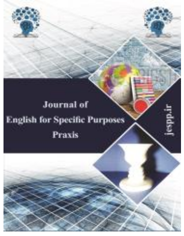
Journal of English for Specific Purposes Praxis ISSN: 3092-7854
https://www.jespp.ir
Vol. 2, No. 3; 2025, July
JESPP



Article Citation:
Alamdar, F. S., Mohammadi, A. H., Afraz, SH., Samimi, F., & Hobouti, M. (2025). Flipped classroom vs. AI-enhanced instruction: Effects on self-monitoring and writing in Iranian EFL learners. *Journal of English for Specific Purposes Praxis, 2*(3), 39–64. https://doi.org/10.22034/jespp.2025.542411.1017

Flipped Classroom vs. AI-Enhanced Instruction: Effects on Self-Monitoring and Writing in Iranian EFL Learners

Fatemeh Sadat Alamdar [1,*], Abdoul Hamid Mohammadi [2], Shahram Afraz [3], Fazlollah Samimi [4], Maryam Hobouti [5]

[1,*]Corresponding author, Department of Education, Ministry of Education, Hormozgan, Iran; f.alamdar67@gmail.com

[2] Faculty Member, Bandar Abbas University of Medical Sciences, Hormozgan, Iran; abdolhamid.mohammadi403@gmail.com

[3] Department of English Language Teaching, Qe.C., Islamic Azad University, Qeshm, Iran; fshahram.afraz@iau.ac.ir

[4] Department of ELT, BA.C., Islamic Azad University, Bandar Abbas, Iran; fazlolahsamimi@iau.ac.ir

[5] Department of Education, Ministry of Education, Gilan, Iran; hoboutimaryam@gmail.com

Original Research Article	Date of Submission: 16 May 2025	Date of Acceptance: 20 July 2025

Abstract

English Foreign Language (EFL) instruction with artificial intelligence (AI) has raised new issues about its value in comparison to the traditional models, including the flipped classroom. Comparative value of AI-enhanced learning to flipped instruction for Iranian EFL learners' ability to self-monitor was investigated in this study. A quasi-experimental design was employed with 120 upper-intermediate learners, who were recruited through simple random sampling and allocated into three groups: control, flipped classroom, and AI-supported learning. The control group received the usual instruction for eight weeks, whereas the intervention groups were exposed to particular interventions. Assessment was carried out by using standardized self-monitoring questionnaires and writing tasks. Analysis of covariance (ANCOVA) showed that both the flipped and the AI-supported conditions were highly superior to the control condition for learning self-monitoring ($p < 0.001$, partial $\eta^2 = 0.924$). Additionally, the AI-supported condition demonstrated superior overall gains in self-monitoring compared to the flipped classroom condition. These findings suggest that while flipped classrooms enable autonomy with teacher support, AI systems enable adaptive scaffolding and learner-specific feedback, which more efficiently empower learners' strategic self-regulation abilities. The study makes theoretical contributions to learner autonomy frameworks and empirical contributions to CALL and self-regulation, demonstrating how AI can be leveraged to augment human teaching.

Keywords: AI-enhanced instruction, flipped classroom, self-monitoring strategies, writing performance

1. Introduction

The use of technology in TESOL has transformed the language teaching profession. Technology-facilitated instruction has ruled the architecture of EFL classrooms for the past decade (Ahmadi, 2018; Cheng et al., 2019). New models prioritize flexibility, student-centered instruction, and active construction over passive reception. Of these, flipped classrooms and AI instruction have been two of the most innovative concepts. Although both strategies aim to increase learning time, learner control, and the development of higher-order skills, there is a notable pedagogical difference between them.

The flipped classroom model is widely sought after since it can extend exposure time and provide opportunities for participation. Students are typically pre-exposed to study materials through video courses or web lessons, and problem-solving, collaboration, and communicative exercises are conducted in the classroom (Afzali & Izadpanah, 2021; González-Gómez et al., 2016). Empirical studies have revealed advantages, including increased motivation, enhanced learner autonomy, and improved knowledge retention (Zainuddin, 2018; Zainuddin & Perera, 2019). For EFL, flipped classrooms can offer more practice, interactive cooperation, and instructor-managed learner autonomy (Mehring, 2017). Such advantages are not necessarily so. Some researchers have identified little effect, essentially where learners were not highly digitally literate or internally motivated (Huang & Hong, 2016).

Meanwhile, artificial intelligence (AI) has become a helpful approach in TESOL. Intelligent tutors, adaptive learning, and natural language processing-based chatbots are AI technologies that can offer instant, customized feedback (Dekhakhena,2025; Hooshyar et al., 2019). They can help students practice grammar, vocabulary, and pronunciation without increasing teachers' workload (Yilmaz & Yilmaz, 2023). AI-based interventions have also been identified as enhancing writing accuracy, building confidence, and promoting learner autonomy through greater control over practice for learners (Gayed et al., 2022; Zandi & Khezrlou, 2023). Large language models such as ChatGPT have recently been employed to support self-regulated learning through real-time adaptive scaffolding (Hsu et al., 2023; Wang & Liang, 2023).

Despite these advancements, most research has consequently addressed flipped classrooms and AI teaching separately. Flipped classroom research only targets motivation and collaboration (Zainuddin, 2018; Afzali & Izadpanah, 2021), while AI research is concerned with personalized feedback and adaptive practice (Dekhakhena,2025; Hooshyar et al., 2019). Few studies directly contrasted the two methods, and these were primarily descriptive in style (e.g., Luo & Wang, 2023; Zandi & Khezrlou, 2023). There is thus not sufficient empirical evidence to draw any conclusion as to which model best leads directly to metacognitive growth in EFL settings.

The significance of this comparison lies in its focus on the crucial role of self-monitoring in language acquisition. Self-monitoring refers to the ability of learners to monitor and control their own performance as they execute tasks (Zimmerman, 2000). Self-monitoring is one of the most crucial metacognitive capabilities, as it entails both awareness and regulation. Students who self-monitor their vocabulary, grammar, and pronunciation can identify gaps, adjust their strategies, and become proficient in these skills over time (Rahimi & Yadollahi, 2017; Abd Elkader, 2023). Previous studies have shown strong correlations between self-monitoring and accuracy, fluency, and performance in EFL settings (Zandi & Khezrlou, 2023). It is still little understood which models of technology-facilitated instruction—e.g., AI environments and flipped classrooms—influence the acquisition of this strategy.

2. Literature Review

2.1. Flipped Classroom in EFL

The flipped classroom methodology reversed the traditional learning sequence by shifting lower-order cognitive work, such as lectures, beyond the classroom, thereby leaving class time for interactive learning (Lee et al., 2017). The model was claimed to enhance motivation, autonomy, and active participation in EFL contexts. For instance, Afzali and Izadpanah (2021) noted that Iranian high school learners in flipped classrooms exhibited significant improvement in speaking fluency and confidence. On a similar note, Alizadeh and Fard Kashani (2020) noted favorable effects on learners' writing skills, with declines in language anxiety. Other studies indicated the model improved engagement and metacognition, especially when gamification was implemented (Zainuddin, 2018). There were also mixed findings, however. Some research has found that students lacking digital literacy or teachers not providing well-designed pre-class preparation materials have limited advantages (Cheng et al., 2019). Methodological flaws persisted, with most research employing short-term intervention with small samples, thus lacking generalizability (Durmuşoğlu & Demirok, 2025). Furthermore, the majority of flipped classroom studies did not investigate how these interventions directly influenced student self-regulation processes, and the mechanisms underlying these changes were underresearched. This limits the explanatory value of findings, even though performance gains have been indicated.

2.2. AI-Enhanced instruction

AI-infused instruction involved the application of innovative tools such as chatbots, adaptive environments, and feedback automation tools in language learning (Lee & Warschauer, 2020). The proof was that AI facilitated learner autonomy and provided high-quality, personalized feedback. For example, Zandi and Khezrlou (2023) showed that AI-supported feedback significantly enhanced Iranian postgraduate learners' academic writing. Equally, Luo and Wang (2023) demonstrated that AI writing feedback based on personal need increased accuracy and aided self-regulation. Beyond writing, AI chatbots and tutors were claimed to reduce anxiety when speaking and enhance motivation (Yilmaz & Yilmaz, 2023). At the same time, problems emerged. Perrotta and Selwyn (2020) argued that agency for teachers and students was in danger of being lost in hyper-automated contexts. Bias, fairness, and ethical use of student data were concerns voiced in several reviews (Holmes et al., 2021; Zhai et al., 2023). Despite these concerns, the overall direction in research indicates AI as a promising force towards more adaptive, personalized, and autonomous EFL learning. However, most AI-based studies have been conducted with short interventions or convenience samples, and they rarely control for learners' baseline strategy use. This leaves it unclear whether the gains can be attributed to AI feedback itself or to novelty effects and additional practice opportunities.

