attention to details such as beauty, fashion, resolution, and overall quality, demonstrated their dedication to this activity. This observation also aligns with Hirata (2023) and Hwang and Lee (2025), who discussed the potential of avatars to influence learners' self-perception and willingness to speak. However, because avatar masking, recorded performance, repeated recording, and teacher support occurred together, the present study cannot determine which factor contributed most to the students' reported positive affective responses. In addition, students' first-time experience with the technology may have contributed to their initial excitement and positive responses.

During the process, students appeared highly engaged and enthusiastically immersed. They paid attention to details of their avatar creation and the whole process of video production, until they lost track of time. These indicators are closely related to the flow state experience, when students are involved immersively in the learning process and lose time awareness (Csikszentmihalyi, 1990; Fauzi et al., 2025). However, because flow was not directly measured in this study, these findings should be interpreted as "flow-like experiences". In addition, students' engagement may have been strengthened by the multimodal nature of the activity, which combined visual design, voice recording, animation, and AI-generated outputs. The combination of these elements may have made the speaking activity more enjoyable and engaging.

Furthermore, they conducted the audio recording repeatedly to achieve the desired audio quality. This effort was confirmed by the open-ended questionnaire, that they valued the recording process since they were allowed to re-record when their pronunciation was wrong.

It reflected students' self-monitoring and an investment in the final product. This behavior is consistent with several characteristics of the flow-like experience, namely goal-oriented activity and a sense of control over task completion (Csikszentmihalyi, 1990; Heutte et al., 2021). Additionally, this finding also echoes Hasan et al (2025), who reported that self-monitoring tasks can support pronunciation improvement. Interestingly, this additional finding emerged beyond the scope of the initial data collection plan, providing valuable insights into students' self-monitoring behaviors. It is also possible that teacher guidance during the production process contributed to students' persistence by helping them overcome pronunciation and technical difficulties while maintaining focus on completing the task.

When the students' AI lip-sync videos were completed, they appeared eager to see the results and said the videos were impressive. Therefore, their eagerness to watch the final video likely reflected not only their satisfaction with their performance but also their curiosity about the AI's results. This curiosity may have complemented their engagement throughout the activity, although this was not directly measured in the study. This interpretation is supported by Schutte and Li (2025), who found that curiosity is one of the relevant factors to students' engagement with AI technology. This observation is also in line with Xu et al., (2024), who examined children's perceptions of AI-enabled interactive media characters and highlighted how interactions with such characters shape their positive experiences with AI-enabled media. Overall, these findings suggest that students' eagerness or curiosity about the AI-generated results may contribute to their engagement throughout the activity.

From a technological perspective, it is important to note that Pippit AI was not originally designed for learning, so using it in an instructional context was relatively challenging. Moreover, this lesson also utilized multiple applications and tools, including Google Translate, ChatGPT/Gemini, a voice recorder, and Pippit AI, resulting in a more complex workflow throughout the learning process. This condition may explain why the questionnaire on usability showed moderate results, and the open-ended questionnaire explained several constraints. Conversely, satisfaction was very high, indicating that the obstacle did not lead to negative feedback on the final product. The students' positive impressions and amazement may have been influenced not only by AI automation, which provides more efficient and instant results than conventional animation software (Hai et al., 2026), but also by the opportunity to interact with a new AI application in a classroom setting. It is remarkably consistent with the findings of Wang et al. (2025), who reported that EFL learners perceived the use of multimodal AI positively in a language-learning context.

Accordingly, this AI-Assisted speaking strategy was positively experienced by students in this class, particularly in terms of affective responses, engagement, and technology-based learning experiences. Unlike previous studies that primarily utilized AI technology to support real-time, live conversations (Hirata, 2023; Jedrasiak et al., 2024; Leong et al., 2023), this study engages students in creating their own avatars and synchronizing their recorded voices with those avatars using AI. This approach integrates speaking practice with a creative process assisted by modern technology, which was associated with students' enjoyment and willingness to participate in the speaking activities. In addition, self-monitoring emerged as a notable serendipitous finding. Although it was not a primary focus of the study, students' repeated review and revision of their recordings provided an additional indication of their active involvement in monitoring their own performance. This finding is noteworthy because it suggests a potential area for further investigation in future research.

However, the positive experiences observed in this study should also be interpreted in light of several possible contributing factors, including the novelty of the AI technology, teacher support during implementation, repeated opportunities for performance improvement, and the multimodal appeal of AI-generated avatar production. Therefore, future studies using longitudinal designs and direct measurement of flow experiences are needed to determine whether these positive experiences can be sustained beyond the initial novelty phase. In practice, it may serve as an alternative instructional practice in EFL learning contexts. The findings also suggest that there may be many unexplored opportunities to utilize AI media across various instructional activities in language learning. Theoretically, these findings encourage further exploration of AI use in EFL teaching contexts and in broader instructional and pedagogical settings.

