Scientific Research Publishing

Open Journal of Social Sciences, 2026, 14(11), 21-41
https://www.scirp.org/journal/jss
ISSN Online: 2327-5960
ISSN Print: 2327-5952

# A Generative AI Teaching Ecosystem and English Majors' Learning Emotions, Classroom Engagement, and Oral Willingness to Communicate: Evidence from an English Audio-Visual-Oral Course

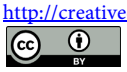
**Yue Zhang**

College of English, Zhejiang Yuexiu University, Shaoxing, China
Email: 20132060@zyufl.edu.cn

**How to cite this paper:** Zhang, Y. (2026). A Generative AI Teaching Ecosystem and English Majors' Learning Emotions, Classroom Engagement, and Oral Willingness to Communicate: Evidence from an English Audio-Visual-Oral Course. *Open Journal of Social Sciences, 14,* 21-41.
https://doi.org/10.4236/jss.2026.1411003

**Received:** October 10, 2026
**Accepted:** November 2, 2026
**Published:** November 5, 2026

Copyright © 2026 by author(s) and Scientific Research Publishing Inc.
This work is licensed under the Creative Commons Attribution International License (CC BY 4.0).
http://creativecommons.org/licenses/by/4.0/



Open Access

## Abstract

Since generative artificial intelligence (AI) entered the foreign language classroom, evidence on its relationship with learners' emotions and participation has come mostly from short-term studies of single tools; semester-long, multi-component course ecosystems remain under-examined. This study compared six intact classes of first-year English majors in an English Audio-Visual-Oral course at an application-oriented university in China: five classes (n = 144) taught within an "AI teaching ecosystem" (graded AI text-dialogue tasks, AI micro-lectures, and AI review songs embedded in the course's online smart course) and one class (n = 30) taught conventionally. Learning emotions, classroom engagement, oral willingness to communicate (WTC), AI literacy, and final examination scores were compared at the end of the semester. Neither primary outcome showed a between-group difference: speaking anxiety, measured by a single retained item that replaced the planned eight-item scale (g = 0.23), and WTC (g = 0.05). Among the other outcomes, the AI-ecosystem classes reported higher private and social enjoyment (g = 0.62 and 0.83). In student-level models that treat students as independent observations, both differences survived correction for multiple comparisons, and social enjoyment did so in every sensitivity analysis; because the conventional condition comprised a single class, these $p$ values do not provide condition-level evidence. Classroom engagement, AI literacy, and examination scores showed no detectable differences. Within the AI-ecosystem group, acceptance of the AI activities correlated positively with enjoyment, engagement, and WTC. Drawing on the asymmetry of antecedents across constructs and a "shared experience" hypoth-

esis, the paper argues that the differences associated with an AI teaching ecosystem systematically embedded in a course concentrate in the emotional ecology of the classroom, particularly collective enjoyment. For anxiety, the null result is consistent with the absence of a direct pathway from text-based dialogue tasks to oral classroom situations; for WTC, a null between-group contrast coexists with a positive within-group association that this design cannot resolve. Because the teaching condition was confounded with class, teacher, and additional graded practice in this intact-class comparison, the associations cannot be interpreted causally.

## Keywords

Generative Artificial Intelligence, English Audio-Visual-Oral Course, Foreign Language Enjoyment, Classroom Engagement, Willingness to Communicate, Foreign Language Speaking Anxiety

## 1. Introduction

Generative artificial intelligence (AI) is reconfiguring the goals, pathways, and teaching relationships of foreign language education. The goals and pathways of building "new foreign languages" in the AI era have been set out systematically (He & Wang, 2025), and a human-machine collaborative foreign language pedagogy is taking shape (Xu, 2026); a systematic review shows that empirical studies of generative AI in foreign language teaching have surged in number while the evidence remains fragmented (Pan et al., 2026), and the tension between technological optimism and pedagogical caution is far from resolved (Zheng, 2024). Speaking is widely recognized as the weakest skill of Chinese learners of English and the area in which generative AI is most expected to help: domestic studies have reported positive outcomes of generative AI-empowered speaking instruction (Wu et al., 2024), a systematic review of chatbot-supported English speaking instruction points to its potential for increasing practice opportunities and alleviating anxiety (Ge & Zheng, 2025), and classroom evidence on speaking instruction assisted by large language models has begun to appear (Wu, 2025). International research places learner emotion at the core of AI-supported speaking outcomes: controlled comparisons have examined the relationships between conversational generative AI chatbots and willingness to communicate, speaking anxiety, and self-perceived communicative competence (Wang et al., 2024); quasi-experimental evidence on AI-mediated interaction, speaking skills, and willingness to communicate is accumulating (Fathi et al., 2024; Wu et al., 2026); and a recent meta-analysis has synthesized the experimental effects of generative AI on positive and negative foreign language emotions (Chen et al., 2026).

The existing evidence, however, has three limitations. First, most studies operationalize "AI" as the use of a single tool over a few weeks, whereas in real classrooms AI is more likely to be embedded in the teaching process as several com-

ponents across an entire semester; comparisons between semester-long, multi-component teaching ecosystems and conventional teaching remain absent (Huang et al., 2022; Lee et al., 2026). Second, outcome measures usually target a single construct and rarely place negative emotion, positive emotion, behavioral intention, classroom engagement, and AI literacy within one comparative framework, which leaves unanswered the theoretically more consequential question of which type of variable is most likely to show a difference under an AI teaching ecosystem. Third, where out-of-class AI use is already widespread (Liu et al., 2025), a "no AI" control condition can hardly be established, and interest in chatbot partners is subject to novelty decay (Fryer et al., 2017); whether the positive findings of short-term single-tool studies transfer to a semester-long course ecosystem has not been tested.

Against this background, the present study draws on the natural teaching arrangement of an English Audio-Visual-Oral course to compare the emotions and participation of English majors under two conditions, an "AI teaching ecosystem" and conventional teaching, and addresses three questions: (1) whether students under the two conditions differ observably in speaking anxiety and oral willingness to communicate; (2) whether they differ observably in foreign language enjoyment, classroom engagement, AI literacy, and final examination scores; and (3) how, within the AI-ecosystem group, students' acceptance of the AI activities relates to these indicators.

## 2. Theoretical Foundations

### 2.1. The Dual Structure of Foreign Language Emotions and the Asymmetry of Antecedents

The affective filter hypothesis holds that a low-anxiety affective environment is a precondition for language input to be internalized (Krashen, 1982), a proposition that has long underpinned "reducing anxiety" as the first-order goal of speaking instruction design. The positive psychology turn in second language acquisition research has shown, however, that the emotional landscape of the foreign language classroom is not anchored at the anxiety pole alone: foreign language enjoyment (FLE) is a positive emotion independent of foreign language classroom anxiety (FLCA), and the two are not two sides of one coin but two constructs with different antecedents and different functions (Dewaele & MacIntyre, 2014). The short form of the Foreign Language Enjoyment Scale distinguishes three dimensions, teacher appreciation, private enjoyment, and social enjoyment (Botes et al., 2021), and its Chinese validation shows the structure to be equally stable among Chinese learners (Li et al., 2018).