2.3. Self-Monitoring in Language Learning

Self-monitoring refers to learners' ongoing capacity to observe, track, and regulate their performance while engaged in a task, distinguishing it from post-task processes such as self-assessment or self-evaluation (Zimmerman, 2000; Pintrich, 2000). Self-monitoring, an application of Zimmerman's (2000) model for self-regulated learning, characterizes learners' ability to self-monitor and evaluate performance during the process of tasks. Unlike self-assessment or self-evaluation, which were typically post-task, self-monitoring was a process-oriented approach that was incorporated into momentary performance (Pintrich, 2000). For EFL research, the Motivated Strategies for Learning Questionnaire

(MSLQ) and Strategy Inventory for Language Learning (SILL) were used to measure self-monitoring behavior in qualities (Rahimi & Yadollahi, 2017). Empirical evidence affirmed its utility in remembering vocabulary, writing accuracy, and independent study (Rezaei & Davoudi, 2016). However, Iranian studies placed more emphasis on the reality that learners found it difficult to maintain consistent self-monitoring in the absence of explicit scaffolding (Zarei & Aghaei, 2019). Theoretically, self-monitoring is situated in Zimmerman's (2000) three-stage SRL model, where learners expect strategies in the forethought stage, self-regulate actions in the performance stage, and analyze outcomes in the reflection stage. However, the majority of EFL studies measured self-monitoring as a straightforward questionnaire measure, lacking any relation to these stages, and limiting theoretical integration.

2.4. Gap and Reason

Although flipped classrooms and AI-based instruction have been separately investigated in relation to self-regulated learning, comparative research is scarce that directly compares their effects on EFL self-monitoring. This absence of comparison was a critical failure, especially in Iran, since both approaches had already been utilized concurrently but not comparatively in depth (Afzali & Izadpanah, 2021; Zandi & Khezrlou, 2023). Since learner autonomy and sustainable academic achievement are products of self-regulation, the question then lies in what type of instructional model best facilitates it. Furthermore, Iran's situational and cultural features, including exam-based curricula and uneven access to technology, make it a good case comparison.

2.5. Empirical Studies Comparing Flipped and AI Approaches

Although direct comparisons between flipped and AI-based learning in EFL are scarce, some empirical studies are reporting preliminary findings. Afzali and Izadpanah (2021) reported the benefits of flipped classrooms for Iranian students' speaking confidence, whereas Zandi and Khezrlou (2023) reported significant writing improvement through AI-based feedback. Luo and Wang (2023) also demonstrated that individualized AI writing assistance promoted accuracy and autonomy, and Alizadeh and Fard Kashani (2020) reported decreased writing anxiety in flipped classrooms. Although these are encouraging findings, these investigations differ considerably in sample size, length, and emphasis, and none examined self-monitoring as a fundamental SRL component. Thus, the comparative evidence base remains patchy, underscoring the need for systematic research that maps instructional innovations onto theoretical models of self-regulated learning. Collectively, this highlights a critical gap where flipped and AI-enhanced approaches need to be compared within a unified design that explicitly evaluates their impact on self-monitoring strategies.

3. Research Question

This study, therefore, sought to bridge this gap by employing MSLQ-based measures in comparing flipped classroom and AI-aided instruction. According to these gaps, the questions of this study are:

RQ1: To what extent did the two instructional approaches differentially affect EFL learners' self-monitoring strategies when controlling for pre-test scores, as measured through an ANCOVA design?

RQ2: To what extent did the two instructional approaches differentially affect EFL learners' language performance when controlling for pre-test scores, as measured through an ANCOVA design?

4. Methodology

4.1. Research Design

The researcher used a three-arm, quasi-experimental pretest-posttest control-group intact EFL class design. The unit of allocation was intact EFL classes, which were randomly assigned at the class level to AI-supported, flipped, or control conditions, generally consistent with quasi-experimental methodology when random allocation to units does not occur at the individual level (Dörnyei, 2010). Baseline equivalence was established using a common pretest; posttests were then used to measure the impact of the intervention while maintaining a level of ecological validity. Primary ANCOVAs were supplemented with sensitivity analyses using linear mixed-effects models with class as a random intercept (and cluster-robust standard errors, SEs), which produced the same pattern of results to mitigate the assumption of independent observations.

The design allowed for the measurement of group differences in changes in self-monitoring strategies without compromising validity due to the application of the same pre-test and post-test measures. The pre-test established baseline equivalence among groups, and the post-test assessed the impact of interventions. Such a design had been widely recommended in second language acquisition research, where random assignment was often not feasible due to institutional limitations (Klassen et al., 2012; Dörnyei, 2010).

This approach allowed cautious causal inferences regarding the relative effectiveness of the two models while preserving ecological validity in authentic EFL classroom contexts. This research protocol was reviewed and approved by the Ethics Committee of Hormozgan University of Medical Sciences, Bandar Abbas. Informed written consent was collected from all participants prior to recruitment. Data were anonymized prior to analysis, and no individually identifiable data were stored in AI tools.

4.2. Participants

Upper-intermediate Iranian EFL learners were used as participants in this study. A sampling frame included the registry of upper-intermediate adult learners attending evening courses at a private language institute. Those qualified adult learners were included if they gave consent, and intact classes of learners were assigned to conditions as mentioned above. One hundred twenty learners were selected, and the sample size was determined using Cochran's formula to provide adequate statistical power and representativeness. It was large enough to identify medium effect sizes in quasi-experimental comparisons while at the same time giving room for attrition.

The male subjects in the study represented 56% of the sample (n = 67) while the female subjects represented the remaining 44% (n = 53). All subjects were students who were ages 18 to 30 (M = 27). This age range (18-30; M = 27) represents adult/young-adult learners, which is normal for private EFL programs and institutes because courses commonly hold on weekends and after-work hours and typically attract the older learner demographic. Again, this demographic aligns with our sampling frame and course scheduling. The institute primarily delivers services to post-secondary students and working adults enrolled in evening and/or intensive programs, making the older-than-school-age demographic typical, in line with private sector EFL delivery. The institute schedules evening and weekend courses, intentionally targeting this demographic segment. The demographic in this study is consistent with the institute page staff reviews used as the sampling frame. For homogeneity in language skills, all the participants completed the Oxford Quick Placement Test (OQPT). Participants whose levels fell in the upper-intermediate band only were taken into account, and learners outside this range were kept outside of the final data.

Volunteers were then randomly assigned to two experimental condition groups: the AI-instruction group and the flipped classroom group. A third control group under traditional instruction was also considered as a measure to promote validity. The assignment was crossed over groups to maintain comparability with respect to gender, age, and ability, thereby rendering the sample statistically adequate and representative of the upper-intermediate EFL population in the given context. Descriptive inspections concerning the pretest comparability of groups were performed, showing no meaningful imbalances as regards ages or gender. The demographics were thus not considered covariates in the main ANCOVA, which speculated on the pretest as the a priori covariate instead, given the theoretical focus and sample size planning.

4.3. Instruments

4.3.1. Oxford Quick Placement Test (OQPT)

The Oxford Quick Placement Test was used to measure the participants' proficiency level and to homogenize the groups. The OQPT was a norm-referenced test widely used in EFL research to categorize learners into bands representing CEFR levels. Those participants scoring only in the upper-intermediate band (B2) were considered in this study. The test was administered in paper form and localized to the context by highlighting the sections of grammar, vocabulary, and reading comprehension most relevant to academic English. Accordingly, the OQPT was considered reliable and valid for homogenizing participants' proficiency levels in this study.

4.3.2. Motivated Strategies for Learning Questionnaire (MSLQ)

To measure learners' self-monitoring, the Metacognitive Self-Regulation subscale of the Motivated Strategies for Learning Questionnaire (MSLQ) was used. The original MSLQ was developed by Pintrich (1991), comprising 81 items across various subscales. However, only 12 of the metacognitive self-regulation-related items were used in this study. Learners were asked to report planning, monitoring, and regulation of learning activities when working on tasks. Slightly earlier forms of the MSLQ employed in Iranian studies of EFL resulted in satisfactory psychometric properties, with Cronbach's alpha coefficients ranging from 0.78 to 0.87 for the metacognitive subscale (Rahimi & Yadollahi, 2017). Some slight wording adjustments were made to adapt items to language-learning contexts (e.g., replacing "course material" with "English tasks"). Responses to items could be rated from a five-point rating scale from "not at all true of me" to "very true of me." Above all, the MSLQ posed a valid and context-sensitive measure of self-monitoring strategies. Items were subjected to an expert review process (n = 3) and back-translation and were adjudged linguistically appropriate to EFL contexts. Further, a small-scale pilot (n = 15) admitted that 5-point scale items were clear and functioned well in eliciting responses. To provide evidence of validity beyond internal consistency, an exploratory factor analysis (EFA) was conducted to assess the 12 items at pretest (KMO = 0.82; Bartlett's $\chi^2(66)$ = 0.80, $p < 0.001$). All items loaded > 0.40 on a single factor focused on monitoring, which explains 48-55% of the variance, indicating unidimensional scoring in this context.