5. CONCLUSION

This study implemented an AI-assisted speaking activity in which students created their own avatars, rehearsed their pronunciation, recorded speaking monologues, and synchronized their recorded voices with the avatars using AI lip-sync technology. The integrated findings of this study indicate that the learning activity was positively experienced by the EFL learners across affective, engagement, and technological dimensions. They valued high enjoyment, repeated practice during the activity, and speaking without face exposure. From a technological perspective, despite several technical difficulties, the students reported high satisfaction with the application and outcome videos. Overall, the findings suggest that AI-assisted avatar lip-

sync activities may serve as an adaptable and appealing alternative for EFL monolog speaking practice.

Despite these positive findings, this study has several limitations. *First*, it was conducted in a single educational context over a relatively short period with female-only participants, thereby limiting the generalizability of the findings. *Second*, the absence of baseline or control data, reliance on self-reported data and researcher observation, use of multiple digital tools, and lack of direct speaking-performance measures also limit the interpretation of the findings. Furthermore, as the participants were minors, the use of AI-based tools requires careful attention to ethical, privacy, and participant-protection considerations. Therefore, future research should examine the longer-term use of AI-assisted avatar learning across more diverse educational settings and participant groups. Future studies could also employ more focused measures informed by psychological and pedagogical constructs such as flow, engagement, and learning autonomy. In practice, this study suggests that EFL teachers may consider AI-assisted avatar lip-sync activities and other multimodal AI approaches as potential classroom activities, with appropriate ethical, privacy, and age-related safeguards.

DECLARATION OF GENERATIVE AI

The authors would like to declare the use of artificial intelligence (AI) tools during the preparation of this manuscript. Specifically, ChatGPT was utilized for brainstorming initial concepts and improving the clarity, grammar, and overall language of the text. All final content was reviewed and approved by the authors, who take full responsibility for the integrity of the work.

REFERENCES

Aldosari, M. S. (2024). Another World with Artificial Intelligence in Speaking Classes: To Delve into the Influences on Willingness to Communicate, Personal Best Goals, and Academic Enjoyment. *Computer-Assisted Language Learning Electronic Journal*, *25*(4), 439–463. https://callej.org/index.php/journal/article/view/470

Almukhaild, H., & King, J. (2024). Scaling the apex of engagement: Anxious language learners' positive flow experiences during speaking tasks. *System*, *124*, 103373. https://doi.org/10.1016/J.SYSTEM.2024.103373

Bakhtiari, E., Suriano, R., & Grasso, G. M. (2026). AI avatars in education: effects on knowledge acquisition, academic motivation, and cognitive load. *Frontiers in Psychology*, *17*, 1860755. https://doi.org/10.3389/FPSYG.2026.1860755

Braun, V., & Clarke, V. (2006). Using Thematic Analysis in Psychology. *Qualitative Research in Psychology*, *3*(2), 77–101. https://doi.org/10.1191/1478088706qp063oa

Bui, G. (2026). Effects of task interactivity and willingness to communicate on speech fluency in second language learning tasks. *RELC Journal*, 1–20. https://doi.org/10.1177/00336882261422018

Creswell, J. W. (2015). *Educational Research: Planning, Conducting, and Evaluating Quantitative and Qualitative Research*. Pearson.

Creswell, J. W., Plano-Clark, V. L., Gutmann, M. L., & Hanson, W. E. (2003). Advanced Mixed Methods Research Designs. In *Handbook of Mixed Methods in Social and Behavioral Research*. SAGE Publications.

Csikszentmihalyi, M. (1990). *Flow: The psychology of optimal experience*. Harper & Row.

Ding, W., & Lian, X. (2020). Examining Young Learners: Research and Practice in Assessing the English of School-Aged Learners. *Language Assessment Quarterly*, *17*(3), 331–335. https://doi.org/10.1080/15434303.2020.1754830

Ebadi, S., Velayati, S., Ramezanzadeh, A., & Rawdhan Salman, A. (2025). Exploring The Impact of AI-Powered Speaking Tasks on EFL Learners' Speaking Performance and Anxiety: An Activity Theory Study. *Acta Psychologica*, *259*, 105391. https://doi.org/10.1016/J.ACTPSY.2025.105391

Fauzi, R., Restendi, D., Nugraha, F., & Ayundhari, V. L. (2025). Creating Immersive Learning Experiences With AI: An Android App Approach. *JENTRE*, *5*(2), 104–109. https://doi.org/10.38075/JEN.V5I2.521