More important is the asymmetry of antecedents. Teacher and learner variables do not predict enjoyment and anxiety with equal strength: enjoyment is shaped mainly by environmental factors that teaching design can alter, such as the teacher, peers, and classroom atmosphere, whereas anxiety is rooted more deeply in learners' individual traits and prior experience and shows greater cross-situational stability

(Botes et al., 2022; Dewaele et al., 2018; Horwitz et al., 1986). It follows that a change of condition at the course level may be more detectable in enjoyment than in anxiety. This asymmetry provides the theoretical basis for interpreting the pattern of results in the present study, but it is not in itself a claim about temporal order, and the study does not test a process of change.

### 2.2. The Layered Structure of Willingness to Communicate and Situational Transfer

The pyramid model of second language willingness to communicate (WTC) places communicative behavior at the apex; its proximal antecedent is situation-specific communicative confidence, while its distal antecedents include motivation, interpersonal and intergroup climate, personality, and social context (MacIntyre et al., 1998). A model of WTC in Chinese English classrooms further confirms that communicative confidence is the most direct predictor of classroom WTC, with the classroom environment operating indirectly through confidence and motivation (Peng & Woodrow, 2010). Two implications follow. First, as a relatively stable disposition, WTC may not show a discernible between-group difference after one semester of altered course conditions. Second, the measurement of WTC is situation-specific: willingness to speak in class in front of real people is not the same situation as willingness to exchange text with a machine in private. The existing positive evidence comes mostly from voice-based or anthropomorphic chatbots in controlled, short-term pretest-posttest settings (Fathi et al., 2024; Wang et al., 2024; Wu et al., 2026); whether text-based AI dialogue practice transfers to oral classroom situations remains to be tested.

### 2.3. The Constitutive Logic of the AI Teaching Ecosystem and the "Shared Experience" Hypothesis

The "AI teaching ecosystem" examined here is the course-level realization, in the English Audio-Visual-Oral course, of the project team's three-way collaborative model of professional teacher generation, AI generation, and student generation (the PAU model): the teacher sets the rules and quality standards, AI generates dialogues, feedback, and resources at scale according to those rules, and students become producers of content and viewpoints through use. The theoretical basis of the model and the process of resource construction are reported elsewhere; this paper treats the model only as a teaching condition taken as a whole. The ecosystem comprises three generative AI components, delivered through the course's online smart course, each resting on its own mechanistic assumption. The AI text-dialogue tasks translate the affective filter hypothesis and the notion of scaffolding (Krashen, 1982; Vygotsky, 1978) into behavioral rules for the AI (a peer "classmate" persona, short outputs, questions accompanied by sentence frames, recasting instead of direct correction, and automatic lowering of difficulty for students with low output), providing low-risk production practice. The AI micro-lectures provide explanations of knowledge points that can be watched repeatedly and serve

as cognitive support. The AI review songs adapt unit vocabulary and sentence patterns into songs that are played to the whole class before lessons, reducing the tedium of review and creating a shared class-level experience.

This gives rise to the "shared experience" hypothesis of the present study. The social enjoyment dimension measures peer relations, shared laughter, and shared classroom fun; from the outset it was derived from the social sources of enjoyment in learners' narratives (Botes et al., 2021; Dewaele & MacIntyre, 2014). AI activities organized around common course content, songs played collectively before class, dialogue tasks prompted by shared social topics, and dialogue anecdotes that can be retold among peers, may convert individual human-machine interaction into class-level shared experience and thereby relate more directly to social enjoyment, whereas low-threshold practice and immediate feedback are more likely to relate to private enjoyment. It should be noted that "ecosystem" here denotes a combination of course conditions rather than a theoretical stance and differs from recent work on the ecological coupling of AI, teachers, peers, and environment (Ge et al., 2026); nor can this study identify the independent contribution of each component, since it can only compare the ecosystem as a whole with conventional teaching.

## 3. Research Design

### 3.1. Participants and Teaching Conditions

Participants were students in six parallel classes of the English Audio-Visual-Oral course for the 2025 cohort of English majors at an application-oriented foreign-language university in Zhejiang Province, China, with 187 students on the class rosters. The six classes were scheduled by the academic affairs system without random assignment. The AI-ecosystem group comprised five classes (156 students), all taught by the author; the conventional group comprised one class (31 students) taught by another teacher. The two groups shared the same textbook, the same teaching schedule, the same credits, total class hours, grading weights (continuous assessment 60% plus final examination 40%), and final examination arrangements, and sat the same uniformly set Paper A final examination (all objective items, scored automatically by the learning platform). Because the actual composition of continuous assessment differed between the groups (**Table 1**), only the uniformly set final examination score is compared here. The conventional group was taught according to the course syllabus, whose listed methods are lecturing, pre-class presentations, group discussion, group drills, and role play, with no in-course AI component; its homework was mainly textbook-based, with one or two oral extension assignments, such as presentations, during the semester.

Within the same teaching schedule, the AI-ecosystem group was taught through an online smart course that the teacher built on the Chaoxing Xuexitong platform, with three components embedded throughout the main teaching weeks of the second semester of the 2025-2026 academic year (**Table 1**). (a) AI dialogue extension

tasks: 14 situational dialogue tasks, roughly one per week, completed after class through the "AI Practice" function of the platform, which runs on the platform's own built-in large language model (the Fanya model). Students typed their turns in English. The required number of turns rose from 10 to 15 and then to 20 over the semester; each task ended with an AI-generated evaluation in which the number of turns accounted for about a quarter of the score. In the final task, students averaged 23 turns and a score of 98%, and a typical student turn ran to about 10 to 20 words. The platform records when a task is opened but not the time spent on it; by the teacher's estimate, one task took about 10 to 15 minutes. The AI Practice scores counted as homework and made up 13% of continuous assessment, that is, 7.8% of the final grade. (b) AI micro-lectures: a series of teacher-designed, AI-assisted videos available for repeated viewing, graded together with the other chapter resources as chapter task points, which made up 8% of continuous assessment. (c) AI review songs: songs generated by AI from unit vocabulary and sentence patterns and played to the whole class before lessons. The course knowledge graph on the platform was built only after students had worked through the corresponding resources; the cohort studied here therefore did not use it, and it is not part of the condition examined. Out-of-class independent AI use by students in the conventional group was unrestricted. The teaching condition was applied at the class level, so teacher and condition cannot be separated; the author was both the teacher of the AI-ecosystem classes and the designer of the intervention, and the resulting problem of demand characteristics is addressed in the Discussion.

