Journal of Information Technology Education: Research

An Official Publication of the Informing Science Institute
InformingScience.org

JITEResearch.org

Volume 25, 2026

BRIDGING PROFICIENCY GAPS: CHATGPT-ENHANCED STAD FOR ACHIEVEMENT, ATTITUDES, AND EQUITY IN MULTILINGUAL HIGHER EDUCATION

Yu-Heng Chen*	Chaoyang University of Technology, Taichung, Taiwan	luegg100@gmail.com
Kimyung Keng	Tamkang University, New Taipei City, Taiwan	kimyung@hawaii.edu

* Corresponding author

ABSTRACT

Aim/Purpose	This study examined whether embedding ChatGPT (GPT-4o) within Student Teams Achievement Divisions (STAD) cooperative learning improves academic achievement, learning attitudes, and interaction equity for linguistically heterogeneous undergraduates.
Background	Linguistically diverse classrooms face an equity problem in which language proficiency operates as a status characteristic. Most existing AI-in-education studies report aggregate effects rather than subgroup outcomes, and few embed AI within structured cooperative architectures.
Methodology	An 18-week quasi-experimental design with three time points was conducted at a Taiwanese university (N = 125; experimental n = 65, control n = 60). Participants were stratified by Test of Chinese as a Foreign Language (TOCFL) proficiency band, with Band A2 as the lower-proficiency threshold. The experimental condition followed a six-phase STAD workflow with tiered prompting and progressive scaffolding fading. Analyses included repeated-measures ANOVA, ANCOVA, and content analysis of reflection logs.
Contribution	This study contributes (a) initial subgroup-level evidence, awaiting multi-site replication, that AI-enhanced STAD may narrow proficiency-based achievement gaps in content-area cooperative learning, (b) a replicable governance-

Accepting Editor Martin D Beer | Received: February 22, 2026 | Revised: May 8, June 5, 2026 | Accepted: June 29, 2026.
Cite as: Chen, Y.-H., & Keng, K. (2026). Bridging proficiency gaps: ChatGPT-enhanced STAD for achievement, attitudes, and equity in multilingual higher education. *Journal of Information Technology Education: Research, 25,* Article 27. https://doi.org/10.28945/5824

	embedded design, and (c) a theoretical extension positioning AI as a more knowledgeable other within the zone of proximal development.
Findings	Five findings emerged. The experimental group exhibited steeper achievement growth (posttest $d = 0.91$). Four of six attitude dimensions improved significantly, led by Learning Confidence ($d = 0.97$). Self-reported prompting purposes differed substantially by proficiency subgroup (Cramér's $V = 0.37$), with lower-proficiency learners reporting more language-clarification queries. The proficiency-based achievement gap narrowed from 16.60 to 4.00 points (Group × Proficiency partial $\eta^2 = 0.050$). Exploratory and preliminary evidence further suggested that female students reported higher residual subject anxiety than male students. However, this finding did not survive Bonferroni correction and is constrained by low subscale reliability.
Recommendations for Practitioners	Practitioners should embed AI access at specific interaction nodes, differentiate scaffolding through tiered prompts, govern AI use through written agreements and reflection logs, plan progressive fading, and monitor affective outcomes for vulnerable subgroups.
Recommendations for Researchers and Future Research	Future research should deploy automated, timestamped AI usage logs; replicate across institutions and disciplines; revise low-reliability subscales using rigorous CFA on larger samples ($N \geq 300$); and conduct longitudinal follow-up to assess persistence of equity effects.
Impact on Society	The equity evidence suggests that blanket AI prohibitions may inadvertently disadvantage linguistically diverse learners. The governance-embedded design offers an adaptable model for institutional and national AI-use policies.

Keywords	ChatGPT, GPT-4o, cooperative learning, STAD, linguistically diverse learners, interaction equity, higher education, generative AI, scaffolding

INTRODUCTION

International student mobility has grown substantially over the past two decades. According to UNESCO data cited in the Organisation for Economic Co-operation and Development's (OECD, 2024) Education at a Glance 2024 report, the global population of internationally mobile tertiary students reached approximately 6.9 million in 2022. This represents a 176% increase from the 2002 baseline of 2.5 million (OECD, 2024). Internationally mobile students now constitute approximately 6% of all tertiary students in OECD countries. More than 4.6 million were enrolled across OECD-country institutions in 2022 (OECD, 2024). Beyond international inbound students, domestic minority-language learners, and students studying through non-native languages of instruction, the population that must engage with disciplinary content in a second language is further amplified. This demographic shift creates a structural challenge. When students of unequal language proficiency collaborate on communication-intensive tasks, language proficiency operates as a de facto status characteristic. It determines who speaks, who is heard, and whose ideas shape group products (Andrade & Valtcheva, 2009).

Cooperative learning structures, such as Student Teams Achievement Divisions (STAD) (Slavin, 1983, 1996), are widely adopted to promote positive interdependence and individual accountability. Meta-analytic evidence supports their positive effects on achievement, motivation, and social skills (Johnson et al., 1998; Kyndt et al., 2013). Yet well-documented interaction imbalances persist in linguistically heterogeneous groups. Higher-proficiency students dominate discussion while lower-proficiency peers withdraw (Gillies & Boyle, 2010; Le et al.,

2018). From a sociocultural perspective (Vygotsky, 1978), linguistic barriers restrict participation in meaning negotiation. The scaffolding mechanisms that cooperative learning is designed to activate become unevenly distributed.

The proliferation of generative artificial intelligence (AI), particularly large language models such as ChatGPT, opens new possibilities for addressing this challenge. Generative AI affords real-time linguistic simplification, concept clarification, bilingual bridging, and reflective prompting (Yan et al., 2024), all of which are directly relevant to linguistically diverse cooperative settings. However, design intentionality predicts learning outcomes more strongly than technology adoption *per se* (Tamim et al., 2011). The pedagogical value of ChatGPT, therefore, depends on the instructional architecture surrounding its use, including role assignments, interaction protocols, and metacognitive prompts.

Despite burgeoning interest, three interrelated gaps limit the field's ability to guide principled AI integration into cooperative learning for linguistically diverse populations: (a) methodological reliance on single-point, between-group designs that obscure learning trajectories; (b) limited evidence on AI scaffolding in content-area cooperative settings (as distinct from foreign-language or writing instruction); and (c) equity claims that are rarely operationalised with subgroup-level evidence. Each gap, including key empirical exemplars and methodological critiques, is examined in detail in the Literature Review. The present study addresses these gaps in one specific configuration of multilingual higher education: a Mandarin-medium content-area course with learners stratified by TOCFL band in a Taiwanese university. We do not claim that the findings generalise to typologically diverse plurilingual settings (e.g., European Erasmus or sub-Saharan African universities), and we return to this scope condition in the Limitations section.

The present study addresses these gaps by investigating how a ChatGPT-enhanced cooperative learning intervention influences achievement, attitudes, and interaction equity. The participants are first-year marketing students with heterogeneous Mandarin proficiency at a Taiwanese university. The study makes three contributions: (a) a replicable intervention design integrating ChatGPT as an adaptive scaffold within a structured cooperative learning workflow with explicit governance; (b) a multi-strand evaluation combining repeated-measures outcome analysis, usability assessment, and equity-focused subgroup analyses; and (c) subgroup-level equity evidence through Group × Proficiency interaction analyses. The latter draws on the computer-supported collaborative learning (CSCL) affordance framework (Jeong & Hmelo-Silver, 2016) to offer transferable design principles.

Research Questions

To enable consistent traceability across the Methodology, Results, Discussion, and Conclusion, six research questions and six sub-questions are enumerated. The hierarchy follows three strands.

Strand 1: Learning Outcomes

- **RQ1.** Does the ChatGPT-enhanced cooperative learning intervention improve academic achievement compared to traditional cooperative learning?
- **RQ2.** Does the intervention improve learning attitudes relative to the control group across six attitudinal dimensions?

Strand 2: Interaction Processes

- **RQ3.** How do students use the AI scaffold?
  - **RQ3a.** Is there a positive dose-response relationship between AI usage and learning outcomes?

  - **RQ3b.** Do students of differing language proficiency use the AI scaffold differently in terms of prompting purpose?
- **RQ4.** Does the intervention improve collaborative interaction quality within cooperative teams?

Strand 3: Equity and Moderation

- **RQ5.** Does the intervention narrow gaps between language-proficiency subgroups?
  - **RQ5a.** Does the achievement gap narrow between higher-proficiency and lower-proficiency subgroups?
  - **RQ5b.** Within the experimental group, do lower- and higher-proficiency learners differ in attitudinal change? (Exploratory within-experimental-group subgroup analysis; not a Group × Proficiency interaction test of differential intervention effects on attitudes.)
- **RQ6.** Are there gender-related differences in posttest outcomes? (Note: This RQ was operationalised through gender main-effect ANCOVAs, a Group × Gender interaction was not formally estimated.)
  - **RQ6a.** Are there gender differences in posttest achievement?
  - **RQ6b.** Are there gender differences in posttest attitudes, particularly subject anxiety?

The conceptual model predicts that the AI scaffold will function as a supplementary, more knowledgeable other (MKO) within the zone of proximal development (ZPD) for lower-proficiency learners. It will simultaneously reduce extraneous linguistic load and free cognitive resources for content processing. This dual mechanism predicts disproportionate gains for the most linguistically disadvantaged students. Detailed hypotheses corresponding to each subquestion appear in the methodology section.

The remainder of the paper is organised as follows. Building on the gaps outlined above, the literature review establishes the foundations of sociocultural and cognitive load theory. It critically evaluates empirical evidence on AI-assisted cooperative learning across diverse contexts and develops the conceptual model. The methodology then specifies the design, participants, intervention, instruments, and analytic plan. The results, discussion, implications, limitations, and conclusion sections follow in sequence.

LITERATURE REVIEW AND THEORETICAL FRAMEWORK

THEORETICAL FOUNDATIONS: SOCIOCULTURAL SCAFFOLDING AND COGNITIVE LOAD THEORY

The theoretical foundation of this study draws on two complementary frameworks: Vygotsky's (1978) sociocultural theory and cognitive load theory (CLT) (Sweller, 1988; Sweller et al., 2019). Both address a common pedagogical concern: how instructional support can be designed when task demands exceed the learner's independent capabilities.

The zone of proximal development (ZPD) posits that learning is most productive in the space between independent and assisted performance. The pedagogical mechanism enabling this is scaffolding: temporary, contingent support that fades as competence grows (Wood et al., 1976). Effective scaffolding is calibrated, responsive, and progressively faded (Belland, 2014; Puntambekar & Hübscher, 2005). Traditionally, scaffolding originates from instructors, capable peers, or structured materials. The introduction of generative artificial intelligence (AI) raises

the possibility that an AI system may function as an additional, more knowledgeable other (MKO). It can provide on-demand linguistic and conceptual support at the moment of need (Kasneci et al., 2023; Yan et al., 2024).

The conceptualisation of generative AI as a form of MKO is supported by recent systematic evidence (Kasneci et al., 2023; Yan et al., 2024). However, the analogy has important limitations. Unlike a human tutor, an AI tool cannot perceive nonverbal cues, sustain relational rapport, or exercise professional judgment about when to withhold support (Kasneci et al., 2023). This limitation underscores the need to embed AI scaffolding within a deliberately designed instructional workflow rather than as an unstructured supplement.