Gong, Z., & Kanai, H. (2025). Effects of appearance modifications on oral presentation anxiety in video conferencing. *Scientific Reports 2025 15:1*, *15*(1), 30090. https://doi.org/10.1038/s41598-025-15716-z

Gupta, A., Kumar, J., Tewary, T., & Virk, N. K. (2022). Influence of Cartoon Characters on Generation Alpha in Purchase Decisions. *Young Consumers*, *23*(2), 282–303. https://doi.org/10.1108/YC-06-2021-1342

Hai, T. T., Mai, D. T. T., & Hanh, N. Van. (2026). A rapid review of using AI-generated instructional videos in higher education. *Frontiers in Computer Science*, *7*, 1721093. https://doi.org/10.3389/fcomp.2025.1721093

Hasan, E., Altakhaineh, A. R. M., & Chahwan, R. (2025). Study on the Effect of Self-Monitoring Tasks on Improving Pronunciation of Foreign Learners of Korean in Blended Courses. *Open Education Studies*, *7*(1). https://doi.org/10.1515/edu-2025-0069

Heutte, J., Fenouillet, F., Martin-Krumm, C., Gute, G., Raes, A., Gute, D., Bachelet, R., & Csikszentmihalyi, M. (2021). Optimal Experience in Adult Learning: Conception and Validation of the Flow in Education Scale (EduFlow-2). *Frontiers in Psychology*, *12*, 828027. https://doi.org/10.3389/fpsyg.2021.828027

Hidayatullah, M. H., Suryati, N., Cahyono, B. Y., & Mawaddah, N. (2025). A Systematic Literature Review of Artificial Intelligence in Academic Writing: Challenges and Opportunities. *Journal of Research on English and Language Learning (J-REaLL)*, *6*(2), 249–266. https://doi.org/10.33474/J-REALL.V6I2.23821

Hirata, Y. (2023). Do avatar-assisted virtual classrooms work for students with low speaking confidence?: A qualitative study. *Interactive Technology and Smart Education*, *20*(3), 335–349. https://doi.org/10.1108/ITSE-10-2022-0142

Huang, F., & Zou, B. (2024). English Speaking with Artificial Intelligence (AI): The Roles of Enjoyment, Willingness to Communicate With AI, and Innovativeness. *Computers in Human Behavior*, *159*, 108355. https://doi.org/10.1016/J.CHB.2024.108355

Huang, S. E., & Liu, Y. T. (2022). How to Talk to Myself: Optimal Implementation for Developing Fluency in EFL Speaking Through Soliloquizing. *English Teaching & Learning*, *47*(2), 145. https://doi.org/10.1007/S42321-022-00110-Z

Hwang, Y., & Lee, J. H. (2025). "Yes, I Am More Confident with My Avatars": Integrating EFL Students' Speaking Practice into Metaverse. *RELC Journal*, *56*(3), 760–771. https://doi.org/10.1177/00336882241251952

Ichino, J., Ide, M., Yokoyama, H., Asano, H., Miyachi, H., & Okabe, D. (2022). "I've talked without intending to": Self-disclosure and Reciprocity via Embodied Avatar. *Proceedings of the ACM on Human-Computer Interaction*, *6*(CSCW2), 1–23. https://doi.org/10.1145/3555583

Jedrasiak, K., Wolanski, R. R., & Bereska, D. (2024). Algorithm of Behaviour of Virtual Avatars During Public Speaking Training in a VR Environment. *SAFETY & FIRE TECHNOLOGY*, *64*(2), 46–54. https://doi.org/10.12845/SFT.64.2.2024.3

Ji, X., Liao, Z., Dong, L., Tang, Y., Li, G., & Mao, M. (2024). 3D facial animation driven by speech-video dual-modal signals. *Complex & Intelligent Systems*, *10*(5), 5951–5964. https://doi.org/10.1007/s40747-024-01481-5

Kabudi, T., Pappas, I., & Olsen, D. H. (2021). AI-Enabled Adaptive Learning Systems: A Systematic Mapping of The Literature. *Computers and Education: Artificial Intelligence*, *2*, 100017. https://doi.org/10.1016/J.CAEAI.2021.100017

Kim, J., & Moon, J. (2025). Determinants of Usefulness of Chat GPT for Learning in Technology Acceptance Model (TAM) Using Information Credibility, Fun, and Responsiveness and Moderating Role of Fun. *Sage Open*, *15*(1). https://doi.org/10.1177/21582440251320173

Krashen, S. D. (1985). *Principles and Practice in Second Language Acquisition*. Longman.

Larson, R., & Farber, B. (2019). *Elementary statistics : picturing the world*. Pearson

Education, Inc.