**Table 1.** Teaching conditions in the two groups.

| Feature | AI-ecosystem classes (5 classes, 156 students) | Conventional class (1 class, 31 students) |
|---|---|---|
| Teacher | The author | Another teacher of the course |
| Shared framework | Same textbook, teaching schedule, credits, and class hours; final grade = continuous assessment 60% + final examination 40%; same uniformly set final examination | Same |
| AI dialogue tasks | 14 situational dialogue tasks in the platform's "AI Practice" function (built-in Fanya model), about one per week after class; typed English; required turns rising from 10 to 15 to 20; AI-generated evaluation, with the turn count about a quarter of the score; graded as homework, 13% of continuous assessment (7.8% of the final grade); about 10 - 15 minutes per task (teacher's estimate) | None |
| AI micro-lectures | Teacher-designed, AI-assisted videos for repeated viewing; graded within chapter task points (8% of continuous assessment, together with other chapter resources) | None |
| AI review songs | AI-generated songs on unit vocabulary and sentence patterns, played to the whole class before lessons | None |

**Continued**

| | | |
|---|---|---|
| Oral or extension homework | The 14 AI dialogue tasks | One or two oral extension assignments, such as presentations; otherwise mainly textbook-based homework |
| Course knowledge graph | Built after students had worked through the corresponding resources; not used by this cohort | None |
| Out-of-class AI use | Not restricted | Not restricted |

Note: Percentages of continuous assessment are those set in the AI-ecosystem classes; the component weights used in the conventional class were set by its teacher and were not available to the author.

### 3.2. Instruments

The questionnaire used a two-form (A/B) design. Both forms measured background information, classroom engagement, speaking anxiety, foreign language enjoyment, oral WTC, frequency of everyday AI use, and AI literacy, with identical wording for the shared scale items; Form A additionally included technology acceptance, evaluations of the three components, self-reported task completion, and two open-ended questions, while Form B ended with two general open-ended questions. WTC items were scored from 1 = completely unwilling to 5 = very willing; all other scale items were scored from 1 = completely disagree to 5 = completely agree. Sources, item counts, and reliabilities in this sample are given in **Table 2**.

**Table 2.** Instruments and reliabilities in this sample.

| Construct | Source | Items | α (pooled/AI ecosystem/conventional) |
|---|---|---|---|
| Foreign language enjoyment (FLE) | S-FLES (Botes et al., 2021); Chinese context (Li et al., 2018) | 9 | 0.894/0.895/0.857 |
| Teacher appreciation | | 3 | 0.905/0.902/0.933 |
| Private enjoyment | | 3 | 0.830/0.812/0.885 |
| Social enjoyment | | 3 | 0.831/0.839/0.704 |
| Classroom engagement (LCES) | Eerdemutu et al. (2024) | 9 | 0.939/0.939/0.937 (behavioral/emotional/cognitive 0.80 - 0.90) |
| Oral willingness to communicate (WTC) | adapted from Peng and Woodrow (2010) | 6 | 0.888/0.887/0.902 (meaning-focused 4 items 0.870) |
| Speaking anxiety | single FLCAS item (Horwitz et al., 1986); see text | 1 | (not estimable for a single item) |
| AI literacy (five-item self-report index) | adapted from AILS-CCS (Ma & Chen, 2024): awareness 1, usage 2, evaluation 2 | 5 | 0.878/0.891/0.733; CR = 0.882 |
| Technology acceptance (AI-ecosystem group only) | PU/PEOU adapted from Davis (1989) and Zou and Huang (2023) | 3 + 3 | PU 0.857; PEOU 0.917 |
| Component evaluations (AI-ecosystem group only) | adapted by the author (satisfaction and perceived usefulness, 1 item each × 3 components) | 6 | within-component r = 0.77 - 0.84 |

Three points require explanation. First, the written analysis plan specified speaking anxiety, measured by an eight-item subset of the FLCAS (Horwitz et al., 1986) targeting speaking up in class, with reference to the Chinese revision (Wang, 2003), as a primary outcome. Seven of the eight items were lost when the questionnaire was imported into the platform, and only one item, "I never feel quite sure of myself when I am speaking in my English class", was administered. Using this single item in place of the planned scale is a deviation from the analysis plan; the decision was made after data collection but before any between-group comparison was computed (Section 3.3). The item is interpreted only, and in a limited way, as a single-item self-report of students' sense of assurance when speaking in class; the lost items included the only reverse-scored item, so the questionnaire contained no reverse-scored items at all. Second, because the two groups had different teachers, the teacher appreciation dimension of the enjoyment scale may carry teacher-specific differences; enjoyment is therefore always reported by dimension, and teacher appreciation is excluded from the main discussion. An exploratory factor analysis of the pooled sample showed that the first two eigenvalues of the nine S-FLES items exceeded 1 (5.12 and 1.27); in the rotated three-factor solution, teacher appreciation and social enjoyment loaded clearly on their own factors, while the three private enjoyment items also had moderate loadings on the teacher appreciation factor; inter-dimension correlations were 0.49 - 0.70, and dimension-level reporting was retained on this basis. Third, the AI literacy index uses five of the 15 AILS-CCS items and does not cover the ethics dimension; a single-factor test of the five items supports combining them into one index (first factor explaining 67.9% of the variance, $\lambda 1/\lambda 2 = 6.1$, loadings 0.69 - 0.83). In accordance with the analysis plan, AI literacy is an outcome rather than a covariate: it was measured after the intervention and is precisely a capacity that the intervention might train.

### 3.3. Data Collection and Analytic Strategy

The questionnaire was distributed to the two groups separately at the end of the semester through the Chaoxing Xuexitong platform by the author, with platform points awarded on completion; the teacher of the conventional group was informed and agreed to its distribution in that course space. Exported data were coded in place of student identifiers and used only for research. Data-cleaning rules were fixed before modeling: respondents who failed the attention-check item and those whose standard deviation across the 30 shared scale items was zero were excluded.

The analysis plan was fixed in writing before data collection. (a) The plan specified two primary outcomes, speaking anxiety (the eight-item FLCAS subset) and the WTC total score, without multiplicity correction; because only one anxiety item was administered, anxiety was analyzed as a single item, a substitution decided after data collection but before any between-group comparison and reported here as a deviation from the plan. The other outcomes formed a family of

ten tests (three enjoyment dimensions; engagement total and three dimensions; the meaning-focused WTC score; AI literacy; final examination score), for which both Holm and Benjamini-Hochberg false discovery rate (BH-FDR) corrections are reported, with Holm as the primary criterion. (b) Covariates were restricted to those determined before analysis: the English score band of the gaokao (national college entrance examination; four bands coded 1 - 4 from low to high, with one valid case of "did not take/do not remember" imputed by mapping from self-rated proficiency at enrollment) and gender (added by matching student numbers to class rosters). Major, although listed in the plan, was not entered into the models because only three students in the sample were non-English majors. (c) Every outcome is reported with both ANCOVA and Welch's t test, with Hedges' g and its 95% confidence interval (CI) as the main axis; under the approximation that student observations are independent, the minimum detectable effect at $\alpha = 0.05$ and 80% power is about $d = 0.56$, so the tests have limited power for effects smaller than this, and non-significant results are uniformly described as "no detectable difference". (d) Because the teaching condition was applied at the class level and the conventional condition comprised only one class, class-level standard errors cannot be estimated; the *p* values and confidence intervals reported below come from student-level models and represent only the modeled contrast in this sample. These *p* values assume that students are independent observations; with one class in the conventional condition, that assumption cannot be examined for the condition contrast, so the *p* values should not be read as condition-level evidence. Class clustering is described through the ICC (1) of each outcome within the five AI-ecosystem classes and through the means of the six classes. (e) Common method bias was checked descriptively with Harman's single-factor test; open-ended responses serve only as a descriptive supplement based on keyword categorization. Gender and the final examination score were pre-planned and merged after the primary outcome results had been produced; the pre-enrollment administrative scores that the plan preferred could not be obtained, so the covariate is the self-reported gaokao band; task completion counts are self-reported.