Cognitive load theory (CLT) provides a complementary lens. CLT distinguishes among intrinsic, extraneous, and germane load (Sweller et al., 2019). In linguistically heterogeneous classrooms, lower-proficiency learners face a dual processing burden. They must decode linguistic input while simultaneously engaging with disciplinary content. The result is elevated extraneous cognitive load that reduces resources for germane processing (Gibbons, 2015). An AI scaffold that simplifies language, provides bilingual bridging, or pre-clarifies terminology can reduce this extraneous load (Yan et al., 2024). Hong and Guo (2025), in a study of AI-enhanced multi-display language teaching systems with 302 EFL learners, reported significant improvements in cognitive load management. However, the study used self-reported load measures. It therefore left unaddressed whether processing efficiency genuinely improved or whether perceived load was reduced through novelty.

Integrating these frameworks yields a coherent rationale. Sociocultural theory explains why AI scaffolding should be embedded at specific interaction nodes (to function as an MKO within the ZPD). CLT explains how such scaffolding operates cognitively (by reducing extraneous linguistic load). Both converge on the principle that scaffolding must be adaptive, contingent, and progressively faded. This principle is also reflected in the SMART (Self-directed, Motivated, Adaptive, Resource-enriched, Technology-embedded) framework for smart education (Zhu et al., 2016). It informs the design heuristic of the present intervention.

Cooperative Learning and AI-Assisted Collaboration

Building on the theoretical foundation above, this section reviews the empirical evidence on cooperative learning in linguistically diverse contexts and the integration of AI tools. Meta-analytic syntheses consistently report positive effects of cooperative learning on achievement, motivation, and social skills (Johnson et al., 1998; Kyndt et al., 2013; Slavin, 1996). The STAD model is among the most widely implemented structures (Slavin, 1983), combining whole-class instruction, heterogeneous team practice, individual quizzes, and team recognition. Its effectiveness depends on positive interdependence and individual accountability (Slavin, 1996).

Despite these benefits, cooperative learning faces persistent challenges in linguistically heterogeneous classrooms. Research consistently documents interaction imbalances. Higher-proficiency students dominate discourse while lower-proficiency peers adopt passive roles (Gillies & Boyle, 2010; Le et al., 2018). Buchs et al. (2017) identified status hierarchies and uneven participation as among the most common obstacles. Notably, much of this evidence comes from European and North American settings. Comparable evidence from Southeast Asian and East Asian multilingual systems has only recently begun to accumulate. In these systems, Content and Language Integrated Learning (CLIL) and English- or Mandarin-medium instruction policies generate similar dynamics. Sahan et al. (2022) surveyed 17 universities across Vietnam and Thailand. They documented that English-only English-medium instruction (EMI) policies produce systematic exclusion of lower-proficiency learners. First-language use was frequently framed as a marker of low proficiency, despite its scaffolding function. Sahan et al. (2023), using a quantitative dataset of 544 EMI students, demonstrated that English-language self-efficacy, motivated behaviour, and the proportion of second-language jointly predict the severity

of language-related challenges. Self-efficacy partially mediated this relationship. Aizawa et al. (2023), studying 264 Japanese undergraduates in EMI international-business courses, found that TOEIC scores significantly predicted language-related challenges, although no clear threshold was observed in their sample. Together, these studies confirm that the proficiency-participation dynamic identified in Western literature generalises to Asian higher education contexts. Culturally distinct moderators such as face-saving norms and stratified language hierarchies warrant further empirical scrutiny.

Despite these advances, the integration of AI tools into cooperative learning workflows offers a potential mechanism for addressing interaction asymmetries. H. Wang et al. (2025) reported that an AI-agent-supported collaborative learning approach outperformed traditional computer-supported collaboration in achievement, self-efficacy, and learning interest. It also reduced cognitive load. However, that study used a small sample (n = 45) at a single university and lacked a no-AI baseline beyond traditional CSCL, limiting causal inference. Güner and Er (2025) identified five distinct student interaction profiles with ChatGPT. They demonstrated that instructional scaffolding shifted students toward more collaborative usage patterns. The cross-sectional design, however, precludes claims about sustained change. Collectively, these studies indicate that the effectiveness of AI-assisted collaboration depends on the instructional design surrounding tool use, rather than mere availability. Methodological limitations constrain the strength of current evidence: short durations (often less than 8 weeks), single-site samples, and reliance on self-report.

Three categories of process indicators have emerged as informative for evaluating AI-assisted cooperative learning: (a) usage intensity and timing, (b) prompting purpose distribution, and (c) collaborative engagement quality (Jeong & Hmelo-Silver, 2016). These indicators, alongside outcome measures, form the evidence chain required to explain not only whether an AI-enhanced intervention works, but also how and for whom it operates.

ChatGPT in Education: Empirical Evidence and Risk Governance

A growing number of studies demonstrate that ChatGPT can enhance cooperative and collaborative learning when embedded in structured pedagogical designs. In foreign-language learning courses, Tsai et al. (2024) reported that ChatGPT-assisted writing produced measurable benefits for English as a Foreign Language (EFL) majors. Karataş et al. (2024), in a study of foreign-language learners using ChatGPT, documented improvements in language learning outcomes, motivation, and engagement when AI use was pedagogically scaffolded. In collaborative-learning contexts, H. Wang et al. (2025) reported similar advantages of AI-agent-supported collaboration, and Chen et al. (2025) demonstrated that generative AI as a reflective scaffold promoted higher-order thinking among first-year STEM project undergraduates. These studies converge on a key design principle: generative AI is most effective when assigned a specific pedagogical role, rather than offered as an unrestricted resource.

For linguistically diverse learners, the scaffolding affordances of ChatGPT are particularly relevant. They include reduced language-processing load (Gibbons, 2015), increased participation through low-stakes pre-formulation (Andrade & Valtcheva, 2009; Eysenck & Calvo, 1992), and metacognitive support for self-regulated learning (F. Wang & Hannafin, 2005). Karataş et al. (2024), Tsai et al. (2024), and Woo et al. (2024) report positive impacts on writing, grammar, vocabulary, and motivation. However, Woo et al. (2024) caution that prompt engineering itself can impose heavy cognitive demands. This caveat highlights the need to scaffold AI use itself.

Despite these advances, three methodological limitations recur across the AI-in-education literature. First, sample sizes are often modest (typical n = 30-80), which limits power for moderation analyses. Second, most studies employ short interventions (often a single semester or less) and do not capture trajectory-level change. Third, AI version variability is rarely addressed:

ChatGPT (GPT-3.5), ChatGPT (GPT-4), GPT-4o, and subsequent models differ substantially in capability, yet cross-version comparisons are largely absent (OpenAI, 2024). These limitations bound the inferential strength of current evidence and motivated the present study's design.

Alongside these opportunities, generative AI introduces governance challenges that require deliberate design responses. Four categories of risk are most prominent. First, over-reliance can create a scaffolding paradox in which the support tool undermines learning (Kasneci et al., 2023). Second, hallucination produces plausible but factually incorrect content (Ji et al., 2023). Third, academic integrity violations arise from the blurred boundary between AI-assisted and AI-generated work (Yan et al., 2024). Fourth, privacy risks emerge when learners input personally identifiable information into commercial platforms. Effective governance mechanisms include structured prompt protocols, AI usage reflection logs, teacher monitoring, and assessment designs that evaluate learner understanding rather than AI output quality (F. Wang & Hannafin, 2005). The governance dimension represents an underreported aspect of existing research. Most studies document productivity gains without specifying mechanisms used to mitigate hallucination, over-reliance, or academic integrity risks (Mohamed, 2024).

Research Gaps and Conceptual Model

Building on the critical review above, the three gaps introduced earlier remain unresolved in the literature. The present study addresses all three through a repeated-measures quasi-experimental design situated in a content-area cooperative-learning context with linguistically heterogeneous students, supplemented by Group × Proficiency interaction analyses with effect sizes and measurement invariance testing. Recent scholarship has begun to address AI bias and equity in higher education (Roshanaei, 2024). However, few studies have operationalised equity claims with the analytic specificity attempted here.

The conceptual model synthesises these foundations into a three-layer causal logic:

- **Layer 1, Intervention design features:** ChatGPT-enhanced STAD with a tiered prompt protocol, governance mechanisms, and progressive scaffolding fading.
- **Layer 2, Interaction processes:** AI usage patterns, prompting-purpose distribution, and collaborative engagement quality.
- **Layer 3, Learning outcomes:** Academic achievement trajectories and attitudinal change.

Language proficiency and gender serve as moderating variables across all layers. The model predicts that the AI scaffold will function as a more knowledgeable other (MKO) within the zone of proximal development (ZPD) for lower-proficiency learners (sociocultural mechanism). It will simultaneously reduce extraneous linguistic load and free cognitive resources for content processing (CLT mechanism). This dual mechanism produces disproportionate gains for the most linguistically disadvantaged students.

METHODOLOGY

Research Design and Participants

This study used a quasi-experimental pretest-midterm-posttest control-group design with repeated measures. Two intact classes of first-year students at a Taiwanese university were assigned to the experimental condition (ChatGPT-enhanced STAD) or control condition (traditional STAD without ChatGPT). The course was a required brand-management module. It was delivered face-to-face in Mandarin Chinese, three sessions per week over 18 weeks (February through June 2025). The same instructor taught both groups using identical syllabi and assessment instruments.

The 18-week schedule comprised four phases. Weeks 1–2 covered orientation, baseline assessment, and ChatGPT training. Weeks 3–8 and Weeks 10–15 (12 weeks total) comprised cooperative learning units with weekly reflection logs. Week 9 was the midterm test. Weeks 16–18 included project preparation, posttest, and usability assessment. The 12-week reflection log period yielded N = 780 entries (65 students × 12 weeks).

Of the 130 enrolled students, five were lost to follow-up (three from the experimental group due to incomplete data; two from the control group due to course withdrawal). The final analytic sample was N = 125 (experimental n = 65; control n = 60; attrition = 3.8%, listwise deletion). Given the low attrition rate (< 5%) and the fact that withdrawals were attributable to course-administrative reasons unrelated to the intervention or to outcome variables, the missingness was treated as approximately Missing Completely at Random (MCAR) (Little & Rubin, 2019). Listwise deletion under MCAR yields unbiased estimates with only modest efficiency loss. Sensitivity analyses with multiple imputation (m = 20) on the attitude data produced substantively identical results and are therefore not separately reported. Participant flow appears in Figure 1. Students were stratified by Test of Chinese as a Foreign Language (TOCFL) Band A2, which corresponds to approximately 1,000 vocabulary items (Steering Committee for the Test of Proficiency-Huayu, 2023). The brand-management curriculum involves abstract concepts such as brand equity. The A2 threshold, therefore, marks a meaningful linguistic disadvantage in this Mandarin-medium environment.

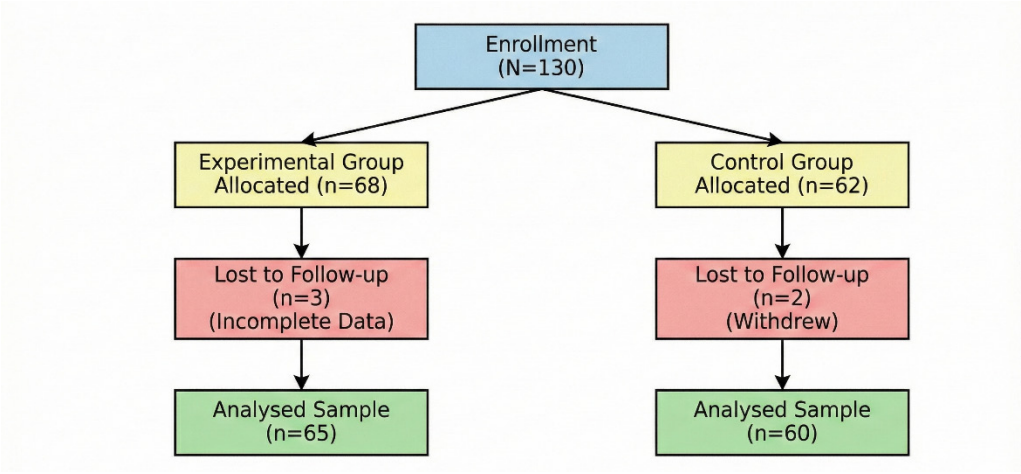

Figure 1. Participant flow diagram (enrolment: N = 130; final analytic sample: experimental n = 65, control n = 60; attrition rate = 3.8%)

Experimental-group students were assigned to heterogeneous teams of four (15 teams of four; one team of five), balanced by academic performance, gender, and language proficiency. Roles rotated weekly: discussion leader, recorder, reporter, and checker. The control group used the same STAD structure without ChatGPT access. The only systematic difference between conditions was ChatGPT availability during individual preparation and group discussion phases.