4.3.3. Language Performance Rubric

In addition to self-report data, scores on a writing performance rubric also acted as the criterion measure. The rubric consisted of three dimensions: accuracy, fluency, and coherence. Accuracy was operationalized as the correct application of grammar, vocabulary, and mechanics; fluency was associated with ease and naturalness of expression and flow of writing; and coherence was viewed as logical organization and cohesion, which involved clarity of ideas being the ultimate goal. Each dimension was rated on a five-point analytic scale adapted from Jacobs, Zinkgraf, Wormuth, Hartfiel, and Hughey's (1981) ESL Composition Profile. The writing tasks were scored by two trained raters blinded to the group assignment at pre-tests and post-tests. Prior to the writing tasks being scored, the raters completed a two-hour calibration, which involved using anchor scripts from pre-tasks, with a full score range calibration; differences between raters were resolved to the criterion ($\Delta \leq 1$ point per trait). All group labels were masked in the writing scripts, and a trained research assistant used a random selection of IDs to code the files. The inter-rater reliability was excellent (ICC[2,k] = 0.84, 95% CI [0.78, 0.89]). Thus, the rubric has been used in multiple studies relating to EFL in Iran and has shown evidence of both content and construct-related validity regarding writing rubrics. Thus, it was considered reasonable to employ the rubric for examining the influence of the treatment on the changes in the learners' writing performance over the course of the study.

4.4. Procedures

The study of eight weeks' duration was conducted in the systematic sequence of pre-testing, intervention, and post-testing. Initially, all participants completed the Oxford Quick Placement Test (OQPT) to determine their upper-intermediate learner status. The Motivated Strategies for Learning Questionnaire (MSLQ – Metacognitive Self-Regulation subscale) was then conducted to obtain baseline measures of learners' self-monitoring strategies. The last steps included completing a writing task that was recreated and coded for accuracy, fluency, and coherence to enhance research reflexivity. These provided the pre-test scores for both dependent variables.

The first experimental group received instruction through the flipped classroom. Before class, students watched pre-recorded video lectures on writing strategies, grammatical items, and organizational structures. Each separate lecture was short and was uploaded to the institute's LMS as a pre-class microlecture. There were 3–5-question readiness checks at the front of each class to check viewing. In class, students engaged in interactive activities that were in a fixed sequence (peer review → guided rewriting practices → specific error analysis). During class, students participated in interactive activities collectively (peer review, guided rewriting practice, and error analysis practice) while the teacher was available. Instruction focused on both writing mechanics and the metacognitive process of monitoring one's own output. Students were also asked to keep short reflective diaries to help them become more conscious of their performance. Micro-lectures (8–12 minutes) were provided on the institute's LMS (Moodle); the micro-lectures covered ideas for organizing an essay, what specific cohesion devices are, and focused on high-frequency grammar issues—each week included two videos for a mean total pre-class viewing time of ≈20–24 minutes.

The second experimental group received instruction delivered in an AI-supported classroom. As students developed their drafts in class, they received AI-generated feedback on grammar, vocabulary use, and coherence. The types of AI-enhanced tools were: chatbots (for interactive speaking practice), adaptive writing practice, and automated writing feedback. The AI environment consisted of a large language model, a textual assistant, and an automated writing feedback module, all of which were available via a web interface. System usage was logged (i.e., number of requests for feedback; number of revision cycles) to verify fidelity. The teacher's role was defined in a protocol prior to implementation, to ensure consistency in facilitation from the first session to the last. In the AI-supported classroom, AI provided immediate feedback, suggestions, potential grading, and revision suggestions so that students could revise their text re-iteratively. The teacher's role was primarily that of a facilitator, and the students needed to consider whether they had critically evaluated and utilized AI feedback responsibly. The emphasis was placed on utilizing AI assistance in developing self-monitoring capabilities, specifically in identifying and correcting errors in writing. The AI environment consisted of a writing assistant based on large language models, accessible via a web interface with security features. It utilized standardized prompts to provide feedback on grammar, vocabulary, and coherence. Students had a maximum of three revision cycles per task; usage logs recorded attempts at feedback and the number of revisions (median cycles).

Both interventions lasted eight weeks, and the students attended two 90-minute classes per week. Pre-class and homework assignments in the flipped classroom, as well as AI-based practice in the AI-enhanced group, were maintained at a consistent level in terms of content coverage and writing assignments to ensure comparability.

At the end of the intervention, all students again completed the MSLQ self-monitoring subscale and a post-test writing task. Writing was evaluated using the assessments from the pre-test based on the same rubric. The self-report and performance-based assessments enabled data triangulation, providing a more comprehensive examination of instructional impact. Fidelity checks included (a) LMS analytics confirming $\geq$ 80% completion of pre-class videos (non-compliance prompted in-class mini-briefings), (b) archived readiness checks ($\geq$ 80% correct threshold), (c) AI usage logs (median= number of interactions per task), and (d) biweekly observations with a 10-item adherence checklist (mean adherence $\geq$ 90%).

4.5. Data Analysis

The analyses proceeded in four phases. (1) Screening: the data were screened for missing data evidence, outliers ($|z| > 3.29$), normality (Shapiro–Wilk), and homogeneity of variances (Levene). Regression slopes homogeneity was tested using the Group×Pretest interaction; the interaction was not significant in all models. (2) Reliability: Cronbach's α was calculated for pre-test and post-test MSLQ subscales, and writing outcome inter-rater reliability was estimated with the two-way random effects intraclass correlation, ICC(2,k). In particular, (3) Primary model (self-monitor): post-test self-monitoring outcome scores were contrasted using one-way ANCOVA with the fixed factors of (AI, flipped, and control) and as the covariates with pre-test self-monitoring scores, with reliance on the homogeneity of regression slopes and linearity assumptions. Adjusted means were established after a pairwise comparison, using Bonferroni correction to uphold the alpha level progressively. With the two major outcomes (writing, self-regulation), familywise error was controlled using the Holm–Bonferroni procedure. Effects were reported as partial η^2 with 95% CI. (4) Outcome: writing: the same ANCOVA was conducted on post-test measures of writing with covariate pre-test writing. The alpha level was 0.05 (two-tailed tests). All statistical analyses were performed within IBM SPSS Statistics 25. Sensitivity analyses via linear mixed-effects models with class as a random intercept (cluster-robust SEs) also replicated the ANCOVA trend. To check sensitivity, ANCOVAs with age and gender added as additional covariates were conducted; significance and effect size pattern were not modified.

5. Results

The results section begins with the assessment of the reliability of the measures used in the research. Since the primary dependent Variable was measured with the help of the Metacognitive Self-Regulation subscale of the MSLQ, it was necessary to examine the internal consistency of the scale for both pre-test administrations and post-test administrations. Reliability was determined using Cronbach's alpha, which has been widely used in education and applied linguistics research to determine if items within a scale all measured the same construct uniformly. Having obtained sufficient reliability at both testing occasions ensured that subsequent statistical comparisons of self-monitoring scores would be valid and meaningful. After the presentation of these coefficients for reliability, descriptive statistics of self-monitoring scores by time and group are provided. Then, inferential analyses, such as ANCOVA, were used to examine the main and interaction effects. The internal consistency of the MSLQ Metacognitive Self-Regulation subscale was evaluated on the pre-test administration.

The 12-item self-monitoring subscale showed an adequate level of reliability at Time 1 ($\alpha = 0.74$). The reliability of the self-monitoring subscale was much higher at Time 2 ($\alpha = 0.90$). The increase in reliability suggests effects from the intervention, as the items were probably more congruent after the participants employed self-monitoring strategies in the intervention (George & Mallery, 2003).