## 4. Results

### 4.1. Sample and Baseline Comparability

A total of 180 questionnaires were returned (AI-ecosystem group 149, response rate 95.5%; conventional group 31, response rate 100%). Six respondents failed the attention-check item (one of whom was also a straight-liner), leaving a valid sample of 174 (144/30). Missing values on closed items were rare; scale scores were computed as the personal mean for respondents who answered at least 80% of the items. Table 3 shows no statistically detectable differences between the groups in gaokao English band, self-rated English proficiency at enrollment, or gender composition, although these measured variables are insufficient to establish pre-intervention equivalence. Frequency of everyday AI use was measured at the end of the semester with different stems on the two forms (Form A restricted it to use "out-

side the learning activities of this course"; Form B did not) and is therefore only described separately: 91.0% of the AI-ecosystem group reported using AI several times a week or more outside the course activities, and 90.0% of the conventional group reported using AI several times a week or more in everyday life. Frequent AI contact existed in both groups; the study compares two teaching conditions, "AI systematically embedded within the course" versus "conventional teaching", not AI users versus non-users.

**Table 3.** Distribution of background variables in the two groups (valid sample: AI ecosystem 144, conventional 30).

| Variable | Category | AI ecosystem n (%) | Conventional n (%) | Test |
|---|---|---|---|---|
| Gaokao English score | 120 and above | 30 (21%) | 7 (23%) | $\chi^2$ (2) = 0.08, $p$ = 0.959[a] |
| | 105 - 119 | 78 (54%) | 16 (53%) | |
| | 90 - 104 | 30 (21%) | 7 (23%) | |
| | 89 and below | 5 (3%) | 0 (0%) | |
| | Did not take/do not remember | 1 (1%) | 0 (0%) | |
| Self-rated English at enrollment | Weak | 33 (23%) | 7 (23%) | $\chi^2$ (2) = 0.05, $p$ = 0.977 |
| | Average | 89 (62%) | 18 (60%) | |
| | Good | 22 (15%) | 5 (17%) | |
| Everyday AI use (Form A excludes course activities; Form B unrestricted) | Never | 1 (1%) | 0 (0%) | (different stems; not tested) |
| | Occasionally | 12 (8%) | 3 (10%) | |
| | Several times a week | 60 (42%) | 15 (50%) | |
| | Almost every day | 71 (49%) | 12 (40%) | |
| Gender | Female | 117 (81%) | 23 (77%) | $\chi^2$ (1) = 0.10, $p$ = 0.747; Fisher $p$ = 0.614 |
| | Male | 27 (19%) | 7 (23%) | |

Note: Percentages are rounded; gender was added by matching student numbers to class rosters. [a]The gaokao test collapses the bands into three (120 and above; 105 - 119; 104 and below), with "did not take/do not remember" treated as missing.

### 4.2. Speaking Anxiety and Willingness to Communicate

Neither primary outcome showed a between-group difference (Table 4). On the single speaking anxiety item, the mean of the AI-ecosystem group exceeded that of the conventional group by 0.22 points, g = 0.23 (95% CI [−0.16, 0.62]); after covariate adjustment, b = 0.20 (95% CI [−0.17, 0.56], $p$ = 0.286), and the Mann-Whitney test agreed ($p$ = 0.326). The WTC total scores of the two groups were close, g = 0.05 (95% CI [−0.34, 0.45]), adjusted b = 0.04 ($p$ = 0.777). The group × gaokao band interactions were non-significant for both outcomes ($p$ = 0.213 and

0.172). The confidence intervals of both outcomes span about 0.8 standard deviations and are compatible with small-to-medium effects as well as with zero; the point estimate for anxiety is slightly higher in the AI-ecosystem group, but given the single-item measure and the wide interval, no substantive interpretation is placed on its direction.

**Table 4.** Between-group comparison of the primary outcomes.

| Outcome | AI ecosystem M ± SD (n = 144) | Conventional M ± SD (n = 30) | Welch t (df) | $p$ | Hedges' g [95% CI] | ANCOVA b [95% CI] | $p$ | partial $\eta^2$ |
|---|---|---|---|---|---|---|---|---|
| Speaking anxiety (single item) | 3.28 ± 0.95 | 3.07 ± 0.91 | 1.19 (43.3) | 0.241 | 0.23 [−0.16, 0.62] | 0.20 [−0.17, 0.56] | 0.286 | 0.007 |
| WTC total score | 3.58 ± 0.74 | 3.54 ± 0.70 | 0.27 (43.6) | 0.785 | 0.05 [−0.34, 0.45] | 0.04 [−0.24, 0.33] | 0.777 | 0.000 |

Note: ANCOVA covariates are gaokao English band and gender; the analysis plan set no overall success criterion across the two outcomes, which are interpreted separately.

### 4.3. Foreign Language Enjoyment, Classroom Engagement, AI Literacy, and Final Examination Scores

The other outcomes displayed a clear pattern (Table 5). On the private and social enjoyment dimensions, the means of the AI-ecosystem group were higher, with effect sizes of g = 0.62 and 0.83; in student-level models, both differences exceeded the significance threshold under Holm and BH corrections for Welch's test and ANCOVA; the Mann-Whitney tests agreed ($p = 0.003$ and $p < 0.001$). Teacher appreciation showed no detectable difference (g = 0.20) and is excluded from the main discussion under the analysis plan. The point estimates for all engagement dimensions were consistently positive (g = 0.21 - 0.34), but none survived correction; the meaning-focused WTC score, AI literacy, and the final examination score showed no detectable differences. In the full examinee sample (155 in the AI-ecosystem group and 31 in the conventional group), the correct rate on textbook-based items was 82.6% versus 82.7% (difference −0.13 percentage points, 95% CI [−4.05, 3.79]), and on listening items at the level of the College English Test Band 6 (CET-6) it was 53.3% versus 49.8% (difference +3.51 percentage points, 95% CI [−2.64, 9.66]); both intervals are wide, and no conclusion can be drawn as to whether the AI activities affected textbook learning.