Baseline equivalence held across all variables (Table 1). The two groups did not differ on pretest achievement ($t(123) = 1.21$, $p = .229$), any pretest attitude dimension (all $p > .50$), gender composition ($\chi^2 = 0.05$, $p = .815$), or language-proficiency distribution ($\chi^2 = 0.19$, $p = .867$).

Limitations of the quasi-experimental design

Random assignment was not feasible, so intact classes were used. This creates internal validity threats, including selection bias and cohort effects. Two design features partially mitigate these

threats: (a) statistical control of pretest scores via ANCOVA, and (b) confirmed baseline equivalence. Nevertheless, instructor expectancy and Hawthorne effects cannot be fully excluded. Findings are interpreted accordingly.

Table 1. Demographic characteristics and baseline scores by group

Variable	Experimental (n = 65)	Control (n = 60)	Test statistic	*p*
Gender (% female)	55.4%	53.3%	$\chi^2 = 0.05$	.815
Language proficiency (% low)	53.8% (n = 35)	50.0% (n = 30)	$\chi^2 = 0.19$	.867
Pretest achievement, *M* (*SD*)	42.07 (18.50)	46.15 (19.20)	$t(123) = 1.21$	.229
Pretest A: Learning Confidence, *M* (*SD*)	3.20 (0.55)	3.25 (0.60)	$t(123) = 0.49$	.625
Pretest B: Perceived Usefulness, *M* (*SD*)	3.33 (0.51)	3.30 (0.54)	$t(123) = 0.33$	.745
Pretest C: Motivation for Exploration, *M* (*SD*)	3.41 (0.50)	3.38 (0.52)	$t(123) = 0.34$	.734
Pretest D: Success Orientation, *M* (*SD*)	3.10 (0.51)	3.15 (0.53)	$t(123) = 0.55$	.582
Pretest E: Social Support, *M* (*SD*)	3.26 (0.46)	3.22 (0.48)	$t(123) = 0.49$	.624
Pretest F: Subject Anxiety, *M* (*SD*)	2.95 (0.52)	2.90 (0.55)	$t(123) = 0.54$	.591

Note. For language proficiency, Fisher's exact test (two-tailed) yielded $p = .722$; the uncorrected Pearson χ^2 is also reported. Both confirm group equivalence ($p > .70$).

Intervention Procedures

The intervention integrated ChatGPT (GPT-4o) as an adaptive scaffold within the STAD workflow. All ChatGPT interactions used the web interface (accessed February through June 2025). Each cooperative learning unit followed a six-phase sequence (Figure 2). ChatGPT was actively embedded at Phases 2 and 3. Phase 6 was a structured AI-reflection phase in which students submitted written logs about their AI use. Consistent with sociocultural theory, the AI tool functioned as a supplementary, more knowledgeable other (MKO) within the zone of proximal development (ZPD).

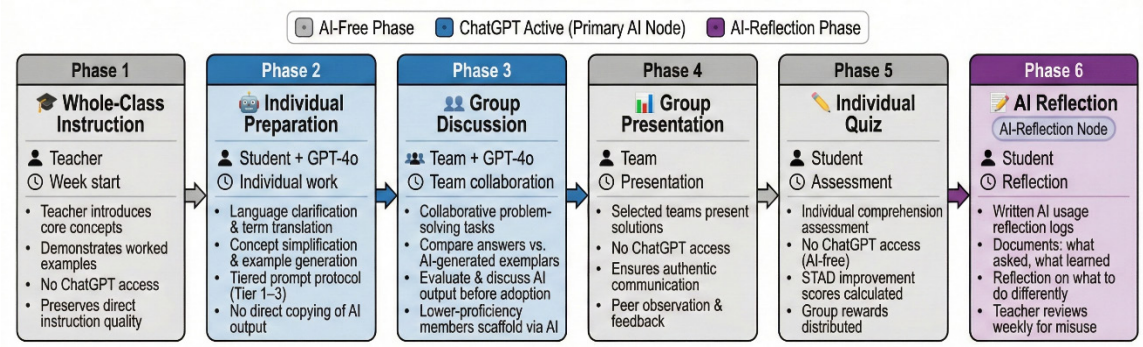

Figure 2. Six-phase ChatGPT-enhanced STAD cooperative learning workflow, showing the two primary ChatGPT interaction nodes (Phases 2 and 3), the AI-reflection phase (Phase 6), and the AI-free phases (Phases 1, 4, and 5)

Phase 1 (whole-class instruction): the teacher introduced core concepts; ChatGPT was not used. Phase 2 (individual preparation): each student used ChatGPT for language clarification, term translation, example generation, and concept simplification, following the tiered prompt protocol. Direct copying of AI output was prohibited. Phase 3 (group discussion): teams used ChatGPT collectively to generate alternative explanations and support lower-proficiency members. AI output was discussed and evaluated by the group before adoption. Phase 4 (group presentation): selected groups presented solutions without ChatGPT. Phase 5 (individual quiz): quizzes assessed comprehension without ChatGPT access; improvement scores generated team rewards. Phase 6 (reflection): students submitted weekly AI usage reflection logs.

The tiered prompt protocol comprised three tiers. Tier 1 (all learners) provided standard prompts for concept clarification. Tier 2 (lower-proficiency learners) added prompts for language simplification and bilingual explanation. Tier 3 (scaffolding fading) encouraged independent task attempts before AI consultation, with progressive reduction across the semester. A Week-2 training session covered ChatGPT capabilities, prompt-engineering basics, and academic-integrity guidelines. Prompt exemplars are available from the corresponding author upon reasonable request.

Governance mechanisms addressed four risk categories. Over-reliance was mitigated through scaffolding fading and AI-free quizzes. For hallucination, students cross-checked AI claims against textbook materials. Academic integrity was maintained through a written AI usage agreement, weekly reflection logs, and AI-free summative assessments. For privacy, students were instructed not to input personally identifiable information. The full intervention specification, including its mapping to the SMART (Self-directed, Motivated, Adaptive, Resource-enriched, Technology-embedded) design principle (Zhu et al., 2016), is available from the corresponding author upon reasonable request.

INSTRUMENTS

Academic achievement

Three paper-based tests were administered in Weeks 1, 9, and 17. Each contained 10 to 15 items covering conceptual understanding, application, and analysis within brand management. The instructor developed test items based on Keller's (2013) textbook (delivered in Chinese translation). Two subject-matter experts reviewed items for content validity. Tests were administered during regular class sessions, with a 40-50-minute response time. Students also completed a final group project assessed using a standardised rubric.

Learning attitudes

A 55-item Composite Learning Attitude Scale was constructed for this study (Table 2). Four dimensions (37 items) were adapted from the Attitudes Toward Mathematics Inventory (ATMI; Tapia & Marsh, 2004): Learning Confidence (A, 10 items), Perceived Usefulness (B, 10 items), Motivation for Exploration (C, 10 items), and Subject Anxiety (F, 7 items). Two supplementary dimensions (18 items) were integrated from the Colorado Learning Attitudes about Science Survey (CLASS; Adams et al., 2006) and the Test of e-Learning Related Attitudes (TeLRA) Scale (Kisanga & Ireson, 2016): Success Orientation (D, 10 items) and Social Support (E, 8 items). All items used a five-point Likert scale.

A pilot study (n = 20) identified six factors explaining 51.7% of variance, consistent with the a priori structure. Confirmatory factor analysis (CFA) on the full pretest sample (N = 130) yielded acceptable fit: $\chi^2/df = 2.45$, CFI = 0.942, TLI = 0.935, RMSEA = 0.068, SRMR = 0.055. All standardised loadings exceeded 0.40. Cronbach's α ranged from 0.79 to 0.88 for five dimensions. The Subject Anxiety dimension (F) yielded $\alpha = 0.574$, below the 0.70 threshold (Nunnally & Bernstein, 1994). All Dimension F findings are therefore treated as preliminary throughout this manuscript.

Table 2. Internal consistency of the Learning Attitude Scale

Dimension	Items	Cronbach's α
A: Learning Confidence	10	0.88
B: Perceived Usefulness	10	0.85
C: Motivation for Exploration	10	0.82
D: Success Orientation	10	0.84
E: Social Support	8	0.79
F: Subject Anxiety	7	**0.574**
Overall	**55**	**0.90**

Note. Scale sources: A, B, C, F adapted from ATMI (Tapia & Marsh, 2004); D adapted from CLASS (Adams et al., 2006); E adapted from TeLRA (Kisanga & Ireson, 2016). See subsection "Learning Attitudes" under Instruments for the Subject Anxiety subscale reliability caveat.

The participant-to-item ratio for the CFA was approximately 1:2.4 (N = 130; 55 items), below the 5:1 minimum recommended by Hair et al. (2019). Following Kline's (2023) view that adequacy depends on indicator reliability and factor definition rather than a rigid ratio, the present six-factor solution is acceptable but suboptimal. Replication with $N \geq 300$ is warranted. Future cycles will report McDonald's ω alongside α (McDonald, 1999).

Measurement invariance testing across language-proficiency subgroups followed Vandenberg and Lance (2000). Configural invariance was supported (CFI = 0.938, RMSEA = 0.072). The configural-to-metric transition produced $\Delta CFI = 0.008$, within the 0.01 threshold (Cheung & Rensvold, 2002). Strict scalar invariance testing was not feasible with subgroup $n < 100$. Subgroup attitudinal comparisons should therefore be interpreted with caution. The complete 55-item Composite Learning Attitude Scale is available from the corresponding author upon reasonable request, subject to the licensing terms of the source instruments (ATMI, CLASS, and TeLRA). The System Usability Scale (SUS; Brooke, 1996) was administered to the experimental group at posttest. The SUS comprises 10 items yielding a composite score (0 to 100); a score of 68 represents the average benchmark, while scores above approximately 80 indicate above-average usability (Bangor et al., 2008).

Interaction process data

Automated, timestamped usage logs were not collected in this evaluation cycle. Two alternative process-data sources were used: (a) content analysis of student reflection logs (N = 780 entries) coded by two independent raters into four prompting-purpose categories (Cohen's $\varkappa = 0.88$), and (b) structured instructor field notes documenting AI-assisted collaborative behaviour. Because these data are self-reported or observer-derived, claims about actual AI use are correspondingly limited.

Data Analysis

Distributional assumptions and baseline equivalence

Shapiro-Wilk tests indicated approximately normal distributions of achievement scores (all $p > .05$). Levene's tests confirmed homogeneity of variance at pretest ($F(1, 123) = 0.42$, $p = .518$), midterm ($p = .353$), and posttest ($p = .268$). Baseline equivalence used independent-samples t-tests for pretest variables and chi-square tests for demographic composition. For attitude comparisons, a Bonferroni-adjusted $\alpha = .0083$ ($= .05/6$) was applied. Achievement and demographic comparisons used $\alpha = .05$.