Table 1

Reliability Statistics

	Cronbach's Alpha	Cronbach's Alpha Based on Standardized Items	N of Items
Self-Monitoring Questionnaire (Pre-test)	0.740	0.737	12
Self-Monitoring Questionnaire (Post-test)	0.899	0.899	12

Following the intervention, the internal consistency of the MSLQ Metacognitive Self-Regulation subscale was also verified for post-test administration. The corrected item–total correlations ranged from 0.522 (Item 11) to 0.711 (Item 9). All were greater than the 0.30 threshold (Pallant, 2020), which suggested that all items were contributing significantly towards the composite self-monitoring construct. Further, Cronbach's Alpha if Item Deleted values ranged between 0.886 and 0.896 to confirm that removal of any item would not significantly improve the general reliability. The implication was that the 12-item scale was highly cohesive at the post-test stage. A post-test means total score of 46.14 (*SD* = 8.38), which was greater than the pre-test means of 35.77. This discrepancy was due to an overall increase in self-monitoring strategies as reported by learners after they were exposed to the interventions. The variance (70.27) was also greater than that of the pre-test, demonstrating wider scatter in the post-test scores and suggesting heterogeneity in the manner in which learners had benefited from the instructional models.

Table 1 presented that Cronbach's alpha coefficient was 0.899, and so was the standardized items' coefficient. This was a significant improvement from the pre-test alpha (0.74), which indicated that there was a very high internal consistency for the post-test. Authors George and Mallery (2003) assert that, on an ideal basis, coefficients of 0.80 and above would be good, and those of 0.90 and above would be excellent, which indicates that the post-test instrument was very reliable to measure self-monitoring. Taken in aggregate, the findings confirmed that the questionnaire functioned even better at the post-test stage, not only confirming the instrument's strength but also showing students' more stable reactions after having participated in the flipped and AI-supported teaching environments.

The descriptive statistics of self-monitoring scores on the pre-test and post-test between groups are given in Table 2. During the pre-test, scores were relatively similar across groups. The control group's mean score was 3.19 (*SD* = 1.07), the flipped classroom group's mean score (referred to as "Self-monitoring" in the output) was 3.11 (*SD* = 1.01), and the AI-augmented group was slightly lower with a mean of 2.81 (*SD* = 1.09). The results indicated that, at the start of the study, all three groups had an equivalent level of self-monitoring so that they were at the same point in terms of baseline equivalence.

Differences emerged at the post-test. The control group reduced slightly to a mean of 2.83 (*SD* = 0.55), reflecting minimal change over time. The flipped classroom group rose to a mean of 4.03 (*SD* = 0.62). In contrast, the AI group recorded the highest mean score of 4.60 (*SD* = 0.38). The AI condition's low standard deviation indicated that both the students in this condition indicated similarly higher post-intervention self-monitoring behaviors.

The descriptive statistics overall results thus indicated that the two experimental conditions helped participants achieve significant growth, especially the AI-instruction-enhanced condition. However, the control condition experienced little to no growth. This means that the following inferential analyses have meaningful trends. Table 2 contains descriptive statistics for self-monitoring from item means (scale 1–5).

Table 2

Descriptive Statistics of Self-Monitoring Scores by Group and Time (Pre-Test and Post-Test)

		Range	Minimum	Maximum	Mean		Std. Deviation	Variance
	Group	Statistic	Statistic	Statistic	Statistic	Std. Error	Statistic	Statistic
Pre-Test	Control	4.00	1.00	5.00	3.187	0.168	1.066	1.137
	Flipped	3.50	1.50	5.00	3.112	0.159	1.009	1.019
	AI	4.00	1.00	5.00	2.812	0.173	1.095	1.201
Post-Test	Control	2.00	2.00	4.00	2.825	0.086	.549	0.302
	Flipped	2.00	3.00	5.00	4.025	0.097	0.619	0.384
	AI	1.50	3.50	5.00	4.600	0.059	0.378	0.144
Valid N (listwise)	120							

Table 3 details the descriptive statistics for writing performance based on the analytic rubric (accuracy, fluency, and coherence combined). The control group had a pre-test mean of 9.45 (*SD* = 2.59), and the pre-test mean in the flipped classroom was 8.80 (*SD* = 2.64). The AI-augmented group's pre-test mean was lower still at 8.28 (*SD* = 2.23). Ranges for the control (11) and flipped (9) groups were similar (8) but reflected moderate within-group variability in students' baseline writing skill levels. Overall, the results here suggest students entered the study with relatively similar writing skills, but the control group appeared to be stronger on average.

At the post-test, there were noticeable patterns between groups. The control group showed minimal improvement, with a mean of 8.70 (*SD* = 1.38), which was slightly lower than its pre-test score. The flipped classroom group, on the other hand, significantly improved to a mean of 11.80 (*SD* = 1.32) in writing performance. The AI-supported group improved most with a post-test mean of 13.55 (*SD* = 1.08). The lower standard deviations among the two experimental groups indicated that students' scores became more consistent after instruction, which means stable and positive effects of the interventions.

On the whole, the descriptive statistics suggested that the control group was fairly constant or slightly declined. By contrast, the two experimental groups showed significant improvements, particularly the group that received AI-powered instruction, which demonstrated the greatest improvements in writing performance.

Table 3

Descriptive Statistics of Writing Scores by Group and Time (Pre-test and Post-test)

	Group	Range	Minimum	Maximum	Mean	Std. Deviation	Variance	N
Pre-writing	Control	11.00	4.00	15.00	9.450	2.591	6.715	40
	Flipped	9.00	4.00	13.00	8.800	2.642	6.985	40
	AI	8.00	4.00	12.00	8.275	2.230	4.974	40
Post-writing	Control	6.00	6.00	12.00	8.700	1.381	1.908	40
	Flipped	5.00	9.00	14.00	11.800	1.324	1.754	40
	AI	4.00	11.00	15.00	13.550	1.084	1.177	40
Valid N	120							120

Table 4 summarizes the descriptive statistics of the post-test self-monitoring ratings among the three groups. For transparency, Table 4 summarizes the same outcome using total scores (12–60). Both metrics are monotonically related and confirm the same group pattern. There were apparent differences among the groups. The control group had the lowest mean score ($M = 35.33$, $SD = 2.63$), showing almost no improvement compared to their pre-test level. The flipped classroom class showed significant improvement, with a mean of 48.63 ($SD = 2.55$), indicating enhanced utilization of metacognitive self-regulation strategies following the intervention. The AI-taught group exhibited the highest frequency of self-monitoring, with a mean of 54.48 ($SD = 1.87$), and possessed the lowest standard deviation, suggesting that students in this class not only performed better but also had more equalized performance.

The mean for the entire sample was 46.14 ($SD = 8.38$), which revealed the significant contribution of the experimental groups to the overall mean of the post-test. These descriptive results revealed a powerful gradient of effect, with AI-supported instruction producing the most gains, followed by the flipped classroom, and the control group at significantly lower levels.

Table 4

Descriptive Statistics of Self-Monitoring (Post-test)

Group	Mean	Std. Deviation	N
Control	35.325	2.625	40
Flipped	48.625	2.548	40
AI	54.475	1.867	40
Total	46.141	8.382	120

Levene's test was conducted to determine the error variances of group equality for post-test self-monitoring scores. As illustrated in Table 5, the test was found to be non-significant when calculated based on the mean, $F(2,117) = 2.54$, $p = 0.083$. Other statistics based on the median ($p = 0.152$), median with adjusted degrees of freedom ($p = 0.153$), and trimmed mean ($p = 0.092$) were also not significant.

Since all p-values were greater than 0.05, there was an assumption of homogeneity of variances. This finding indicated that variation within the post-test self-monitoring scores for the control, flipped class, and AI-enriched groups was equal, thus justifying the use of ANCOVA.

Table 5

Levene's Test of Equality of Error Variances for Post-test Self-Monitoring Scores

		Levene Statistic	df1	df2	Sig.
Post-SM	Based on Mean	2.541	2	117	0.083
	Based on Median	1.914	2	117	0.152
	Based on the median and with adjusted df	1.914	2	107.562	0.153
	Based on the trimmed mean	2.433	2	117	0.092

An ANCOVA was conducted to establish whether post-test self-monitoring scores varied across groups after controlling for pre-test scores (Pre_ SM). Results are shown in Table 6. The analysis showed that there was a significant covariate effect of Pre_ SM, $F(1,116) = 4.28$, $p = 0.041$, partial $\eta^2 = 0.036$, such that learners' pre-test self-monitoring played a relatively small but significant role in determining performance on the post-test.

Most importantly, there was an exceedingly significant main effect of group, $F(2,116) = 701.65$, $p < 0.001$, partial $\eta^2 = 0.924$. The considerable effect size revealed that instruction type had a strong influence on learners' self-monitoring performance, after controlling for pre-test differences. The effect accounted for more than 92% of the variance of the dependent Variable, which is an exceedingly strong instructional effect.