**Table 5.** Between-group comparison of the other outcomes (family of ten tests with correction).

| Outcome | AI ecosystem M ± SD | Conventional M ± SD | Welch t (df) | g [95% CI] | Welch $p$ | Holm $p$ | BH q | ANCOVA b [95% CI] | ANCOVA corrected $p$ (Holm/BH) | partial $\eta^2$ |
|---|---|---|---|---|---|---|---|---|---|---|
| FLE: teacher appreciation | 4.55 ± 0.58 | 4.43 ± 0.47 | 1.16 (49.2) | 0.20 [−0.19, 0.60] | 0.250 | 1.000 | 0.357 | 0.12 [−0.10, 0.34] | 1.000/0.365 | 0.007 |
| **FLE: private enjoyment** | 4.47 ± 0.53 | 4.14 ± 0.52 | 3.16 (42.7) | **0.62 [0.22, 1.02]** | 0.003 | **0.026** | **0.014** | 0.33 [0.12, 0.53] | **0.016/0.009** | 0.056 |

**Continued**

| | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| **FLE: social enjoyment** | 4.16 ± 0.69 | 3.59 ± 0.62 | 4.47 (45.3) | **0.83 [0.43, 1.24]** | <0.001 | **<0.001** | **<0.001** | 0.57 [0.30, 0.84] | **<0.001/<0.001** | 0.092 |
| Engagement: total | 4.01 ± 0.65 | 3.82 ± 0.53 | 1.69 (49.0) | 0.30 [−0.10, 0.69] | 0.098 | 0.683 | 0.224 | 0.19 [−0.06, 0.44] | 0.901/0.321 | 0.014 |
| Engagement: behavioral | 3.93 ± 0.73 | 3.78 ± 0.60 | 1.19 (48.9) | 0.21 [−0.19, 0.60] | 0.240 | 1.000 | 0.357 | 0.15 [−0.12, 0.43] | 1.000/0.365 | 0.007 |
| Engagement: emotional | 4.08 ± 0.73 | 3.88 ± 0.58 | 1.62 (49.9) | 0.28 [−0.11, 0.67] | 0.112 | 0.683 | 0.224 | 0.20 [−0.08, 0.47] | 0.963/0.321 | 0.012 |
| Engagement: cognitive | 4.03 ± 0.65 | 3.81 ± 0.57 | 1.84 (46.0) | 0.34 [−0.06, 0.73] | 0.072 | 0.575 | 0.224 | 0.22 [−0.03, 0.47] | 0.629/0.262 | 0.018 |
| WTC: meaning-focused | 3.52 ± 0.77 | 3.49 ± 0.75 | 0.16 (43.0) | 0.03 [−0.36, 0.42] | 0.875 | 1.000 | 0.875 | 0.03 [−0.27, 0.33] | 1.000/0.859 | 0.000 |
| AI literacy (five items) | 4.04 ± 0.57 | 4.00 ± 0.37 | 0.47 (62.7) | 0.07 [−0.32, 0.46] | 0.639 | 1.000 | 0.710 | 0.05 [−0.16, 0.26] | 1.000/0.709 | 0.001 |
| Final examination score (/100) | 68.3 ± 11.7 | 66.6 ± 10.6 | 0.77 (44.9) | 0.14 [−0.25, 0.54] | 0.445 | 1.000 | 0.556 | 2.24 [−1.88, 6.36] | 1.000/0.365 | 0.007 |

Note: Bold marks results that reach the significance threshold after correction in student-level models (Holm primary, BH secondary); ANCOVA covariates are gaokao English band and gender; the valid sample for private enjoyment is 143/30 and for the other outcomes 144/30; the final examination score is out of 100, all other outcomes are on 5-point scales.

Because the social enjoyment items refer to the class as a whole and thus describe a class-level attribute, the class means deserve attention (Table 6). Within the five AI-ecosystem classes, the ICC (1) point estimates of all outcomes were close to zero (anxiety −0.002, WTC −0.020, private enjoyment −0.001, social enjoyment −0.003, teacher appreciation 0.002, engagement −0.009); the private enjoyment (4.14) and social enjoyment (3.59) of the conventional class were lower than those of all five AI-ecosystem classes (4.35 - 4.62 and 4.04 - 4.32 respectively). The homogeneity of the five classes is compatible both with "the same teaching ecosystem" and with "the same teacher", so the class-level description cannot rule out teacher or class explanations, but it shows that the between-group differences on the two enjoyment dimensions were not driven by any single AI-ecosystem class. Near-zero ICCs among the five AI-ecosystem classes do not, however, show that the student-level $p$ values are valid for the condition contrast: the conventional condition contains a single class, so its between-class variance cannot be estimated. If the six classes are treated as the units of analysis, a randomization test cannot yield a one-sided $p$ value below 1/6 (about 0.17); that is the value obtained here, because the conventional class had the lowest private- and social-enjoyment means of the six classes.

Table 6. Means of the six classes on the outcomes (valid sample).

| Class | n | Speaking anxiety | WTC total | Private enjoyment | Social enjoyment | Teacher appreciation | Engagement total |
|---|---|---|---|---|---|---|---|
| AI-ecosystem class 1 | 31 | 3.19 | 3.49 | 4.45 | 4.12 | 4.58 | 4.03 |
| AI-ecosystem class 2 | 28 | 3.43 | 3.62 | 4.46 | 4.32 | 4.55 | 4.12 |
| AI-ecosystem class 3 | 25 | 3.20 | 3.53 | 4.35 | 4.05 | 4.37 | 3.82 |
| AI-ecosystem class 4 | 30 | 3.50 | 3.72 | 4.62 | 4.04 | 4.69 | 4.01 |
| AI-ecosystem class 5 | 30 | 3.10 | 3.56 | 4.46 | 4.26 | 4.52 | 4.05 |
| Conventional class | 30 | 3.07 | 3.54 | 4.14 | 3.59 | 4.43 | 3.82 |

Note: The private enjoyment score of AI-ecosystem class 3 is based on 24 students.

### 4.4. Robustness Checks and Response Style

Four sets of sensitivity analyses left the judgment of "no detectable difference" on the two primary outcomes unchanged. (a) When all 180 questionnaires were included, private and social enjoyment remained the two largest between-group contrasts (g = 0.50 and 0.85); after recomputing the ten-test family corrections in student-level models, social enjoyment still passed (Holm and BH both < 0.001), whereas private enjoyment did not (Welch: Holm $p = 0.109$, BH q = 0.061; ANCOVA: Holm $p = 0.099$, BH q = 0.055). (b) Further excluding students attending from other cohorts or majors (N = 169) did not change the conclusions. (c) Replacing the gaokao band with self-rated proficiency (with gender) gave b = 0.21 ($p = 0.231$) for anxiety and b = 0.03 ($p = 0.837$) for WTC; with the gaokao band alone (without gender), the two coefficients were 0.20 and 0.05. (d) Using the full examinee sample for the final examination score (155/31, including six students who did not submit the questionnaire) gave g = 0.15 (95% CI [−0.24, 0.53], $p = 0.431$). In Harman's single-factor test, the first factor explained 41.5% of the variance, so common method bias cannot be ruled out.

The two groups differed in response style: across all agreement-anchored items, 35.3% of the responses of the AI-ecosystem group were "completely agree" compared with 17.6% in the conventional group, and on items unrelated to enjoyment (classroom engagement, AI literacy, and the anxiety item) the figures were 26.2% versus 12.2%. When the proportion of extreme responses on non-enjoyment items was added as an exploratory covariate, the estimated contrasts decreased: social enjoyment from 0.57 to 0.39 (95% CI [0.15, 0.62]) and private enjoyment from 0.33 to 0.19 (95% CI [0.02, 0.37]). Because this index is derived from the same post-intervention responses as the outcomes, the adjustment cannot separate response bias from substantive differences in experience.