Primary analyses

To address RQ1 (achievement), a repeated-measures ANOVA tested the Time × Group interaction. The Greenhouse-Geisser correction was applied where sphericity was violated. Post hoc pairwise contrasts were Bonferroni-adjusted. Effect sizes are reported as partial η^2 and Cohen's d with 95% confidence intervals.

To address RQ2 (attitudes), one-way ANCOVA was conducted for each of the six dimensions. The dependent variable was the posttest score; the covariate was the corresponding pretest score. Homogeneity of regression slopes was verified (all interaction $p > .10$). Bonferroni correction ($\alpha = .0083$) controlled the family-wise error rate.

To address RQ5a (proficiency-based achievement gaps), a two-way ANCOVA tested the Group × Proficiency interaction; significant interactions were followed by simple-effects analyses. RQ5b (proficiency-based attitude differences) was addressed exploratorily through within-experimental-group ANCOVAs ($n = 65$), examining whether proficiency moderated attitudinal change inside the experimental condition; a fully crossed Group × Proficiency interaction on attitudes was not estimated because subgroup cell sizes (lower-proficiency control $n = 30$; higher-proficiency control $n = 30$) yielded insufficient statistical power for stable interaction estimates given the multidimensional attitude structure. Findings for RQ5b are therefore reported as exploratory rather than as direct tests of differential intervention effects on attitudes. To address RQ6a (gender main effects on outcomes), one-way ANCOVAs with Bonferroni correction were conducted. RQ6b (gender-moderated intervention effects) was likewise addressed exploratorily; a Group × Gender interaction was not formally tested for the same power-related reasons. Partial η^2 is reported for omnibus F-tests; ω^2 is reported for subgroup analyses (Olejnik & Algina, 2003).

Power analysis

An a priori power analysis was conducted using G*Power 3.1 (Faul et al., 2009) before data collection to determine the minimum sample size requirements. With $\alpha = .05$ and target power = .80, the analysis indicated that $N = 52$ would suffice for the repeated-measures ANOVA assuming a medium effect ($f = 0.25$), and $N = 85$ would suffice for the six ANCOVAs under Bonferroni-adjusted $\alpha = .0083$. The achieved sample of $N = 125$, therefore, satisfied a priori sample-size adequacy for the primary analyses. Separately, post hoc inspection of observed effects shows that the four-cell Group × Proficiency model produced a significant interaction (partial $\eta^2 = 0.050$); this is reported as a descriptive observation rather than as confirmation of a priori power. A larger sample would benefit finer-grained moderation analyses, particularly fully crossed Group × Gender × Proficiency models.

Clustering and design effect

Because students were nested within cooperative teams in the experimental group, the intraclass correlation coefficient (ICC) at the team level was computed. The experimental group comprised 16 teams (15 of four; one of five) with a mean cluster size of $m = 4.06$. The ICC for posttest achievement was 0.038. Following Killip et al. (2004) and Snijders and Bosker (2012), the design effect was:

$$DE = 1 + (m - 1) \times ICC = 1 + (4 - 1) \times 0.038 = 1 + 0.114 = \mathbf{1.114}$$

The DE of 1.114 falls below 1.5 and well below 2.0, above which clustering would warrant multilevel modelling (Lai & Kwok, 2015). However, Lai and Kwok caution that the $DE < 2$ rule should not apply when cluster size is small ($m < 10$). The present $m = 4$ falls in this exception. A single-level strategy was retained for primary analyses. Cluster-robust standard errors were computed as a sensitivity check. The substantive pattern of results was unchanged. The control group also followed a STAD architecture and, therefore, involved team-level nesting in principle. Because instructor-led grouping in the control condition was less stable across the

semester and team-level outcome data were not consistently recorded, an ICC was not computed for the control group; this asymmetry is acknowledged as a limitation of the present clustering analysis. More fundamentally, the study employed a quasi-experimental pretest-midterm-posttest design using two pre-existing intact classes. Students were not randomly assigned to experimental conditions. Instead, one intact class was assigned to the ChatGPT-enhanced STAD condition and the other to the traditional STAD condition. Within each class, students were stratified by TOCFL band and gender to support balanced team composition for STAD groupings, but this stratification operated within a single condition and does not constitute random assignment between conditions. Consequently, treatment effects cannot be fully disentangled from class-level effects (instructor differences, scheduling, room characteristics, peer-cohort dynamics), and all causal claims should be interpreted cautiously. We address this constraint explicitly in the Limitations section and reiterate that multi-site, multi-class replication is required before any of the present findings can be interpreted as design-attributable rather than class-attributable.

Robustness and software

Robustness checks included ANCOVA re-estimation, outlier analysis (±3 SD), and non-parametric alternatives where assumptions were violated. Analyses used SPSS v.28; CFA and measurement invariance used AMOS v.28; effect sizes were verified with G*Power 3.1.

Ethical Considerations

All participants provided written informed consent. Non-participation did not affect course grades. Under Taiwan's regulations, studies evaluating pedagogical practices in general educational settings are eligible for exemption from formal Institutional Review Board (IRB) review. This low-risk classroom-based pedagogical evaluation was therefore exempt. The study was reviewed and approved by the department-level academic affairs committee. Strict ethical protocols were maintained: numerical-code de-identification, encrypted data storage, and a prohibition on entering personally identifiable information into ChatGPT. Future research cycles will seek formal IRB approval.

Generative AI in the research and writing process

The authors disclose the following uses of generative AI. First, ChatGPT (GPT-4o) was deployed as the experimental intervention; its use is the object of investigation rather than a writing aid. Second, in manuscript preparation, the authors used ChatGPT and Gemini for language-editing assistance only. No AI-generated text was incorporated verbatim. Third, AI tools were not used to generate or interpret empirical data. All statistical analyses were conducted by the authors using SPSS, AMOS, and G*Power. The authors take full responsibility for the content, accuracy, and originality.

RESULTS

Overview

Building on the analytic plan specified above, this section reports findings in the sequence in which they were tested. It addresses the six research questions and six sub-questions across three strands: learning outcomes (RQ1, RQ2), interaction processes (RQ3a, RQ3b, RQ4), and equity and moderation (RQ5a, RQ5b, RQ6a, RQ6b). Five principal findings emerge: (a) the experimental group exhibited a significantly steeper achievement trajectory (RQ1, posttest d = 0.91); (b) four of six attitude dimensions improved significantly after Bonferroni correction (RQ2); (c) self-reported prompting purposes differed substantially by proficiency subgroup (RQ3b, Cramér's V = 0.37); (d) the proficiency-based achievement gap narrowed from 16.60

to 4.00 points (RQ5a, partial $\eta^2 = 0.050$); and (e) female students reported higher residual subject anxiety than male students (RQ6b). All findings are presented descriptively below; theoretical interpretation is reserved for the discussion section.

Academic Achievement (RQ1)

Table 3 presents achievement descriptive statistics across three time points. A repeated-measures ANOVA with time (three levels) within-subjects, and group between-subjects revealed a significant Time × Group interaction, $F(1.88, 231.24) = 15.42$, $p < .001$, partial $\eta^2 = 0.11$, 95% CI [0.05, 0.18], exceeding the medium-effect threshold (Cohen, 1988).

Post hoc Bonferroni-adjusted pairwise comparisons confirmed no significant group difference at pretest (mean difference = 4.08, $p = .229$). By midterm, the experimental group surpassed the control (mean difference = 7.20, $p = .012$, $d = 0.44$). At posttest, the experimental group outperformed the control by 14.20 points, 95% CI [8.50, 19.90], $d = 0.91$, exceeding the large-effect threshold (Cohen, 1988). The achievement trajectory is visualised in Figure 3.

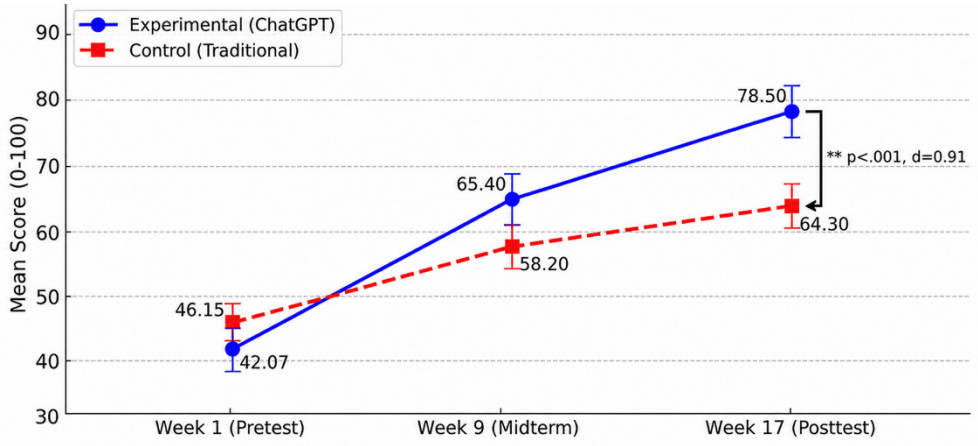

Figure 3. Mean academic achievement scores by group at the three measurement points: pretest (Week 1), midterm (Week 9), and posttest (Week 17) (error bars represent 95% confidence intervals)

Table 3. Academic achievement scores by group and time point

Time point	Experimental *M* (*SD*)	95% CI	Control *M* (*SD*)	95% CI
Pretest (Week 1)	42.07 (18.50)	[37.50, 46.60]	46.15 (19.20)	[41.30, 51.00]
Midterm (Week 9)	65.40 (15.50)	[61.60, 69.20]	58.20 (17.10)	[53.90, 62.50]
Posttest (Week 17)	78.50 (14.20)	[75.00, 82.00]	64.30 (16.80)	[60.00, 68.60]

Note. Time × Group interaction: $F(1.88, 231.24) = 15.42$, $p < .001$, partial $\eta^2 = 0.11$.

Greenhouse-Geisser correction applied ($\varepsilon = 0.94$). Bonferroni-adjusted contrasts: pretest $p = .229$; midterm $p = .012$, $d = 0.44$; posttest $p < .001$, $d = 0.91$. Effect sizes are reported as Hedges' g with the small-sample bias correction $J = 1 - 3/(4 \cdot df - 1) = 0.994$ applied to the pooled-SD Cohen's d. The uncorrected $d \approx 0.92$ from the descriptive statistics; the bias-corrected $g \approx 0.91$. Per APA 7 convention, d and g are presented under the same symbol. Means in Table 3 are unadjusted observed values. ANCOVA-adjusted subgroup means are reported in Table 5 and may differ slightly from the marginal means implied by Table 3. For example, the

experimental-group posttest mean of 78.50 in Table 3 corresponds to an ANCOVA-adjusted weighted mean of approximately 78.35 in Table 5 owing to pretest covariate adjustment.

Learning Attitudes (RQ2)

Table 4 presents ANCOVA results for each attitude dimension (posttest as the dependent variable; pretest as the covariate). Homogeneity of regression slopes was verified for all dimensions (all interaction $p > .10$).