The total adjusted model was also significant, $F(3,116) = 470.77$, $p < 0.001$, with an R^2 of 0.924 (adjusted $R^2 = 0.922$), indicating that the model explained a significant proportion of the variance in post-test self-monitoring scores. ANCOVA results combined confirmed that the three groups differed significantly on post-test self-monitoring and that the instructional condition was the primary factor explaining the difference. The pre-test covariate was significant, suggesting that baseline self-monitoring predicted part of the variance in post-test scores. This indicates some continuity of individual differences in self-monitoring over time, even after instruction.

Figure 1 shows the marginal means of post-test self-monitoring scores for the three groups with pre-test self-monitoring scores statistically controlled as a covariate. The figure shows an apparent linear increase in adjusted post-test means from the control group ($M \approx 35$) to the flipped classroom group ($M \approx 49$), and at a peak in the AI-supported instruction group ($M \approx 55$). Estimated marginal means of post-test self-monitoring scores by group (covariate: pre-test self-monitoring). Error bars indicate 95%

confidence intervals. Bonferroni-adjusted pairwise comparisons showed significant differences among all three groups (AI > Flipped > Control, all ps < 0.001).

Table 6

ANCOVA Results for Self-Monitoring (Post-test as Dependent Variable, Pre-test as Covariate)

Source	Type III Sum of Squares	df	Mean Square	F	Sig.	Partial Eta Squared
Corrected Model	7727.864a	3	2575.955	470.770	0.000	0.924
Intercept	6250.432	1	6250.432	1142.302	0.000	0.908
Pre-SM	23.398	1	23.398	4.276	0.041	0.036
Group	7678.532	2	3839.266	701.648	0.000	0.924
Error	634.727	116	5.472			
Total	263849.000	120				
Corrected Total	8362.592	119				

a. R Squared = 0.924 (Adjusted R Squared = 0.922)

This trend indicates that while students in the control condition reported slight gains in self-monitoring, the flipped classroom treatment yielded a significant improvement. The AI-enhanced instruction had the most considerable effect, and students performed better than their peers in both the control and flipped classroom groups. The sharp slope between the control and experimental groups demonstrates the efficacy of the instructional interventions, and most importantly, the noticeable advantage of AI integration in building metacognitive self-regulation.

The trend on the positive side across groups further indicates that instructional design played a significant role in shaping learners' post-test self-monitoring scores, in line with the ANCOVA result of the strongly significant main effect of group membership on self-monitoring when pre-test performance was controlled.

Figure 1

Estimated Marginal Means of Post-SM

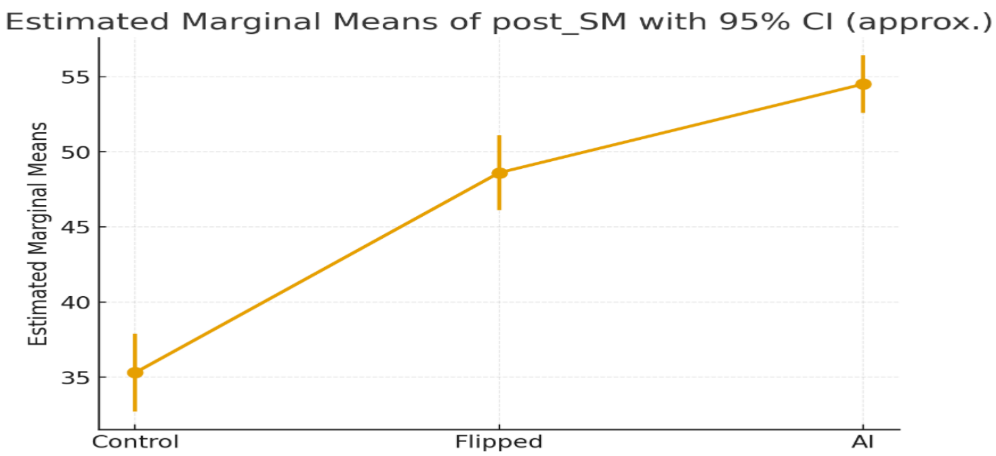


Table 7 indicates descriptive statistics of post-test level writing performance for the three groups. The control group had the lowest mean (*M* = 8.70, *SD* = 1.38), which had scarcely any gain over the pre-test level. The flipped classroom group showcased concrete evidence of a considerable gain with a mean of 11.80 (*SD* = 1.32). This score indicated a whole lot better in terms of accuracy, fluency, and coherence. The AI-supported group scored better than both groups with a mean score of *M* = 13.55, *SD* = 1.08. Since all participants showed improvement, with a slight standard deviation, it indicated that the improvement was uniform among all participants.

The average across all participants was 11.35 (*SD* = 2.38). This provided a clear ordering of outcomes at the summary level, showing that AI-augmented instruction resulted in greater improvement, the flipped classroom resulted in moderate improvement, and the improvement in the control group was significantly lower.

Table 7

Descriptive Statistics of Writing Scores by Group (Post-test)

Group	Mean	Std. Deviation	N
Control	8.700	1.381	40
Flipped	11.800	1.324	40
AI	13.550	1.084	40
Total	11.350	2.375	120

Levene's test was conducted to determine whether the error variances of the post-test writing scores of the three instruction groups were equal. As shown in Table 8, the test was not significant by the mean, $F(2,117) = 1.48$, $p = 0.232$. The alternative statistics, based on the median ($p = 0.233$), the median with adjusted degrees of freedom ($p = 0.233$), and the trimmed mean ($p = 0.220$), were also not significant.

Since all the significance values were greater than the 0.05 level, the homogeneity of variances was satisfied. This ensured that post-test writing scores variability was the same in the control, flipped classroom, and AI-supported groups, confirming the ANCOVA results.

Table 8

Levene's Test of Equality of Error Variances for Post-test Writing Scores

		Levene Statistic	df1	df2	Sig.
Post_Writing	Based on Mean	1.479	2	117	0.232
	Based on Median	1.474	2	117	0.233
	Based on the median and with adjusted df	1.474	2	116.735	0.233
	Based on the trimmed mean	1.532	2	117	0.220

ANCOVA was utilized to examine post-test writing scores across groups, controlling for pre-test writing ability. Results are contained in Table 9. The pre-test writing that was used as the covariate was not significant, $F(1,116) = 0.028$, $p = 0.868$, partial $\eta^2 = 0.000$, which showed that pre-test writing scores did not affect post-test writing performance. So, the differences on the pre-test level in writing could not account for differences on the post-test level in writing.

Conversely, the main effect of group was powerful, $F(2,116) = 143.52$, $p < 0.001$, partial $\eta^2 = 0.712$. This considerable effect size indicated that the instructional conditions accounted for approximately 71% of the variance in the post-test writings. As a whole, the corrected model was significant, $F(3,116) = 98.92$, $p < 0.001$, with an R^2 of 0.719 (adjusted $R^2 = 0.712$), showing that the model accounted for a very high percentage of variance. Taken as a whole, these findings confirmed that instruction type (control, flipped classroom, AI-aided) significantly affected students' post-test writing ability, independent of pre-test ability. The covariate for pre-test performance did not reach significance for writing, indicating that baseline writing ability did not influence post-test performance. This suggests that the interventions had effects that were independent of participants' original writing skills. Bonferroni-adjusted pairwise comparisons showed significant differences between groups (AI > Flipped > Control, ps < 0.001).