### 4.5. Acceptance of the AI Activities and Learning Indicators within the AI-Ecosystem Group

Within the AI-ecosystem group (Table 7), perceived usefulness, perceived ease of use, and the three component evaluations correlated positively, from weakly to

strongly, with WTC, enjoyment, engagement, and AI literacy (r = 0.25 - 0.74, $p < 0.01$), with perceived usefulness most strongly associated with engagement (r = 0.69) and AI literacy (r = 0.74); all acceptance and evaluation indicators correlated only weakly and non-significantly with the speaking anxiety item (r = −0.01 to −0.13). Self-reported task completion was almost unrelated to any indicator ($|r| \leq 0.16$), but 74.3% of the students reported completing all 14 tasks, so the variance of this variable is severely restricted. Descriptive categorization of the open-ended responses showed that the AI dialogue extension task was the most frequently mentioned AI activity in all five classes; students mentioned lower pressure to express themselves, timely feedback, and the possibility of repeated attempts, and also raised concerns about over-reliance and inaccurate AI output. Responses from the conventional group referred to activities such as speeches, reading aloud, group discussion, and listening practice.

**Table 7.** Correlations of AI activity acceptance and component evaluations with learning indicators within the AI-ecosystem group (pairwise n = 140 - 144).

| | Speaking anxiety | WTC total | Private enjoyment | Social enjoyment | Engagement total | AI literacy |
|---|---|---|---|---|---|---|
| Perceived usefulness (PU) | −0.08 | 0.59*** | 0.52*** | 0.57*** | 0.69*** | 0.74*** |
| Perceived ease of use (PEOU) | −0.12 | 0.47*** | 0.25** | 0.43*** | 0.58*** | 0.73*** |
| AI dialogue task evaluation | −0.13 | 0.56*** | 0.43*** | 0.54*** | 0.66*** | 0.63*** |
| AI micro-lecture evaluation | −0.01 | 0.49*** | 0.45*** | 0.64*** | 0.69*** | 0.55*** |
| AI review song evaluation | −0.04 | 0.55*** | 0.47*** | 0.52*** | 0.64*** | 0.60*** |
| Self-reported task completion | −0.03 | 0.11 | 0.16 | 0.08 | 0.14 | −0.01 |

Note: *$p < 0.05$, **$p < 0.01$, ***$p < 0.001$ (uncorrected); correlations are computed on the non-missing cases of each pair of variables.

## 5. Discussion

### 5.1. Asymmetric Antecedents and Intervention Modality: Interpreting the Null Primary Results

Neither primary outcome showed a between-group difference, whereas the two enjoyment dimensions showed differences that were stable across the student-level analyses. This pattern can first be explained by the asymmetry of antecedents set out in the theoretical foundations: enjoyment is shaped by environmental factors such as the teacher, peers, and classroom atmosphere and is the variable that a change in course conditions is most likely to reach, whereas anxiety is rooted in learners' individual traits and WTC has communicative confidence as its proximal antecedent, so neither necessarily shows a discernible between-group change within one semester (Botes et al., 2022; Dewaele et al., 2018; MacIntyre et al., 1998). The results of this study are consistent with this asymmetry. It must be made clear, however, that this is an interpretation of the pattern obtained from a single end-of-semester measurement, not an inference about the order of change.

The second explanation is the mismatch between the modality of the interven-

tion and the situation of the outcome measures. Of the three components, the one closest to speaking, the AI dialogue task, used a text modality; students typed their turns, and the course offered no spoken interaction with the AI, whereas the speaking anxiety item and all six WTC situations refer to speaking in class in front of real people. Studies reporting positive results have mostly used voice-based or anthropomorphic chatbots in controlled, short-term pretest-posttest settings (Fathi et al., 2024; Wang et al., 2024; Wu et al., 2026), and the meta-analytic evidence likewise consists mainly of such experimental studies (Chen et al., 2026). The present study is a semester-long intact-class comparison on trait-like outcomes; it answers a different question from those studies, and the results cannot be converted directly. This mismatch also means that the expectation raised in the Introduction, that an AI dialogue partner would alleviate speaking anxiety, lacked a direct pathway under a text modality in the first place. The within-group data are consistent with this for anxiety, which correlated near zero with all acceptance indicators, but not for WTC: students who rated the AI activities more favorably also reported higher WTC ($r = 0.47$ - $0.59$), as well as higher enjoyment and engagement. For WTC, a null between-group contrast therefore coexists with a positive within-group association, which may reflect a genuine link among students who engaged with the tasks, shared method variance, or self-selection; the present design cannot distinguish among these.

In addition, with only 30 students in the conventional group, the minimum detectable effect under the independent-student approximation is about $d = 0.56$, and the confidence intervals of both primary outcomes are compatible with small-to-medium effects; because the AI-use items differed between the two forms, the direction and magnitude of the influence of other AI exposure on the between-group differences cannot be determined. The conclusion on the primary outcomes therefore stops at "no detectable difference" rather than "no difference between the two teaching conditions".

### 5.2. "Shared Experience" and Social Enjoyment: The Collective Dimension of the AI Teaching Ecosystem

The between-group difference in social enjoyment ($g = 0.83$) is the most stable finding of this study: in student-level models it remained significant after correction under every sensitivity specification, and the conventional class scored lower than all five AI-ecosystem classes. Judged by the content of the social enjoyment dimension, this difference points to the collective mood and shared experience of the class rather than to individual satisfaction with the course. The AI teaching ecosystem provided design preconditions for such shared experience: the AI review songs played to the whole class before lessons were common events experienced by everyone at the same time, the dialogue tasks prompted by social topics created common topics to talk about after class, and amusing moments in the dialogues could be retold among peers. Foreign language enjoyment research has from the outset regarded peers, laughter, and group belonging as social sources of

enjoyment (Dewaele & MacIntyre, 2014), and the social enjoyment dimension of the S-FLES is precisely the operationalization of that source (Botes et al., 2021). On this basis, the "shared experience pathway" is a hypothesis consistent with the pattern of the data and directly testable through multi-wave measurement, classroom interaction records, and comparisons between individual and shared AI tasks; because this study did not record whether students actually shared such content, the pathway is advanced as a hypothesis to be tested rather than as a confirmed mechanism.

The difference in private enjoyment ($g = 0.62$) is compatible with the design of low-threshold practice and immediate feedback: the rule of the AI dialogue task that "speaking for even one minute counts as success", together with the immediate evaluation, makes it easier for students to experience "learning something interesting" and "feeling proud of my progress". The private enjoyment difference, however, did not survive the family correction when the excluded cases were included; it is less stable than social enjoyment and should be interpreted more cautiously.