Table 4. ANCOVA results for learning attitude dimensions: Experimental versus Control Group

Dimension	Exp. Post *M* (*SD*)	Ctrl. Post *M* (*SD*)	*F*(1, 122)	*p*	Partial η^2	Cohen's *d*	95% CI for *d*
A: Learning Confidence	3.88 (0.47)	3.35 (0.58)	32.18	< .001**	.209	0.97	[0.60, 1.34]
B: Perceived Usefulness	3.72 (0.49)	3.42 (0.52)	12.85	< .001**	.095	0.59	[0.23, 0.95]
C: Motivation for Exploration	3.60 (0.46)	3.45 (0.50)	3.52	.063	.028	0.31	[-0.04, 0.67]
D: Success Orientation	3.42 (0.48)	3.20 (0.50)	7.62	.007**	.059	0.45	[0.09, 0.80]
E: Social Support	3.56 (0.45)	3.32 (0.47)	9.45	.003**	.072	0.52	[0.16, 0.87]
F: Subject Anxiety†	2.72 (0.49)	2.88 (0.53)	4.15	.044	.033	-0.31	[-0.67, 0.05]

Note. ** indicates significance at Bonferroni-adjusted $\alpha = .0083$. *M* (*SD*) values are unadjusted observed posttest means; Cohen's *d* values are computed from ANCOVA-adjusted means. For Dimension F, negative *d* indicates lower anxiety in the experimental group.

†Caution: The Subject Anxiety dimension (F) yielded $\alpha = 0.574$, below the conventional 0.70 threshold; findings are preliminary rather than confirmatory.

After Bonferroni correction, four dimensions showed significantly greater improvement in the experimental group: Learning Confidence ($d = 0.97$, large), Perceived Usefulness ($d = 0.59$, medium), Success Orientation ($d = 0.45$, small to medium), and Social Support ($d = 0.52$, medium). Motivation for Exploration did not reach conventional significance ($p = .063$); the 95% CI for *d* [-0.04, 0.67] marginally included zero at the lower bound. *Subject Anxiety did not survive Bonferroni correction ($p = .044$, $d = -0.31$; CI crossed zero) and is treated as suggestive (see †).*

Equity and Moderation (RQ5a, RQ5b, RQ6a, RQ6b)

Achievement gap by language proficiency (RQ5a)

A two-way ANCOVA, with posttest achievement as the dependent variable, group and language proficiency as between-subject factors, and pretest achievement as the covariate, tested the gap-narrowing hypothesis (Table 5).

The Group × Proficiency interaction was significant, $F(1, 120) = 6.30$, $p = .013$, partial $\eta^2 = 0.050$ (small to medium). Simple-effects analysis revealed that the proficiency-based achievement gap persisted in the control group (higher-proficiency $M = 72.60$ versus lower-proficiency $M = 56.00$; gap = 16.60 points) but was markedly reduced in the experimental group

(higher-proficiency M = 80.50 versus lower-proficiency M = 76.50; gap = 4.00 points). Subgroup means are ANCOVA-adjusted. The overall model accounted for approximately 44% of variance in posttest achievement (R^2 = 0.44). This pattern is visualised in Figure 4.

Table 5. Two-Way ANCOVA:
Posttest achievement by group and language proficiency

Source	*SS*	*df*	*MS*	*F*	*p*	Partial η^2
Pretest (covariate)	2450.5	1	2450.5	45.20	< .001	.273
Group	1850.2	1	1850.2	34.10	< .001	.221
Proficiency	520.4	1	520.4	9.60	.002	.074
Group × Proficiency	340.8	1	340.8	6.30	.013	.050
Error	6500.0	120	54.2			

Table 5a. Covariate-adjusted posttest achievement means by group and proficiency subgroup

Group	Proficiency	n	Adjusted M	SE	95% CI
Experimental	Lower	35	76.50	1.25	[74.02, 78.98]
Experimental	Higher	30	80.50	1.35	[77.83, 83.17]
Control	Lower	30	56.00	1.35	[53.33, 58.67]
Control	Higher	30	72.60	1.35	[69.93, 75.27]

Note. Adjusted means and 95% confidence intervals are estimated marginal means from the two-way ANCOVA model reported in Table 5 (posttest achievement as the dependent variable; pretest achievement as the covariate). The proficiency-based gap is 4.00 points in the experimental condition and 16.60 points in the control condition.

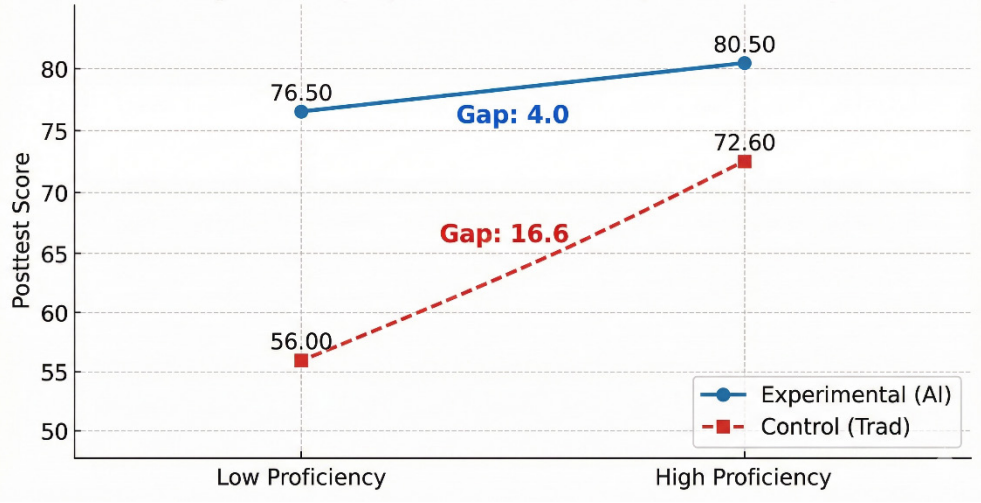

Figure 4. Covariate-adjusted posttest achievement (estimated marginal means from the two-way ANCOVA reported in Table 5) by group and proficiency subgroup

Although the y-axis is labelled "Posttest Score" for brevity, the plotted values are ANCOVA-adjusted means rather than raw scores. The Group × Proficiency interaction (partial η^2 = 0.050) indicates disproportionate gains for lower-proficiency learners. Numerical 95% confidence intervals for each subgroup mean are reported in Table 5a (immediately above) and in the accompanying text.

Attitude differences by language proficiency (RQ5b)

Within the experimental group, ANCOVAs controlling for pretest indicated significant proficiency effects for Learning Confidence ($F(1, 62) = 5.44$, $p = .023$, $\omega^2 = 0.07$) and Subject Anxiety ($F(1, 62) = 3.79$, $p = .046$, $\omega^2 = 0.08$), with lower-proficiency students showing greater improvement on both. The remaining four dimensions showed no significant proficiency effects (all $p > .05$).

Gender effects (RQ6a, RQ6b)

To address RQ6a (gender differences in achievement), a one-way ANCOVA was conducted with gender as the between-subjects factor and pretest achievement as the covariate. No significant gender difference in posttest achievement was found ($F(1, 122) = 0.85$, $p = .358$). The Group × Gender interaction was not tested separately. The reason was insufficient statistical power for a fully crossed factorial model with the current sample size. This analysis is planned for the next evaluation cycle.

To address **RQ6b** (gender differences in attitudes), one-way ANCOVAs across the six attitude dimensions identified Subject Anxiety as the only dimension approaching significance ($F(1, 122) = 5.45$, $p = .021$, $\omega^2 = 0.04$), with female students reporting higher anxiety. This effect did not survive Bonferroni correction. This RQ6b finding (gender-based Subject Anxiety) is analytically distinct from the experimental-versus-control comparison of Subject Anxiety reported in Table 4 under RQ2.

Supplementary Evidence: Interaction Processes and Usability (RQ3a, RQ3b, RQ4) — Preliminary and Exploratory

Automated usage logs were not available in this evaluation cycle; two preliminary process-data sources address RQ3a, RQ3b, and RQ4. Content analysis of student reflection logs ($N = 780$ entries; 65 students × 12 weeks of cooperative learning units) revealed a distribution of self-reported prompting purposes: language clarification (30.1%), concept explanation (27.4%), example generation (25.9%), and metacognitive self-checking (16.6%). The distribution differed significantly by proficiency subgroup (RQ3b), $\chi^2(3) = 108.45$, $p < .001$, Cramér's $V = 0.37$. Lower-proficiency students reported substantially more language-clarification queries (45.2%) than higher-proficiency students (12.5%), while higher-proficiency students reported proportionally more concept-explanation and metacognitive self-checking queries. Inter-rater reliability was excellent (Cohen's $\kappa = 0.88$). This differential usage pattern provides preliminary support for the differential-usage hypothesis underlying RQ3b, although the self-reported nature of the data limits causal inference. With respect to RQ3a (dose-response), automated log data are required to test this rigorously and were not available in this cycle; the SUS-achievement correlation reported below provides only indirect proxy evidence.

To address RQ4 (collaborative interaction quality), instructor field notes documented improved engagement among lower-proficiency students. Those who consulted ChatGPT during Phase 2 contributed more actively in Phase 3 than in prior semesters without AI support. Across 48 observed discussion sessions, 112 instances of group evaluation of AI output and 18 teacher interventions for AI misuse were recorded. These observations provide preliminary descriptive evidence for RQ4. Implementing automated, timestamped usage logs is the highest-priority improvement for the next design cycle.

The mean System Usability Scale (SUS) score was 72.5 (SD = 12.4). This exceeds the average SUS benchmark of 68 (Bangor et al., 2008); scores above approximately 80 are considered above average. SUS scores correlated positively with achievement gains ($r = .35$, $p = .004$). No significant difference in SUS scores was found between proficiency subgroups ($t(63) = 1.12$, $p = .267$). Causal direction cannot be inferred from this correlation alone.

Summary of Hypothesis Testing

Table 6 consolidates results across all research questions and sub-questions.

Table 6. Summary of research questions, hypotheses, and results

RQ	Hypothesis tested	Statistic	*p*	Effect size	Decision
RQ1	Achievement Time × Group	$F(1.88, 231.24) = 15.42$	< .001	partial $\eta^2 = 0.11$	Supported
RQ2	Attitudes Exp. versus Ctrl. (six dimensions)	See Table 4	Mixed	$\lvert d \rvert = 0.31$ to 0.97	Supported (4 of 6)
RQ3a	Usage to outcome dose-response	Not formally tested	N/A	N/A	No automated data available
RQ3b	Differential usage by proficiency	$\chi^2(3) = 108.45$	< .001	$V = 0.37$	Preliminary support
RQ4	Collaborative quality	Field notes (descriptive)	N/A	N/A	Preliminary support
RQ5a	Achievement gap narrows	$F(1, 120) = 6.30$	.013	partial $\eta^2 = 0.050$	Supported
RQ5b	Exploratory: within-experimental-group proficiency differences in attitudes†	$F(1, 62) = 5.44$ / 3.79	.023 / .046	$\omega^2 = 0.07$ / 0.08	Partially supported
RQ6a	Gender main effect on achievement (no Group × Gender interaction tested)	$F(1, 122) = 0.85$	.358	$\omega^2 \approx 0$	No significant gender effect detected
RQ6b	Gender main effect on anxiety (no Group × Gender interaction tested)†	$F(1, 122) = 5.45$	.021	$\omega^2 = 0.04$	Inconclusive (low subscale reliability)

Note. Partial η^2 is reported for omnibus and interaction effects; ω^2 is reported for subgroup analyses. RQ5b values F(1, 62) reflect within-experimental-group ANCOVAs (n = 65) examining whether language proficiency moderated attitudinal change in the experimental condition; they do not represent a Group × Proficiency interaction. †RQ5b (Subject Anxiety component) and RQ6b findings are preliminary owing to low subscale reliability ($\alpha = 0.574$); see subsection "Learning Attitudes" under Instruments.

Synthesis

Together, these findings address RQ1 through RQ6. They demonstrate differential gains across proficiency strata, attitudinal shifts, and narrowed proficiency-based subgroup differences in achievement. Residual gender-related affective barriers warrant further investigation.