Table 9

ANCOVA Results for Writing Performance (Post-test as Dependent Variable, Pre-test as Covariate)

Source	Type III Sum of Squares	df	Mean Square	F	Sig.	Partial Eta Squared
Corrected Model	482.645a	3	160.882	98.923	0.000	0.719
Intercept	1099.991	1	1099.991	676.363	0.000	0.854
Pre-Writing	.045	1	0.045	0.028	0.868	0.000
Group	466.829	2	233.414	143.522	0.000	0.712
Error	188.655	116	1.626			
Total	16130.000	120				
Corrected Total	671.300	119				

a. R Squared = 0.719 (Adjusted R Squared = 0.712)

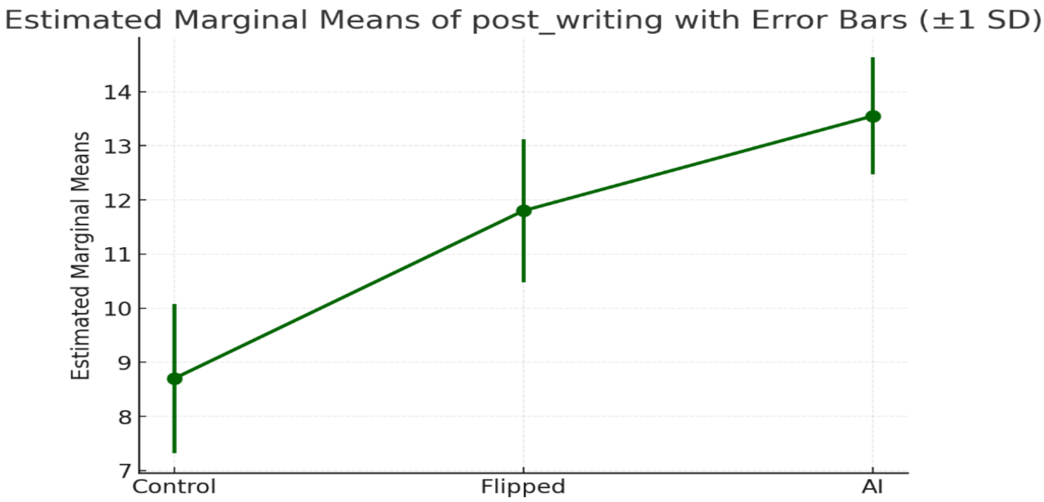
Figure 2 presents the estimated marginal means of post-test writing achievement scores for the three instruction groups with the pre-test writing score as a covariate. The trends in the graph are clear and consistent: the control group had the lowest adjusted post-test mean ($M \approx 8.7$); the flipped classroom group experienced an impressive increase in their post-test mean ($M \approx 11.8$); the AI-assisted instruction group had the highest mean score ($M \approx 13.6$). Estimated marginal means of post-test writing scores by group (covariate: pre-test writing). Error bars are 95% confidence intervals.

This pattern indicates that while the control group of students made little progress with time, both experimental interventions allowed for significant boosts in writing skill. The flipped classroom method significantly enhanced students' writing skills. However, instruction supported by AI had the most substantial impact, not just on average performance but also on the reliability of students' results, as reflected in the minor standard deviations observed in the descriptive statistics.

Overall, the graph provides visual confirmation of the ANCOVA results, which showed instructional group as a highly significant main effect and accounted for a high proportion of variance in post-test writing quality. The steep slope from control to experimental conditions bears witness to the potential of innovative instructional methods to reverse the learning process, particularly when supplemented by AI-driven feedback and practice.

Figure 2

Estimated Marginal Means of Post-Writing



6. Discussion

The first research question compared the relative impact of flipped classroom instruction and AI-supported instruction on students' self-monitoring. The results demonstrated that both experimental conditions significantly enhanced students' metacognitive self-regulation; however, the AI-supported condition produced the most significant gains. Students who went through AI-based environments consistently generated high post-test self-monitoring scores, with reduced variability, not only improvement, but also consistency among participants. The flipped classroom class also had more self-monitoring, though less so than did the AI class.

The advantage of teaching with AI was that learners received continuous and adaptive feedback, facilitating the assessment of their approach and the optimization of learning activities. Flipped classroom, on the other hand, was not as effective, primarily due to infrastructural limitations in Iran, including low digital literacy and unstable internet connectivity, which hindered learners' access to pre-classwork and provided little scope for social in-class monitoring (Huang & Hong, 2016; Cheng et al., 2019).

Besides these infrastructural issues, a chain of pedagogical procedures might also account for the smaller impacts of the flipped model. To begin with, flipped instruction relies on students' pre-class obedience and self-regulatory willingness; low motivation or conformity deficiency can affect in-class results (Dörnyei, 2010). Second, the model offloads cognitive load onto pre-class phases: students must study new material in advance of linking it to performance. This task will not work if the materials are not appropriately scaffolded for strategy monitoring. Third, delayed feedback in flipped classrooms relies on teacher time or peer discussion, which can limit timely error detection and micro-adjustments. Fourth, teachers' scaffolding pedagogical quality mediates effectiveness—encompassing pre-class prompts, in-class monitoring cues, and guided post-task reflection. Iranian research has demonstrated that such scaffolding has a considerable impact on outcomes (Afzali & Izadpanah, 2021; Alizadeh & Fard Kashani, 2020). Lastly, in assessment-focused environments, students tend to opt for test-related material rather than reflection activities, which undermines the proposed metacognitive advantages of flipped learning (Burton, 2002).

By contrast, AI-based tools such as intelligent tutors, writing feedback software, and chatbots like GPT-3 offer real-time, personalized feedback on coherence, vocabulary, and grammar. This enabled the students to revise their work in real time and undergo successive rounds of improvement (Luo & Wang, 2023). By offering individualized cues in real-time, AI aligned more closely with the forethought–performance–reflection cycle of self-monitoring. Accordingly, whereas both methods permitted self-regulation, flipped classrooms provided conditional and indirect gains, but AI provided ongoing, focused scaffolding with more reliable gains. Nonetheless, we should also recognize that AI feedback is limited as much of it can be shallow (e.g., grammar or lexical accuracy) and not always truly represent the pragmatic depth or discourse-level coherence, and also the possibility that learners depend too much on the automated scaffolding, and do not develop their own monitoring mechanisms. Simultaneously, AI feedback is often form-focused, concentrating on grammar and lexis, and neglecting higher-level features of writing, such as content development, creativity, or pragmatic appropriateness. This may lead to an ingrained, accuracy-oriented approach to writing, where learners may not fully pay attention to discourse-level monitoring. These results align with current evidence demonstrating that AI-delivered interventions enhance self-regulation and motivation by providing adaptive scaffolding and feedback (Choi & Lee, 2024; Hazaymeh et al., 2025; Du & Daniel, 2024; Wang & Liang, 2023). For comparison purposes, flipped classroom approaches were also identified as positively affecting self-monitoring, specifically by preparing students beforehand for class and engaging them in reasoning during class activities. Iranian research has supported the idea that inverted designs increase learners' autonomy and engagement in monitoring approaches; however, the effect size varies depending on the instructional design and teacher scaffolding (Afzali & Izadpanah, 2021; Alizadeh & Fard Kashani, 2020). These findings suggest that flipped learning fosters monitoring indirectly by causing structural changes in instruction provision, whereas AI provides direct, moment-by-moment regulatory support.

Furthermore, it is vital to look not just at instructional factors, but also at the larger cultural and educational context of Iran that could have affected the effects produced by the current instructional methods. Iran's exam-oriented education, inconsistent access to technology, and varying degrees of digital literacy among students presented numerous obstacles to students' ability to fully immerse themselves in either the flipped approach or the AI-based approach. For example, as many students in Iran commonly do not have regular access to functioning and reliable internet services or digital devices, this may have limited their participation in the online elements of the flipped classroom model. Furthermore, I believe the rigorous evaluation model, evident in Iran's exam-driven culture, may have increased such differences. In flipped designs, learners rarely engaged with activities prior to class that were not tied to their scores, which detracted from a model that dealt with reflection and preparation.

In contrast, the AI-based tools provided useful, tangible corrections and also showed visible short-term gains in accuracy and coherence that corresponded more with learners' score-oriented expectations. This correspondence may explain some of the differences in the regularity of support students used between the AI and the flipped design. This limitation, along with other contextual limitations, may have made flipped classrooms less important than AI-supported models, which do not rely too heavily on internet connection requirements (Huang & Hong, 2016; Cheng et al., 2019).

The present results, therefore, add to previous research by showing that, although both interventions were effective, AI-based learning yielded more substantial and consistent improvements. This result aligned with systematic reviews demonstrating strong effects of AI compared to traditional and blended approaches on metacognitive processes and learner regulation (Wang et al., 2024; Zawacki-Richter et al., 2019). The incorporation of adaptive feedback systems and automated performance tracking appeared to be key differentiators that distinguished AI-driven interventions from flipped classrooms. While considering the effect sizes as statistically significant ($\eta^2 > 0.70$), we must be cautious when interpreting effect sizes as large as this. We must consider that the significant effect sizes may be the result of extraneous variables, such as participant homogeneity, teachers holding the same biases, and the novelty/practice effect due to the AI tools used in this study. These are analogous to concerns raised about effect sizes in previous SLA intervention studies, which suggest that inflated effect sizes may be related to participant sample size or intact-class designs (Dörnyei, 2010). Furthermore, for this study, it is likely that the relatively homogenous sample also contributed to the effect sizes. Future research should attempt to replicate the findings of this study using a more heterogeneous sample to enhance the validity of the current results.

The second research question probed the effect of the flipped classroom and AI-guided learning on students' writing performance. Findings showed that both experimental groups outperformed the control group significantly, with the AI-supported instruction having the highest scores in the post-test. The flipped classroom condition showed impressive gains in accuracy, fluency, and coherence; however, the AI condition performed considerably better, recording greater gains and more stable performance across the students.