### 5.3. Alternative Explanations and the Boundaries of the Evidence

The above interpretations must be set alongside four alternative explanations. The first is the confounding of teacher and class. The conventional condition had only one class, taught by a different teacher; teacher style, class group dynamics, and selection differences could each account for all of the between-group results. The absence of a detectable difference in teacher appreciation ($g = 0.20$, 95% CI [−0.19, 0.60]) cannot serve as evidence against teacher confounding: the confidence interval of that dimension is still compatible with a medium-sized difference, and the dimension is excluded from the main discussion under the analysis plan. The second is response style. The proportion of "completely agree" responses in the AI-ecosystem group was about twice that in the conventional group, and this tendency was equally present on items unrelated to enjoyment; because the author was also the teacher of the AI-ecosystem classes and organized the administration in both groups, demand characteristics and social desirability may have inflated positive evaluations asymmetrically in the AI-ecosystem group. Exploratory adjustment for response style reduced the estimated contrasts, more for private than for social enjoyment; because the adjustment variable comes from the same post-intervention responses, it cannot separate response bias from substantive differences in experience, and since the questionnaire contained no reverse-scored items, this alternative explanation cannot be excluded. The third is the temporal nature of a single measurement. One end-of-semester measurement cannot distinguish novelty effects (Fryer et al., 2017) from stable change, nor determine the sequence in which the indicators changed. The fourth is practice dose and grading incentives. The AI-ecosystem classes completed 14 graded after-class dialogue tasks worth 13% of continuous assessment, whereas the conventional class had one or two oral extension assignments; the observed differences may therefore partly re-

flect additional practice and assessment incentives rather than the AI components as such, and the design cannot separate the two. Taken together, all between-group results of this study can be interpreted only as associations; the differences on the two enjoyment dimensions are "stable observed differences consistent with the AI teaching ecosystem", not confirmed teaching effects.

The "no detectable difference" results for AI literacy and the final examination score are informative in their own right. Both groups reported high levels of AI literacy and everyday AI use, which suggests that where generative AI has already permeated students' daily lives, an in-course AI teaching ecosystem does not automatically translate into differences in general AI literacy. The final examination tested only listening comprehension and mastery of textbook content and included no oral production; it therefore cannot be used to evaluate the effect of the AI activities on speaking performance, nor can the absence of a detectable difference be taken to show that textbook learning was unaffected, since the interval for the textbook-item correct rate still allows a difference of about four percentage points.

### 5.4. Pedagogical Implications

Within the boundaries of the evidence, four implications can be proposed. First, evaluations of in-course AI integration should measure positive emotion by dimension in addition to anxiety, WTC, and achievement, and should report effect sizes alongside confidence intervals. This study shows that the detectability of the same teaching ecosystem can differ sharply across emotional constructs; judging success solely by whether students "dare to speak" or by examination scores would miss the emotional ecology of the classroom, the level most likely to be reached first. Second, the collective dimension of AI activities deserves deliberate design. AI songs played to the whole class before lessons, dialogue tasks built around common topics, and the display of excellent dialogues with names removed can all turn one-to-one human-machine interaction into something shared by the class; if later research confirms the "shared experience pathway" on the basis of recorded sharing behavior, this would be a line of AI foreign language classroom design distinct from "AI as personal tutor". Third, on this evidence, text-based AI dialogue tasks should not be expected to reduce anxiety about speaking in class. No between-group difference was detected in anxiety or WTC. Within the AI-ecosystem group, students who rated the AI activities more favorably reported higher WTC ($r = 0.47$ - $0.59$) but not lower anxiety ($r = -0.01$ to $-0.13$); because these correlations are cross-sectional, they cannot show that the tasks raised WTC. Voice-based tasks, or task designs linked to classroom speaking, are the natural next step. Fourth, the concerns about over-reliance and inaccurate AI output raised spontaneously by students suggest that norms for AI use, strategies for checking output, and teacher review of AI-generated content should be supplied alongside the AI activities; the evidence from this combined package also does not support the expectation that adopting any single component alone would yield the same results.

## 6. Conclusion and Limitations

In this semester-long intact-class comparison, the primary outcomes of speaking anxiety and oral willingness to communicate showed no detectable between-group differences; among the other outcomes, the AI-ecosystem group reported higher private and social enjoyment, differences that were significant in student-level models and, for social enjoyment, stable across all sensitivity analyses, while classroom engagement, AI literacy, and the final examination score showed no detectable differences; within the AI-ecosystem group, acceptance of the AI activities correlated positively with enjoyment, engagement, and WTC. Theoretically, this pattern is consistent with the asymmetry of antecedents between enjoyment on the one hand and anxiety and WTC on the other, and, for anxiety, with the absence of a direct pathway from text-based AI dialogue tasks to oral classroom situations; the "shared experience pathway" offers a testable explanation for the collective dimension of the AI teaching ecosystem. Because outcomes were measured only at the end of the semester, the conventional condition comprised only one class, and the teaching condition was confounded with class, teacher, additional graded practice, and differences in response tendency, the results reflect only associations among the classes observed here and do not support a causal interpretation of the AI teaching ecosystem.

The limitations of the study fall into three areas. In design, there was no pretest, the control consisted of a single intact class, teacher and condition were confounded, and the AI teaching ecosystem was a combined condition, including additional graded practice, whose components' contributions cannot be separated. In measurement, speaking anxiety was reduced to a single item in deviation from the analysis plan, the questionnaire contained no reverse-scored items, all questionnaire indicators were self-reported once at the end of the semester, and the final examination included no oral production. In sampling, the study was confined to first-year English majors at a single university. Future research should add pretests and measures of oral production, include at least two classes taught by different teachers under each teaching condition, record classroom sharing behavior and objective dosage, and compare text-based and voice-based AI dialogue tasks, so as to test the asymmetric-antecedents explanation and the "shared experience pathway" hypothesis proposed here.

## Acknowledgements

This work was supported by:

Zhejiang Yuexiu University Research Project "An Empirical Study on the Development and Optimization of Multi-modal Digital Resources for Ideological and Political Education in Foreign Language Courses Based on the PAU Model" (No. N2024014);

Zhejiang Province Second Batch of Provincial Undergraduate Education and Teaching Reform Projects during the 14th Five-Year Plan "'Artificial Intelligence + Cultural Intelligence': Research on the Digital-Intelligent Construction Path of Ide-

ological and Political Education in Foreign Language Courses" (No. JGBA2024598);

Zhejiang Higher Education Association 2026 Higher Education Research Project "Research on an Application Model for English Audio-Visual-Oral Teaching Integrating AI Knowledge Bases and Multi-layer Course Graphs" (No. KT2026511);

2025 Commissioned Research Project of the Institute of World Literature and Area & Country Studies, Zhejiang Yuexiu University "Research on the Multimodal Pragmatic Identity Construction and Communication Ecosystem of Zhejiang-Made Brands Going Global from the Perspective of 'AI + Teacher' Collaboration".

This paper is a research outcome of the above projects.

## Conflicts of Interest

The author declares no conflicts of interest regarding the publication of this paper.