Discussion

Equity Narrowing as the Headline Finding (RQ5a)

Building on the descriptive findings reported above, this section interprets the evidence in light of the conceptual model and the existing literature. The most consequential finding is that the

proficiency-based achievement gap narrowed from 16.60 to 4.00 points in the experimental group (Group × Proficiency interaction, partial $\eta^2 = 0.050$). To our knowledge, and acknowledging that any structured search has limits, this study provides initial subgroup-level evidence combining four design features within one study: (i) the STAD cooperative-learning architecture, (ii) GPT-4o as the AI scaffold, (iii) a content-area Mandarin-medium higher-education course, and (iv) proficiency-stratified analysis of achievement-gap closure in a multilingual EFL undergraduate cohort. Each feature has been examined in adjacent literatures (e.g., AI in language-learning courses, AI in cooperative learning at the aggregate level, equity in EMI), but their joint operationalisation with subgroup-level equity evidence appears not to have been previously reported.

This novelty claim is supported by a structured literature search executed by the authors. The search covered Google Scholar, Semantic Scholar, and ERIC-indexed sources from the public release of ChatGPT (30 November 2022) through April 2026. Four Boolean strings were used: (a) ("ChatGPT" OR "generative AI" OR "GPT-4") AND ("STAD" OR "Student Teams Achievement Divisions") AND ("equity" OR "achievement gap" OR "proficiency gap"); (b) "ChatGPT" AND "cooperative learning" AND "language proficiency" AND "content-area"; (c) "ChatGPT" AND "Student Teams Achievement Divisions"; and (d) "AI" AND "cooperative learning" AND "EFL" AND "achievement gap". The search returned no peer-reviewed empirical study that simultaneously met three conditions: (i) deployed a generative-AI tool within a STAD cooperative-learning architecture; (ii) operationalised the intervention in a content-area higher-education setting (i.e., not language instruction itself); and (iii) reported subgroup-level outcomes disaggregated by learner language proficiency.

Adjacent empirical work has examined related topics, but not their joint operationalisation. Examples include AI-supported collaborative learning at the aggregate level (e.g., H. Wang et al., 2025), AI-assisted writing in EFL contexts without cooperative-learning structures (e.g., Tsai et al., 2024), and aggregate cognitive-load and motivation effects of AI-enhanced language platforms (e.g., Hong & Guo, 2025). To our knowledge, and acknowledging the limits of any structured search, the present study therefore offers initial subgroup-level evidence, awaiting multisite replication, that AI-enhanced STAD is associated with the narrowing of proficiency-based achievement gaps in a content-area cooperative-learning context. Given the perfect confounding between condition and class noted in the Methodology section, this association should not be interpreted as established causal evidence. This evidence speaks directly to a long-standing concern in linguistically diverse higher education. Cooperative tasks can paradoxically widen rather than close participation gaps when language proficiency operates as a status characteristic.

The finding suggests, rather than proves, a sociocultural mechanism. During Phase 2 (individual preparation), lower-proficiency learners appear to have used the AI tool to pre-formulate contributions and clarify unfamiliar terminology. They also rehearsed ideas in a low-stakes environment before entering group discussion. The differential prompting pattern observed in the reflection-log data is consistent with the contingency principle of effective scaffolding (Belland, 2014). Specifically, Cramér's $V = 0.37$; lower-proficiency learners reported 45.2% language-clarification queries versus 12.5% for higher-proficiency learners. Those furthest from independent mastery sought the most scaffolding support. We interpret this as an indication, not confirmation, that the AI scaffold functioned as a supplementary, more knowledgeable other (MKO) within the zone of proximal development (ZPD) of Vygotsky (1978).

Achievement and Attitudinal Gains (RQ1, RQ2)

Regarding **RQ1**, the experimental group exhibited significantly steeper achievement growth ($d = 0.91$), consistent with gains reported in AI-agent-supported learning (H. Wang et al., 2025)

and generative AI scaffolding in STEM (Chen et al., 2025). Cognitive load theory (CLT) provides a complementary explanation. By offloading linguistic processing tasks to the AI scaffold, lower-proficiency learners may have freed working-memory resources for content learning and higher-order thinking. The disproportionate gains in this subgroup align with this CLT-based interpretation and with Hong and Guo's (2025) findings on generative AI and cognitive-load management.

Regarding RQ2, four of six attitude dimensions improved significantly. The large effect on Learning Confidence (d = 0.97) is interpretable through both frameworks. The AI scaffold expanded the range of tasks that lower-proficiency learners could accomplish successfully, reinforcing perceived competence (sociocultural mechanism). Simultaneously, the reduction in extraneous linguistic load reduced frustration and overload (CLT mechanism). These experiences typically diminish self-efficacy (Eysenck & Calvo, 1992). The smaller but significant gain in Success Orientation suggests that metacognitive scaffolding through Tier 3 prompts and reflection logs may have contributed to learners' sense of agency and strategic awareness.

Unexpected and Contradictory Findings

Two findings deviated from the hypothesised pattern. Motivation for Exploration (Dimension C) did not reach Bonferroni-adjusted significance (p = .063); the 95% CI for d [-0.04, 0.67] marginally included zero at the lower bound, indicating that a null effect cannot be ruled out. A plausible explanation is a ceiling effect. Both groups began with relatively high motivation (M > 3.38 on a 5-point scale), which left limited room for growth. The STAD structure itself may also have promoted exploratory motivation in both conditions, attenuating the incremental ChatGPT effect. Subject Anxiety (Dimension F) similarly did not survive Bonferroni correction (d = -0.31; CI crossed zero). The hypothesis that the intervention would reduce anxiety was not confirmed. This suggests that affective anxiety mechanisms may require a more targeted or longer intervention.

The finding that female students reported higher residual learning anxiety (RQ6b) warrants attention as a design limitation. It sits within the same equity framing as the proficiency-narrowing finding: while the intervention closes one equity gap, it does not address another. Several alternative explanations merit consideration. First, prior literature attributes gender differences in technology-related anxiety to socialisation patterns and stereotype threat (Cai et al., 2017; Huang, 2013). The unstructured nature of AI interaction may activate these anxieties differentially. Second, female students in this cohort may have engaged more deeply with reflection prompts and thereby reported more affective awareness, rather than experiencing more anxiety per se. This self-report-bias account cannot be ruled out without behavioural data. Third, sample-specific cultural and disciplinary characteristics (e.g., the gender composition of marketing programmes, regional norms around technology use) may interact with intervention features in ways that single-site studies cannot disentangle. The Foreign Language Classroom Anxiety Scale (Horwitz et al., 1986) and related instruments would provide a more theoretically grounded measure for future cycles. Future iterations should incorporate explicit norm-setting, optional affective-support prompts within Tier 2 scaffolding, and peer mentoring to support learners experiencing heightened anxiety related to technology.

Positioning Within the Literature

The present findings both corroborate and extend recent studies. The overall achievement effect (d = 0.91) is comparable to gains reported by H. Wang et al. (2025) and Chen et al. (2025). The present study extends these findings in two ways. First, to our knowledge, the equity evidence is the initial subgroup-level demonstration suggesting that AI-enhanced cooperative learning may be associated with the narrowing of achievement gaps for linguistically disadvantaged learners in a content-area course. This addresses a recognised gap in the literature, although confirmatory multi-site replication is required. Second, whereas existing studies typically

evaluate AI as a standalone supplement, the present study embeds ChatGPT within a structured STAD model with phase restrictions, governance, and progressive scaffolding fading. This suggests that deliberate AI integration may serve as an equity-promoting tool when differentiated by learner need. The positive correlation between SUS scores and achievement gains ($r = .35$) provides indirect proxy evidence for the dose-response relationship proposed in RQ3a. Systematic log data would be needed to substantiate the claim.

Policy Call to Action

Taken together, the findings suggest that institutional and national AI-use policies in linguistically diverse higher-education systems would benefit from a shift away from blanket prohibition toward governance-embedded design. In such a design, AI access is deliberately structured at specific pedagogical interaction nodes, differentiated by learner need, and accompanied by transparent governance mechanisms. The present study provides an initial empirical basis for such a shift; cross-institutional and longitudinal replication is now needed.

IMPLICATIONS

Theoretical Implications

The study contributes to educational technology theory in three ways. First, it provides empirical evidence for conceptualising generative artificial intelligence as a more knowledgeable other (MKO) within Vygotsky's (1978) zone of proximal development (ZPD). This extends sociocultural scaffolding theory to AI-mediated learning environments. The differential usage patterns across proficiency subgroups are consistent with the contingency principle of effective scaffolding (Belland, 2014). Second, it provides empirical support for the cognitive load theory (CLT) prediction that reducing extraneous linguistic load through AI scaffolding can free cognitive resources for germane processing. This produces disproportionate achievement gains for learners with the highest extraneous load (Sweller et al., 2019). Third, it contributes to the cooperative learning literature by demonstrating that AI tools can be integrated as a functional element within structured cooperative learning models. Such integration complements positive interdependence, individual accountability, and promotive interaction (Johnson et al., 1998).

Practical Implications

Five recommendations emerge for practitioners integrating generative AI into cooperative learning with linguistically diverse students. First, AI access should be embedded at specific interaction nodes rather than offered in an always-on manner; phase-structured use preserves pedagogical integrity during accountability phases while providing support where most needed. Second, scaffolding should be differentiated by learner proficiency through tiered prompt protocols; pre-designed Tier 2 prompts reduce linguistic and prompt-engineering barriers for lower-proficiency learners. Third, AI use should be governed through explicit rules, reflection logs, and teacher monitoring to mitigate over-reliance, consistent with Güner and Er (2025). Fourth, scaffolding fading should be planned from the outset, with Tier 3 prompts encouraging independent attempts at the task before AI consultation. Fifth, affective outcomes should be monitored for subgroups experiencing technology-related anxiety; norm-setting activities and gender-responsive affective-support prompts are recommended.

The findings also carry policy implications. The equity evidence suggests that blanket prohibitions of AI tools may inadvertently disadvantage linguistically diverse students who stand to benefit most from AI scaffolding. The governance framework developed here, comprising usage agreements, prompt protocols, reflection logs, and AI-free assessments, provides an adaptable model for institutional AI-use policies in diverse educational settings.

LIMITATIONS AND FUTURE RESEARCH

Several limitations must be acknowledged.

First, the quasi-experimental design used two pre-existing intact classes at a single Taiwanese university, with one class assigned to the ChatGPT-enhanced STAD condition and the other to the traditional STAD condition. Because students were not randomly assigned between conditions, treatment effects and class-level effects (instructor expectancy, scheduling, peer-cohort dynamics, Hawthorne effects) are perfectly confounded at the class level and cannot be statistically disentangled in this single-site sample. Multi-site, multi-class replication with random assignment at the class or cluster level is therefore needed before the present findings can be interpreted as design-attributable rather than class-attributable. This is the most important methodological constraint of the present study.

Second, the sample size (N = 125) was adequate for detecting observed main effects but limited the power of fully crossed Group × Gender × Proficiency analyses.

Third, the Subject Anxiety subscale yielded $\alpha = 0.574$, below the conventional 0.70 threshold (Nunnally & Bernstein, 1994). All findings involving Subject Anxiety (Table 4, Dimension F; RQ5b proficiency comparison; RQ6b gender comparison) are therefore preliminary. Item-level revision using validated instruments such as the Foreign Language Classroom Anxiety Scale (Horwitz et al., 1986) is planned. Future replications will also report McDonald's ω alongside Cronbach's α (McDonald, 1999).

Fourth, the participant-to-item ratio for the CFA was approximately 1:2.4 (N = 130; 55 items). This falls below the conventional 5:1 minimum and the preferred 10:1 ratio (Hair et al., 2019; Tabachnick & Fidell, 2019). Following Kline's (2023) guidance that adequate sample size depends on indicator reliability and factor definition rather than a rigid ratio, the present six-factor solution is acceptable but suboptimal. Future replication of the measurement model with N ≥ 300 (ratio ≥ 5:1) is warranted before strong inferences are drawn from the factor structure.