The superiority of AI-supported pedagogy can be attributed to the adaptive, real-time scaffolding and feedback that intelligent systems can offer, allowing for the resubmission of work for revision, the detection of errors, and enhanced cohesion. This corroborated previous studies that revealed AI-based feedback tools improved the quality of learners' writing by facilitating lexical accuracy, syntactic sophistication, and overall composition (Hazaymeh et al., 2025; Luo & Wang, 2023; Wang & Liang, 2023). Research also found that AI-assisted tools facilitated personalized writing practice, leading to more sustainable improvement in writing skills compared to traditional feedback mechanisms (Du & Daniel, 2024; Choi & Lee, 2024).

Although the flipped classroom model yielded benefits for writing performance, including preparing students for class and facilitating collaboration, it was less consistently effective than AI-supported instruction. This inconsistency is largely due to the differences in feedback from each model. Flipped classrooms are designed to allow students to discuss with their peers or the teacher. However, they may result in limited capacity for providing students with real-time feedback, depending on the teacher's availability or the students' preparedness. Conversely, AI tools provided immediate and ongoing feedback, resulting in more meaningful and continuous revisions made more authentically by the students. Instead of only cycling through iterative corrections, the fundamental difference is how each model (albeit indirectly) was promoting self-monitoring. The immediacy of AI feedback compelled learners to focus on errors, adjust their approach, and reapply in the moment, which continuously reinforced the metacognitive cycle of forethought, performance, and reflection. The advantages of the flipped model included the use of autonomy and the delay of reflection; however, all benefits are indirect. Most learners had to wait for things to be validated by peers or teachers, which is a very different type of moment-to-moment feedback in their monitoring processes. This course of activities led me to understand the salient differences in feedback type and how real-time adaptive feedback is integral to student self-regulation, which is an important aspect of writing performance. Iranian studies have confirmed that flipped writing instruction increases the autonomy and self-regulated manner in which students approach drafting and revising activities, while results are often mediated by teacher scaffolding and learner motivation (Afzali & Izadpanah, 2021; Alizadeh & Fard Kashani, 2020). Whereas flipped classrooms overhauled the writing process and provided space for reflection, they did not provide the same degree of personalized instant feedback as in AI-environmented conditions, which was the explanation for the gap in magnitude of improvement.

Overall, the findings supported previous research by showing that not only did AI instruction improve writing skills more effectively than traditional methods of communication, but it also created a safe environment with fair and constructive feedback for all students. The results supported more extensive reviews, which also noted the revolutionary potential for AI in language education, especially for productive skills in which real-time correction and development are highly important (Wang et al., 2024; Zawacki-Richter et al., 2019). The range of effects observed between the flipped classroom and AI-supported learning models suggests that AI supported the improvement process through quicker, real-time feedback designed to provide personalized remediation, helping to build self-regulation and writing progress. The AI-enabled feedback allowed students to work through an iterative process, enhancing their work based on real-time, immediate feedback. The feedback cycle was more consistent and immediate than in a flipped classroom, where feedback was somewhat sporadic and primarily carried over a space and time-delayed remediation strategy. The consistency and stability of the improvements seen in the AI group illustrate the importance of immediacy for personalized feedback in language learning. Nonetheless, we must field the findings in an Iranian EFL context. Factors such as the exam-centered curriculum, contingent (rather than continual or inconsistent-indeterminate) technology infrastructure, and students' digital literacy likely influenced both the flipped and AI-supported models. For example, the lack of accommodating and dependable, stable internet access and devices may have limited the potential of the flipped model. However, it has accelerated the AI model dependent on stimulus novelty. All things considered, the nature of these contextual factors and the generalizable nature of our results should be considered for parallels in similar socio-educational contexts.

7. Conclusion

This study compared the effects of AI-supported and flipped classroom instruction on Iranian EFL learners' self-monitoring and writing performance. The results substantiated that both new approaches improved the performance of learners compared to the control group. However, AI-facilitated instruction produced the most significant gain, both in writing ability and metacognitive self-regulation, with effects being higher and more consistent across participants.

To the best of our knowledge, this study is among the first attempts to empirically compare flipped and AI-supported instruction on self-monitoring in the Iranian EFL context. While flipped classrooms have been thoroughly investigated in terms of language learning strategies and writing development, this current study offers fresh evidence that AI-based instruction possesses an even greater advantage in cultivating self-regulatory behaviors.

Even as higher achievements were reported for AI-augmented learning, the findings must be considered in moderation. Flipped and AI-augmented models both emerged as effective in enhancing learners' self-monitoring and writing skills, which means that the two models can enhance traditional pedagogy in important ways. However, the adaptive feedback and customized scaffolding engineered in AI tools position them as individually potent in optimizing learner regulation and performance, thereby underscoring their capacity to transform future EFL instruction.

The patterns of learner gains demonstrated in this study suggest that, even though AI feedback was relatively more effective overall, the collaborative activity encouraged in flipped spaces also provided benefits, and suggest that there may be value in trying to combine these approaches. The findings had important pedagogical implications for practice. For EFL teachers, the findings suggest that AI technologies cannot replace or complement flipped practices. Flipped learning provided a framework that encouraged students to do pre-class work with materials and apply practice in cooperative contexts. AI technologies provided adaptive, personalized feedback to inform students' monitoring processes. Blending the two methods would enable instructors to craft hybrid classrooms where pre-class introduction and in-class peer collaboration were augmented with ongoing, individualized scaffolding. Blending the two methods could maximize the impact of instruction while preserving a people-oriented classroom environment.

For students, both methods emphasized the value of autonomy and agency in the learning process. Flipped designs fostered pre-class accountability, while artificial intelligence systems guided learners to monitor their strategies, detect errors, and adjust their performance in real-time. The results suggested that students learned most effectively when they were actively involved, using technological affordances to regulate their own progress rather than relying exclusively on the teacher. Such practices facilitated the development of long-term self-regulation skills that extend beyond the immediate instructional setting.

The study also contributed to broader theoretical frameworks of CALL and self-regulation. The results supported the argument that self-monitoring was a key process within Zimmerman's model of self-regulated learning, illustrating that well-targeted instructional design could enhance this metacognitive process directly. These results expand self-regulated learning theory by demonstrating how technology can serve as a delivery medium and an adaptive regulatory scaffold in the monitoring phase of Zimmerman's cycle. This contributes to CALL by demonstrating how AI tools can function not only as content delivery systems but also as dynamic agents that affect learner agency and monitoring behavior.

From a CALL perspective, the study enriched existing models by situating AI-supported teaching not only as a practice tool but as a metacognitive model. Lacing flipped approaches with AI highlighted how technological mediation could go beyond content conveyance to affect learners' agency over their learning processes. These findings implied that next-generation CALL models must consider the twofold role of technology: as a medium for delivery and as a control agent that reconfigures learner autonomy.

Certain limitations have to be remembered while interpreting the findings. First, the sample size was minimal, which restricted the external validity of the findings for more comprehensive EFL populations. Secondly, a single institutional setting was used for the study, thereby limiting the external validity of the conclusions. Thirdly, the intervention was over a short period, possibly missing the long-term effects of flipped and AI-driven instruction on students' self-regulation and writing skills. Lastly, the use of self-report measures for metacognitive regulation raises a question about the potential for implicit biases to affect students' perceptions and how students self-report on their work. Despite these limitations, the study's contribution remains intact but needs further fine-tuning and replication of outcomes. In addition, the specific AI tools used (e.g., GPT-3, automated writing feedback) limit the generalizability of the results to platforms outside of what is possibly written here, and some discrepancies in citation indicate overall imprecision, which must be improved in any other study that documents work with AI. In order to improve the possibility of building more coalesced understanding of AI-assisted and flipped flipping experiences, future studies will need to address these limitations in order to understand the effectiveness of the study's findings, specifically sustained effects over time, the implications of longitudinal studies with longer periods of intervention are warranted; also, mixed methods would yield better information because a combined quantitative share will balance account qualitative reports of learner experiences. Future studies require additional researchers to examine different types of AI technologies, including adaptive writing assistants, speech recognition programs, and the use of multimodal and/or multiple feedback systems, to understand better the overall success of learner uptake of AI technologies in selected areas of skill. Finally, conducting the original learning activities or stimulating the research study with larger learning groups that have a wider range of educational or cultural backgrounds would increase both the study's transferability and reliability.

In summary, the research emphasizes that although flipped and AI-based instruction have the potential to improve EFL learners' self-regulatory competencies and writing, the adaptive immediacy built into the AI distinction makes both approaches potentially very powerful, but particularly in hybrid courses where the benefits of artificial intelligence complement rather than replace human-centered pedagogical designs.