## References

Botes, E., Dewaele, J. M., & Greiff, S. (2021). The Development and Validation of the Short Form of the Foreign Language Enjoyment Scale. *The Modern Language Journal, 105,* 858-876. https://doi.org/10.1111/modl.12741

Botes, E., Dewaele, J. M., & Greiff, S. (2022). Taking Stock: A Meta-Analysis of the Effects of Foreign Language Enjoyment. *Studies in Second Language Learning and Teaching, 12,* 205-232. https://doi.org/10.14746/ssllt.2022.12.2.3

Chen, S., Yu, L., Huang, R., Bai, X., & Qin, J. (2026). The Effects of Generative AI on Foreign Language Emotions: A Meta-Analysis. *Acta Psychologica, 268,* Article 107336. https://doi.org/10.1016/j.actpsy.2026.107336

Davis, F. D. (1989). Perceived Usefulness, Perceived Ease of Use, and User Acceptance of Information Technology. *MIS Quarterly, 13,* 319-340. https://doi.org/10.2307/249008

Dewaele, J. M., & MacIntyre, P. D. (2014). The Two Faces of Janus? Anxiety and Enjoyment in the Foreign Language Classroom. *Studies in Second Language Learning and Teaching, 4,* 237-274. https://doi.org/10.14746/ssllt.2014.4.2.5

Dewaele, J. M., Witney, J., Saito, K., & Dewaele, L. (2018). Foreign Language Enjoyment and Anxiety: The Effect of Teacher and Learner Variables. *Language Teaching Research, 22,* 676-697. https://doi.org/10.1177/1362168817692161

Eerdemutu, L., Dewaele, J. M., & Wang, J. (2024). Developing a Short Language Classroom Engagement Scale (LCES) and Linking It with Needs Satisfaction and Achievement. *System, 120,* Article 103189. https://doi.org/10.1016/j.system.2023.103189

Fathi, J., Rahimi, M., & Derakhshan, A. (2024). Improving EFL Learners' Speaking Skills and Willingness to Communicate via Artificial Intelligence-Mediated Interactions. *System, 121,* Article 103254. https://doi.org/10.1016/j.system.2024.103254

Fryer, L. K., Ainley, M., Thompson, A., Gibson, A., & Sherlock, Z. (2017). Stimulating and Sustaining Interest in a Language Course: An Experimental Comparison of Chatbot and Human Task Partners. *Computers in Human Behavior, 75,* 461-468. https://doi.org/10.1016/j.chb.2017.05.045

Ge, L., Xia, Y., & Yang, C. (2026). Beyond Tool Use: How Ecological Coupling Configures AI's Empowerment for Language Learning Engagement. *Frontiers in Psychology, 17,* Article 1747355. https://doi.org/10.3389/fpsyg.2026.1747355

Ge, N., & Zheng, C. P. (2025). A Systematic Review of Intelligent Chatbots Supporting English Speaking Instruction. *Foreign Language Education, 46,* 56-63.

He, N., & Wang, S. R. (2025). The Goals and Pathways of Building "New Foreign Languages" in the Era of Artificial Intelligence. *Foreign Language World, 2,* 11-15.

Horwitz, E. K., Horwitz, M. B., & Cope, J. (1986). Foreign Language Classroom Anxiety. *The Modern Language Journal, 70,* 125-132. https://doi.org/10.1111/j.1540-4781.1986.tb05256.x

Huang, W., Hew, K. F., & Fryer, L. K. (2022). Chatbots for Language Learning—Are They Really Useful? A Systematic Review of Chatbot-Supported Language Learning. *Journal of Computer Assisted Learning, 38,* 237-257. https://doi.org/10.1111/jcal.12610

Krashen, S. D. (1982). *Principles and Practice in Second Language Acquisition*. Pergamon Press.

Lee, S., Choe, H., Zou, D., & Jeon, J. (2026). Generative AI (GenAI) in the Language Classroom: A Systematic Review. *Interactive Learning Environments, 34,* 335-359. https://doi.org/10.1080/10494820.2025.2498537

Li, C., Jiang, G., & Dewaele, J. M. (2018). Understanding Chinese High School Students' Foreign Language Enjoyment: Validation of the Chinese Version of the Foreign Language Enjoyment Scale. *System, 76,* 183-196. https://doi.org/10.1016/j.system.2018.06.004

Liu, G. L., Darvin, R., & Ma, C. (2025). Exploring AI-Mediated Informal Digital Learning of English (AI-IDLE): A Mixed-Method Investigation of Chinese EFL Learners' AI Adoption and Experiences. *Computer Assisted Language Learning, 38,* 1632-1660. https://doi.org/10.1080/09588221.2024.2310288

Ma, S., & Chen, Z. (2024). The Development and Validation of the Artificial Intelligence Literacy Scale for Chinese College Students (AILS-CCS). *IEEE Access, 12,* 146419-146429. https://doi.org/10.1109/access.2024.3468378

Macintyre, P. D., Clément, R., Dörnyei, Z., & Noels, K. A. (1998). Conceptualizing Willingness to Communicate in a L2: A Situational Model of L2 Confidence and Affiliation. *The Modern Language Journal, 82,* 545-562. https://doi.org/10.1111/j.1540-4781.1998.tb05543.x

Pan, H. Y., Li, J. J., & Ma, X. Y. (2026). Applications and Research of Generative Artificial Intelligence in Foreign Language Teaching: A Systematic Review. *Foreign Language Teaching and Research, 58,* 250-262.

Peng, J. E., & Woodrow, L. (2010). Willingness to Communicate in English: A Model in the Chinese EFL Classroom Context. *Language Learning, 60,* 834-876. https://doi.org/10.1111/j.1467-9922.2010.00576.x

Vygotsky, L. S. (1978). *Mind in Society: The Development of Higher Psychological Processes*. Harvard University Press.

Wang, C. K. (2003). A Test Report of the Foreign Language Classroom Anxiety Scale (FLCAS) among College Students. *Journal of Psychological Science, 2,* 281-284.

Wang, C., Zou, B., Du, Y., & Wang, Z. (2024). The Impact of Different Conversational Generative AI Chatbots on EFL Learners: An Analysis of Willingness to Communicate, Foreign Language Speaking Anxiety, and Self-Perceived Communicative Competence. *System, 127,* Article 103533. https://doi.org/10.1016/j.system.2024.103533

Wu, J. H., Zhou, W. T., & Cao, C. (2024). An Empirical Study of Generative Artificial Intelligence Technology Empowering Speaking Instruction. *China Educational Technology, 4,* 105-111.

Wu, L. J. (2025). An Empirical Study of Large Language Models Empowering English Speaking Instruction. *Modern Educational Technology, 35,* 33-41.

Wu, T. T., Hapsari, I. P., & Huang, Y. M. (2026). Effects of Incorporating AI Chatbots into Think-Pair-Share Activities on EFL Speaking Anxiety, Language Enjoyment, and Speak-

ing Performance. *Computer Assisted Language Learning, 39,* 1408-1446. https://doi.org/10.1080/09588221.2025.2478271

Xu, J. J. (2026). A Human-Machine Collaborative Foreign Language Pedagogy Driven by the Dual Cores of Humanistic and Intelligent Literacy. *Foreign Language World, 2,* 2-8.

Zheng, Y. Y. (2024). Generative Artificial Intelligence in Foreign Language Education: Key Controversies and Theoretical Construction. *Foreign Language Education, 45,* 48-53.

Zou, M., & Huang, L. (2023). To Use or Not to Use? Understanding Doctoral Students' Acceptance of ChatGPT in Writing through Technology Acceptance Model. *Frontiers in Psychology, 14,* Article 1259531. https://doi.org/10.3389/fpsyg.2023.1259531