Fifth, automated, timestamped interaction-process data were not collected. The assessment of collaborative interaction quality (RQ4) relied on instructor field notes rather than systematic quantitative coding. Claims about AI usage are therefore based on self-reports and observer-derived indicators, which limit causal inference. Relatedly, the chi-square test for differential prompting purposes (RQ3b) treated the 780 reflection-log entries as independent, although they were nested within 65 students. This violation of the independence assumption may inflate the test statistic and underestimate p-values, although Cramér's V remains valid as a descriptive effect-size indicator. The RQ3b finding is therefore reported as preliminary; future cycles will aggregate prompting purposes at the student level (n = 65) or apply multilevel categorical models. Establishing the full causal chain from design features to outcomes will require both automated logs and validated coding schemes. Future cycles will incorporate a triangulated CSCL coding architecture drawing on De Wever et al. (2006) for methodological orientation, Gunawardena et al. (1997) for the social construction of knowledge, Garrison et al. (2001) for cognitive presence, and Hmelo-Silver and Barrows (2008) for facilitation analysis.

Sixth, the study employed GPT-4o (OpenAI, 2024); future replications should compare outcomes systematically across model versions.

Seventh, strict scalar measurement invariance was not feasible owing to subgroup $n < 100$ (Vandenberg & Lance, 2000); only configural and partial metric invariance could be established. Interpretation of subgroup attitudinal comparisons should be cautious accordingly.

Eighth, the cohort was drawn from a single Taiwanese university, where first-year marketing students use Mandarin as the medium of instruction, with linguistic proficiency stratified by TOCFL band. The term "multilingual higher education" used in the title therefore denotes a

specific configuration: Mandarin-medium content instruction with TOCFL-stratified learners predominantly from East and Southeast Asia. It should not be over-extrapolated to typologically diverse plurilingual contexts (e.g., European Erasmus settings, sub-Saharan African plurilingual universities) without dedicated replication. Replication across STEM, social-science, and professional-education contexts, and across genuinely diverse linguistic configurations, is required before broad generalisability claims can be made.

Four future research directions emerge from these limitations: (a) automated AI usage logs with timestamped prompting-purpose coding will enable dose-response analysis and validation of proposed mechanisms; (b) cross-institutional, cross-disciplinary replications are needed across diverse linguistic contexts and educational stages; (c) the Subject Anxiety subscale will be revised with rigorous CFA on a larger sample. McDonald's ω will be reported alongside Cronbach's α, and gender-responsive scaffolding will be incorporated and evaluated; (d) longitudinal follow-up will assess whether gains and equity effects persist post-intervention, and whether progressive scaffolding fading produces sustained independent performance; and (e) collaborative interaction quality (RQ4) will be measured through a triangulated CSCL coding architecture. This architecture combines De Wever et al.'s (2006) methodological review, Gunawardena et al.'s (1997) interaction analysis model, Garrison et al.'s (2001) practical inquiry framework, and Hmelo-Silver and Barrows's (2008) collaborative-knowledge-building model, rather than relying solely on instructor field notes.

CONCLUSION

This study investigated whether a ChatGPT-enhanced cooperative learning intervention could improve academic achievement, learning attitudes, and interaction equity among linguistically diverse university students. The evidence addresses some research questions directly and others more indirectly: RQ1, RQ2, RQ5a, and RQ6a are answered through formal hypothesis testing, whereas RQ3a (dose-response), RQ4 (collaborative interaction quality), RQ3b (differential prompting purposes), RQ5b (proficiency-moderated attitudes), and RQ6b (gender-moderated intervention effects) are supported only by preliminary, exploratory, or self-reported evidence and warrant replication with automated process data and adequately powered factorial designs. With this calibration in mind, the study's findings can be summarised as follows.

RQ1 (achievement): achievement trajectories diverged progressively, reaching a large effect at posttest ($d = 0.91$). RQ2 (attitudes): four of six attitude dimensions improved significantly, led by Learning Confidence ($d = 0.97$); the subject anxiety dimension is treated as preliminary owing to low subscale reliability. RQ3a (dose-response): this could not be rigorously tested without automated logs; the SUS-achievement correlation provides an indirect proxy ($r = .35$), and the finding is therefore exploratory. RQ3b (differential usage): self-reported prompting purposes differed substantially by proficiency (Cramér's $V = 0.37$); lower-proficiency learners reported substantially more language-clarification queries; this finding is preliminary owing to the independence-assumption violation discussed in Limitations.

RQ4 (collaborative quality): the instructor's field notes provide preliminary, descriptive, observer-derived evidence of improved engagement among lower-proficiency students. RQ5a (achievement gap): the proficiency-based achievement gap narrowed from 16.60 to 4.00 points (Group × Proficiency partial $\eta^2 = 0.050$). To our knowledge, and as supported by the structured Boolean literature search reported in the Discussion subsection "Equity Narrowing as the Headline Finding." This provides initial subgroup-level evidence, awaiting multi-site replication, that AI-enhanced STAD may be associated with the compression of proficiency-based achievement disparities in content-area cooperative learning within a Mandarin-medium TOCFL-stratified configuration. Given the perfect confounding between condition and class noted in the Methodology, this finding should be read as a starting hypothesis for confirmatory

replication rather than as established causal evidence. RQ5b (within-experimental-group proficiency differences in attitudes): an exploratory within-experimental-group analysis indicated that lower-proficiency students showed greater improvement in Learning Confidence; the Subject Anxiety component is preliminary, and this finding does not constitute a direct test of differential intervention effects on attitudes relative to controls. RQ6a (gender main effect on achievement): no significant gender difference in achievement was detected. RQ6b (gender main effect on anxiety, exploratory): within the same equity framing as RQ5a, female students reported higher residual subject anxiety. This indicates that the design does not address all affective barriers equally; this finding is exploratory and preliminary owing to subscale reliability and the absence of a formal Group × Gender interaction test.

While these findings are based on a single-site study and warrant cross-institutional replication, they provide an initial evidence base for design-focused AI-use policies in linguistically diverse higher education. Institutions should adopt this governance-embedded design, integrating phase-structured AI access, tiered prompting, scaffolding fading, and AI-free accountability, to scale equity in multilingual classrooms.

Future cycles will refine progressive scaffolding-fading protocols for sustained learner independence. They will deploy automated usage logs to establish causal mechanisms. They will revise the Subject Anxiety measurement instrument with rigorous CFA on a larger sample. Finally, they will incorporate gender-responsive features to address residual affective barriers.

ACKNOWLEDGEMENT OF AI ASSISTANCE IN WRITING

During the preparation of this work, the authors used ChatGPT and Gemini solely to improve language readability and assist in formatting conceptual diagrams. The authors thoroughly reviewed and edited the manuscript and take full responsibility for its content, accuracy, and originality.

REFERENCES

Adams, W. K., Perkins, K. K., Podolefsky, N. S., Dubson, M., Finkelstein, N. D., & Wieman, C. E. (2006). New instrument for measuring student beliefs about physics and learning physics: The Colorado Learning Attitudes about Science Survey. *Physical Review Special Topics: Physics Education Research, 2*(1), 010101. https://doi.org/10.1103/PhysRevSTPER.2.010101

Aizawa, I., Rose, H., Thompson, G., & Curle, S. (2023). Beyond the threshold: Exploring English language proficiency, linguistic challenges, and academic language skills of Japanese students in an English medium instruction programme. *Language Teaching Research, 27*(4), 837–861. https://doi.org/10.1177/1362168820965510

Andrade, H., & Valtcheva, A. (2009). Promoting learning and achievement through self-assessment. *Theory Into Practice, 48*(1), 12–19. https://doi.org/10.1080/00405840802577544

Bangor, A., Kortum, P. T., & Miller, J. T. (2008). An empirical evaluation of the System Usability Scale. *International Journal of Human–Computer Interaction*, *24*(6), 574–594. https://doi.org/10.1080/10447310802205776

Belland, B. R. (2014). Scaffolding: Definition, current debates, and future directions. In J. M. Spector, M. D. Merrill, J. Elen, & M. J. Bishop (Eds.), *Handbook of research on educational communications and technology* (4th ed., pp. 505–518). Springer. https://doi.org/10.1007/978-1-4614-3185-5_39

Brooke, J. (1996). SUS: A "quick and dirty" usability scale. In P. W. Jordan, B. Thomas, B. A. Weerdmeester, & I. L. McClelland (Eds.), *Usability evaluation in industry* (pp. 189–194). Taylor & Francis.

Buchs, C., Filippou, D., Pulfrey, C., & Volpé, Y. (2017). Challenges for cooperative learning implementation: Reports from elementary school teachers. *Journal of Education for Teaching, 43*(3), 296–306. https://doi.org/10.1080/02607476.2017.1321673

Cai, Z., Fan, X., & Du, J. (2017). Gender and attitudes toward technology use: A meta-analysis. *Computers & Education, 105*, 1–13. https://doi.org/10.1016/j.compedu.2016.11.003

Chen, S.-Y., Chen, W.-C., & Lai, C.-F. (2025). Generative AI as a reflective scaffold in a UAV-based STEM project: A mixed-methods study on students' higher-order thinking and cognitive transformation. *Education and Information Technologies, 30*, 24787–24814. https://doi.org/10.1007/s10639-025-13758-4

Cheung, G. W., & Rensvold, R. B. (2002). Evaluating goodness-of-fit indexes for testing measurement invariance. *Structural Equation Modeling, 9*(2), 233–255. https://doi.org/10.1207/S15328007SEM0902_5

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Lawrence Erlbaum Associates.

De Wever, B., Schellens, T., Valcke, M., & Van Keer, H. (2006). Content analysis schemes to analyze transcripts of online asynchronous discussion groups: A review. *Computers & Education, 46*(1), 6–28. https://doi.org/10.1016/j.compedu.2005.04.005

Eysenck, M. W., & Calvo, M. G. (1992). Anxiety and performance: The processing efficiency theory. *Cognition and Emotion, 6*(6), 409–434. https://doi.org/10.1080/02699939208409696

Faul, F., Erdfelder, E., Buchner, A., & Lang, A. G. (2009). Statistical power analyses using G*Power 3.1: Tests for correlation and regression analyses. *Behavior Research Methods, 41*(4), 1149–1160. https://doi.org/10.3758/BRM.41.4.1149

Garrison, D. R., Anderson, T., & Archer, W. (2001). Critical thinking, cognitive presence, and computer conferencing in distance education. *American Journal of Distance Education, 15*(1), 7–23. https://doi.org/10.1080/08923640109527071

Gibbons, P. (2015). *Scaffolding language, scaffolding learning: Teaching English language learners in the mainstream classroom* (2nd ed.). Heinemann.

Gillies, R. M., & Boyle, M. (2010). Teachers' reflections on cooperative learning: Issues of implementation. *Teaching and Teacher Education, 26*(4), 933–940. https://doi.org/10.1016/j.tate.2009.10.034

Gunawardena, C. N., Lowe, C. A., & Anderson, T. (1997). Analysis of a global online debate and the development of an interaction analysis model for examining social construction of knowledge in computer conferencing. *Journal of Educational Computing Research, 17*(4), 397–431. https://doi.org/10.2190/7MQV-X9UJ-C7Q3-NRAG

Güner, H., & Er, E. (2025). AI in the classroom: Exploring students' interaction with ChatGPT in programming learning. *Education and Information Technologies, 30*(9), 12681–12707. https://doi.org/10.1007/s10639-025-13337-7

Hair, J. F., Black, W. C., Babin, B. J., & Anderson, R. E. (2019). *Multivariate data analysis* (8th ed.). Cengage.