Lee, J. S., & Chiu, M. M. (2023). Modeling EFL Learners' Willingness to Communicate: The Roles of Face-to-Face and Digital L2 Communication Anxiety. *Annual Review of Applied Linguistics*, *43*, 64–87. https://doi.org/10.1017/S0267190523000090

Leong, J., Perteneder, F., Rajvee, M. R., & Maes, P. (2023). "Picture the Audience...": Exploring Private AR Face Filters for Online Public Speaking. *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, 1–13. https://doi.org/10.1145/3544548.3581039

Liu, G., & Ma, C. (2024). Measuring EFL learners' use of ChatGPT in informal digital learning of English based on the technology acceptance model. *Innovation in Language Learning and Teaching*, *18*(2), 125–138. https://doi.org/10.1080/17501229.2023.2240316

Liu, L., Wang, J., Chen, S., & Li, Z. (2024). VividWav2Lip: High-Fidelity Facial Animation Generation Based on Speech-Driven Lip Synchronization. *Electronics 2024, Vol. 13, Page 3657*, *13*(18), 3657. https://doi.org/10.3390/ELECTRONICS13183657

Philp, J., & Duchesne, S. (2016). Exploring Engagement in Tasks in the Language Classroom. *Annual Review of Applied Linguistics*, *36*, 50–72. https://doi.org/10.1017/S0267190515000094

Schutte, N. S., & Li, H. (2025). The role of self-efficacy and curiosity in student use of artificial intelligence (AI). *International Journal of Educational Technology in Higher Education 2025 22:1*, *22*(1), 73-. https://doi.org/10.1186/S41239-025-00574-6

Shi, W., & Shakibaei, G. (2025). Insights Into the Effectiveness of Artificial Intelligence-Integrated Speaking Instruction in Enhancing Speaking Skills and Social–Emotional Competence as Well as Reducing Demotivation and Shyness. *European Journal of Education*, *60*(3), e70174. https://doi.org/10.1111/ejed.70174

Sundqvist, P., & Uztosun, M. S. (2024). Extramural English in Scandinavia and Asia: Scale Development, Learner Engagement, and Perceived Speaking Ability. *TESOL Quarterly*, *58*(4), 1638–1665. https://doi.org/10.1002/tesq.3296

Susoy, Z. (2026). Reducing anxiety and enhancing performance: the impact of AI chatbots versus human facilitation on EFL speaking assessment outcomes. *Frontiers in Psychology*, *16*, 1745942. https://doi.org/10.3389/fpsyg.2025.1745942

Szmigiel, R., & Lee, D. (2016). Avatars in E- and U-Learning. In *Web Design and Development* (Vols. 2–3, pp. 792–815). IGI Global. https://doi.org/10.4018/978-1-4666-8619-9.ch037

Szyszka, M., Peltonen, P., & Lintunen, P. (2026). Unravelling the relationship between language anxiety and foreign language speech fluency in a monologue production. *Journal of Multilingual and Multicultural Development*, *47*(1), 323–337. https://doi.org/10.1080/01434632.2024.2387149

Wang, C., Du, Y., & Zou, B. (2025). Learners' Acceptance and Use of Multimodal Artificial Intelligence (AI)-Generated Content in AI-Mediated Informal Digital Learning of English. *International Journal of Applied Linguistics*, *36*(1), 927–940. https://doi.org/10.1111/ijal.12827

Xing, C., & Saeed, M. A. (2025). A Systematic Review on Artificial Intelligence (AI) Technologies in ESL/EFL Speaking Skills. *International Journal of TESOL Studies*, *8*(2), 240–270. https://doi.org/10.58304/ijts.250908

Xu, Y., Thomas, T., Li, Z., Chan, M., Lin, G., & Moore, K. (2024). Examining children's perceptions of AI-enabled interactive media characters. *International Journal of Child-Computer Interaction*, *42*, 100700. https://doi.org/10.1016/J.IJCCI.2024.100700

Yoneyama, J., Fujimoto, Y., Okazaki, K., Sawabe, T., Kanbara, M., & Kato, H. (2024). Augmented conversations: AR face filters for facilitating comfortable in-person interactions. *Journal on Multimodal User Interfaces 2024 19:1*, *19*(1), 57–74. https://doi.org/10.1007/S12193-024-00446-9

Zhong, J., Ismail, L., & Lin, Y. (2025). Investigating EFL students' engagement in project-based speaking activities: from a multi-dimensional perspective. *Frontiers in Psychology*, *16*, 1598513. https://doi.org/10.3389/fpsyg.2025.1598513

Zou, B., Xie, S., & Wang, C. (2025). Students' Willingness to Communicate (WTC) in Using Artificial Intelligence (AI) Technology in English-Speaking Practice. *International Journal of Information and Communication Technology Education*, *21*(1), 1–18. https://doi.org/10.4018/IJICTE.375387