Hmelo-Silver, C. E., & Barrows, H. S. (2008). Facilitating collaborative knowledge building. *Cognition and Instruction, 26*(1), 48–94. https://doi.org/10.1080/07370000701798495

Hong, X., & Guo, L. (2025). Effects of AI-enhanced multi-display language teaching systems on learning motivation, cognitive load management, and learner autonomy. *Education and Information Technologies, 30*(12), 17155–17189. https://doi.org/10.1007/s10639-025-13472-1

Horwitz, E. K., Horwitz, M. B., & Cope, J. (1986). Foreign language classroom anxiety. *The Modern Language Journal, 70*(2), 125–132. https://doi.org/10.1111/j.1540-4781.1986.tb05256.x

Huang, C. (2013). Gender differences in academic self-efficacy: A meta-analysis. *European Journal of Psychology of Education, 28*, 1–35. https://doi.org/10.1007/s10212-011-0097-y

Jeong, H., & Hmelo-Silver, C. E. (2016). Seven affordances of computer-supported collaborative learning: How to support collaborative learning? How can technologies help? *Educational Psychologist, 51*(2), 247–265. https://doi.org/10.1080/00461520.2016.1158654

Ji, Z., Lee, N., Frieske, R., Yu, T., Su, D., Xu, Y., Ishii, E., Bang, Y. J., Madotto, A., & Fung, P. (2023). Survey of hallucination in natural language generation. *ACM Computing Surveys, 55*(12), Article 248. https://doi.org/10.1145/3571730

Johnson, D. W., Johnson, R. T., & Smith, K. A. (1998). Cooperative learning returns to college: What evidence is there that it works? *Change: The Magazine of Higher Learning, 30*(4), 26–35. https://doi.org/10.1080/00091389809602629

Karataş, F., Abedi, F. Y., Ozek Gunyel, F., Karadeniz, D., & Kuzgun, Y. (2024). Incorporating AI in foreign language education: An investigation into ChatGPT's effect on foreign language learners. *Education and Information Technologies, 29*(15), 19343–19366. https://doi.org/10.1007/s10639-024-12574-6

Kasneci, E., Sessler, K., Küchemann, S., Bannert, M., Dementieva, D., Fischer, F., Gasser, U., Groh, G., Günnemann, S., Hüllermeier, E., Krusche, S., Kutyniok, G., Michaeli, T., Nerdel, C., Pfeffer, J., Poquet, O., Sailer, M., Schmidt, A., Seidel, T., … Kasneci, G. (2023). ChatGPT for good? On opportunities and challenges of large language models for education. *Learning and Individual Differences, 103*, 102274. https://doi.org/10.1016/j.lindif.2023.102274

Keller, K. L. (2013). *Strategic brand management: Building, measuring, and managing brand equity* (4th ed.). Pearson.

Killip, S., Mahfoud, Z., & Pearce, K. (2004). What is an intracluster correlation coefficient? Crucial concepts for primary care researchers. *Annals of Family Medicine, 2*(3), 204–208. https://doi.org/10.1370/afm.141

Kisanga, D. H., & Ireson, G. (2016). Test of e-Learning Related Attitudes (TeLRA) scale: Development, reliability and validity study. *International Journal of Education and Development using Information and Communication Technology, 12*(1), 20–36.

Kline, R. B. (2023). *Principles and practice of structural equation modeling* (5th ed.). Guilford Press.

Kyndt, E., Raes, E., Lismont, B., Timmers, F., Cascallar, E., & Dochy, F. (2013). A meta-analysis of the effects of face-to-face cooperative learning. Do recent studies falsify or verify earlier findings? *Educational Research Review, 10*, 133–149. https://doi.org/10.1016/j.edurev.2013.02.002

Lai, M. H. C., & Kwok, O. M. (2015). Examining the rule of thumb of not using multilevel modeling: The "design effect smaller than two" rule. *The Journal of Experimental Education, 83*(3), 423–438. https://doi.org/10.1080/00220973.2014.907229

Le, H., Janssen, J., & Wubbels, T. (2018). Collaborative learning practices: Teacher and student perceived obstacles to effective student collaboration. *Cambridge Journal of Education, 48*(1), 103–122. https://doi.org/10.1080/0305764X.2016.1259389

Little, R. J. A., & Rubin, D. B. (2019). *Statistical analysis with missing data* (3rd ed.). Wiley. https://doi.org/10.1002/9781119482260

McDonald, R. P. (1999). *Test theory: A unified treatment*. Lawrence Erlbaum.

Mohamed, A. M. (2024). Exploring the potential of an AI-based chatbot (ChatGPT) in enhancing English as a Foreign Language (EFL) teaching: Perceptions of EFL faculty members. *Education and Information Technologies, 29*(3), 3195–3217. https://doi.org/10.1007/s10639-023-11917-z

Nunnally, J. C., & Bernstein, I. H. (1994). *Psychometric theory* (3rd ed.). McGraw-Hill.

Olejnik, S., & Algina, J. (2003). Generalized eta and omega squared statistics: Measures of effect size for some common research designs. *Psychological Methods, 8*(4), 434–447. https://doi.org/10.1037/1082-989X.8.4.434

OpenAI. (2024). *GPT-4o system card.* https://openai.com/index/gpt-4o-system-card/

Organisation for Economic Co-operation and Development. (2024). *Education at a glance 2024: OECD indicators*. https://doi.org/10.1787/c00cad36-en

Puntambekar, S., & Hübscher, R. (2005). Tools for scaffolding students in a complex learning environment: What have we gained and what have we missed? *Educational Psychologist, 40*(1), 1–12. https://doi.org/10.1207/s15326985ep4001_1

Roshanaei, M. (2024). Towards best practices for mitigating artificial intelligence implicit bias in shaping diversity, inclusion and equity in higher education. *Education and Information Technologies, 29*(14), 18959–18984. https://doi.org/10.1007/s10639-024-12605-2

Sahan, K., Galloway, N., & McKinley, J. (2022). 'English-only' English medium instruction: Mixed views in Thai and Vietnamese higher education. *Language Teaching Research, 29*(2), 657–676. https://doi.org/10.1177/13621688211072632

Sahan, K., Kamaşak, R., & Rose, H. (2023). The interplay of motivated behaviour, self-concept, self-efficacy, and language use on ease of academic study in English medium education. *System, 114*, 103016. https://doi.org/10.1016/j.system.2023.103016

Slavin, R. E. (1983). When does cooperative learning increase student achievement? *Psychological Bulletin, 94*(3), 429–445. https://doi.org/10.1037/0033-2909.94.3.429

Slavin, R. E. (1996). Research on cooperative learning and achievement: What we know, what we need to know. *Contemporary Educational Psychology, 21*(1), 43–69. https://doi.org/10.1006/ceps.1996.0004

Snijders, T. A. B., & Bosker, R. J. (2012). *Multilevel analysis: An introduction to basic and advanced multilevel modeling* (2nd ed.). Sage.

Steering Committee for the Test of Proficiency-Huayu. (2023). *Test of Chinese as a Foreign Language (TOCFL): Vocabulary list and proficiency band descriptors*. Ministry of Education, Republic of China (Taiwan). https://tocfl.edu.tw/

Sweller, J. (1988). Cognitive load during problem solving: Effects on learning. *Cognitive Science, 12*(2), 257–285. https://doi.org/10.1207/s15516709cog1202_4

Sweller, J., van Merriënboer, J. J. G., & Paas, F. (2019). Cognitive architecture and instructional design: 20 years later. *Educational Psychology Review, 31*(2), 261–292. https://doi.org/10.1007/s10648-019-09465-5

Tabachnick, B. G., & Fidell, L. S. (2019). *Using multivariate statistics* (7th ed.). Pearson.

Tamim, R. M., Bernard, R. M., Borokhovski, E., Abrami, P. C., & Schmid, R. F. (2011). What forty years of research says about the impact of technology on learning: A second-order meta-analysis and validation study. *Review of Educational Research, 81*(1), 4–28. https://doi.org/10.3102/0034654310393361

Tapia, M., & Marsh, G. E. (2004). An instrument to measure mathematics attitudes. *Academic Exchange Quarterly, 8*(2), 16–21.

Tsai, C. Y., Lin, Y. T., & Brown, I. K. (2024). Impacts of ChatGPT-assisted writing for EFL English majors: Feasibility and challenges. *Education and Information Technologies, 29*(17), 22427–22445. https://doi.org/10.1007/s10639-024-12722-y

Vandenberg, R. J., & Lance, C. E. (2000). A review and synthesis of the measurement invariance literature: Suggestions, practices, and recommendations for organizational research. *Organizational Research Methods, 3*(1), 4–70. https://doi.org/10.1177/109442810031002

Vygotsky, L. S. (1978). *Mind in society: The development of higher psychological processes*. Harvard University Press.

Wang, F., & Hannafin, M. J. (2005). Design-based research and technology-enhanced learning environments. *Educational Technology Research and Development, 53*(4), 5–23. https://doi.org/10.1007/BF02504682

Wang, H., Wang, C., Chen, Z., Liu, F., Bao, C., & Xu, X. (2025). Impact of AI-agent-supported collaborative learning on the learning outcomes of university programming courses. *Education and Information Technologies, 30*(12), 17717–17749. https://doi.org/10.1007/s10639-025-13487-8

Woo, D. J., Susanto, H., Yeung, C. H., Guo, K., & Fung, A. K. Y. (2024). Exploring AI-generated text in student writing: How does AI help? *Language Learning & Technology, 28*(2), 183–209. https://doi.org/10.64152/10125/73577

Wood, D., Bruner, J. S., & Ross, G. (1976). The role of tutoring in problem solving. *Journal of Child Psychology and Psychiatry, 17*(2), 89–100. https://doi.org/10.1111/j.1469-7610.1976.tb00381.x

Yan, L., Sha, L., Zhao, L., Li, Y., Martinez-Maldonado, R., Chen, G., Li, X., Jin, Y., & Gašević, D. (2024). Practical and ethical challenges of large language models in education: A systematic scoping review. *British Journal of Educational Technology, 55*(1), 90–112. https://doi.org/10.1111/bjet.13370

Zhu, Z. T., Yu, M. H., & Riezebos, P. (2016). A research framework of smart education. *Smart Learning Environments, 3*, Article 4. https://doi.org/10.1186/s40561-016-0026-2

AUTHORS

Dr. Yu-Heng Chen is an Assistant Professor in the Department of Marketing and Distribution Management at Chaoyang University of Technology, Taiwan. Driven by a deep commitment to education, she brings to her scholarship a diverse portfolio of professional experience spanning the business, technology, entertainment, and health sectors. Her research centres on international student education, with a particular focus on translating empirical findings into actionable insights for educators and learners navigating cross-border academic environments. Dr. Chen's interdisciplinary academic foundations in education, innovation and entrepreneurship, and business, inform her work on cooperative learning, AI-assisted instruction, and equity in linguistically diverse higher education.

Dr. Kimyung Keng is an Assistant Professor in the Department of Global Politics and Economics at Tamkang University, Taiwan, where he serves as a dedicated instructor of English-Medium Instruction (EMI) courses. His primary research interests centre on EMI pedagogy, instructional methodology, second-language teaching, and experiential learning, with a particular focus on translating classroom practice into evidence-based pedagogical frameworks. Beyond education, Dr. Keng investigates the implications of regional geopolitics for technology development and industrial transformation, examining how political dynamics shape innovation trajectories and cross-border industrial linkages. This dual research agenda enables him to bridge pedagogical inquiry with applied policy analysis.