Acknowledgment

We all thank the participants for their time, energy, and cooperation.

Authors' Contributions

All authors have conducted the study, collected data, analyzed and interpreted the data, and written up the manuscript.

Funding

The study did not receive any funding.

Competing Interests

The authors declare that there is no conflict of interest.

References

Abd Elkader, S. M. A. (2023). The impact of using artificial intelligence on enhancing EFL language fluency and self-regulation for the preparatory stage students in distinguished governmental language schools [Doctoral dissertation, Ain Shams University]. https://files.eric.ed.gov/fulltext/ED630026.pdf

Afzali, Z., & Izadpanah, S. (2021). The effect of flipped classroom on EFL learners' speaking skill development: Iranian high school context. *Education and Information Technologies*, *26*, 4047–4061. https://doi.org/10.1007/s10639-021-10554-7

Ahmadi, M. R. (2018). The use of technology in English language learning: A literature review. *International Journal of Research in English Education,* *3*(2), 115–125. https://doi.org/10.29252/ijree.3.2.115

Alizadeh, M., & Fard Kashani, A. (2020). The effect of flipped classroom on Iranian EFL learners' writing performance and anxiety. *Education and Information Technologies, 25*, 3899–3919. https://doi.org/10.1007/s10639-020-10129-8

Cheng, L., Ritzhaupt, A. D., & Antonenko, P. (2019). Effects of the flipped classroom instructional strategy on students' learning outcomes: A meta-analysis. *Educational Technology Research and Development, 67*(4), 793–824. https://doi.org/10.1007/s11423-018-9633-7

Choi, J., & Lee, S. (2024). The impact of chatbots based on large language models on second language vocabulary acquisition. *Heliyon*, *10*(3), e25370. https://doi.org/10.1016/j.heliyon.2024.e25370

Dekhakhena, A. (2025). AI-Powered personalized learning in EFL acquisition: exploring adaptive instruction and feedback systems. *Journal of Studies in Language, Culture and Society (JSLCS)*, *8*(1), 111–131.

Dörnyei, Z. (2010). Researching motivation: From integrativeness to the ideal L2 self. *Introducing applied linguistics: Concepts and skills*, *3*(5), 74–83.

Du, X., & Daniel, S. (2024). Transforming language education: A systematic review of AI-powered chatbots for English as a foreign language speaking practice. *Computers and Education: Artificial Intelligence*, *5*, 100123. https://doi.org/10.1016/j.caeai.2024.100123

Durmuşoğlu, H., & Demirok, M. S. (2025). Effects of flipped classrooms on the academic achievements, individualized education plan competencies and quality of related preparation of pre-service teachers. *Behavioral Sciences (Basel, Switzerland)*, *15*(4), 438. https://doi.org/10.3390/bs15040438

Gayed, J. M., Carlon, M. K. J., Oriola, A. M., & Cross, J. S. (2022). Exploring an AI-based writing assistant's impact on English language learners. *Heliyon*, *8*(1), e08733. https://doi.org/10.1016/j.heliyon.2022.e08733

González-Gómez, D., Jeong, J. S., Airado Rodríguez, D., & Cañada-Cañada, F. (2016). Performance and perception in the flipped learning model: An initial approach to evaluate the effectiveness of a new teaching methodology in a general science classroom. *Journal of Science Education and Technology*, *25*, 450–459. https://doi.org/10.1007/s10956-016-9605-9

Hashemian, M., & Heidari, A. (2013). The relationship between L2 learners' motivation/attitude and success in L2 writing. *Procedia-Social and Behavioral Sciences*, *70*, 476–489.

Hazaymeh, W. A., Alshaikhi, T., Abdul Wahab, M. O., & Khasawneh, M. A. (2025). Evaluating AI applications for enhancing listening comprehension among EFL students in English language in real-world scenarios. *World Journal of English Language, 15*(6), 209–223. http://wjel.sciedupress.com

Holmes, W., Porayska-Pomsta, K., Holstein, K., Sutherland, E., Baker, T., Shum, S. B., & Santos, O. C. (2021). Ethics of AI in education: Towards a community-wide framework. *International Journal of Artificial Intelligence in Education*, *31*, 1–23. https://doi.org/10.1007/s40593-020-00239-1

Hooshyar, D., Yousefi, M., & Lim, H. (2019). The effectiveness of intelligent tutoring systems on student performance: A meta-analysis. *Computers & Education*, *136*, 1–14. https://doi.org/10.1016/j.compedu.2019.03.005

Hsu, W.-C., Ching, Y.-H., & Grabowski, B. L. (2023). The effects of AI-based vocabulary learning tools on EFL learners' vocabulary acquisition. *Journal of Educational Technology & Society*, *26*(2), 1–15. https://www.jstor.org/stable/48632759

Huang, Y. N., & Hong, Z. R. (2016). The effects of a flipped English classroom intervention on students' information and communication technology and English reading comprehension. *Educational Technology Research and Development*, *64*(2), 175–193.

Jacobs, P. A. (1981). Mutation rates of structural chromosome rearrangements in man. *American Journal of Human Genetics*, *33*(1), 44.

Klassen, A. C., Creswell, J., Plano Clark, V. L., Smith, K. C., & Meissner, H. I. (2012). Best practices in mixed methods for quality of life research. *Quality of life Research*, *21*(3), 377–380.

Lee, J., & Warschauer, M. (2020). The use of AI-powered writing feedback in English as a second language writing instruction. *Language Learning & Technology*, *24*(1), 1–15. https://hdl.handle.net/10125/44767

Lee, J., Lim, C., & Kim, H. (2017). Development of an instructional design model for flipped learning in higher education. *Educational Technology Research and Development*, *65*(2), 427–453. https://doi.org/10.1007/s11423-016-9502-1

Luo, N., & Wang, J. (2023). Personalized AI feedback in English writing: Effects on accuracy and learner autonomy. *System*, *117*, 102912. https://doi.org/10.1016/j.system.2023.102912

Mehring, J. (2017). Technology as a teaching and learning tool in the flipped classroom. In *Digital Language Learning and Teaching* (pp. 67–78). Routledge.

Nunnally, J. C. (1978). *Psychometric theory* (2nd ed.). McGraw-Hill.

Pallant, J. (2020). *SPSS Survival Manual: A step-by-step guide to data analysis using IBM SPSS*. Routledge.

Perrotta, C., & Selwyn, N. (2020). AI and education: The importance of teacher and student agency. *Learning, Media and Technology, 45*(3), 251–264. https://doi.org/10.1080/17439884.2020.1686017

Pintrich, P. R. (1991). *A manual for the use of the Motivated Strategies for Learning Questionnaire (MSLQ)*. Ann Arbor, MI: University of Michigan, National Center for Research to Improve Postsecondary Teaching and Learning. https://eric.ed.gov/?id=ED338122

Pintrich, P. R. (2000). Multiple goals, multiple pathways: The role of goal orientation in learning and achievement. *Journal of Educational Psychology*, *92*(3), 544.

Rahimi, M., & Yadollahi, S. (2017). Iranian EFL learners' attitudes towards self-regulated learning in blended environments. *Computers in Human Behavior, 68,* 93–104. https://doi.org/10.1016/j.chb.2016.11.032

Rezaei, A., & Davoudi, M. (2016). The impact of blended learning on EFL learners' motivation and reading comprehension. *Journal of Language Teaching and Research, 7*(3), 590–598. https://doi.org/10.17507/jltr.0703.20

Wang, W., Chen, L., Ye, H., & Zhang, J. (2024). Artificial intelligence in education: A systematic review and bibliometric analysis. *Expert Systems with Applications, 252*, 124167. https://doi.org/10.1016/j.eswa.2024.124167

Wang, Y., & Liang, J. (2023). Artificial intelligence in language instruction: Impact on English learning achievement, L2 motivation, and self-regulated learning. *Frontiers in Psychology*, *14*, 1261955. https://doi.org/10.3389/fpsyg.2023.12619556

Yilmaz, R. M., & Yilmaz, F. G. K. (2023). The effect of AI-powered chatbots on EFL learners' speaking anxiety and motivation. *Computer Assisted Language Learning, 36*(6), 1234–1252. https://doi.org/10.1080/09588221.2021.2006213

Zainuddin, Z. (2018). Students' learning performance and perceived motivation in gamified flipped-classroom learning environments: A comparison between fully flipped and semi-flipped classrooms. *International Journal of Educational Technology in Higher Education, 15*, 9. https://doi.org/10.1186/s41239-018-0024-x

Zainuddin, Z., & Perera, C. J. (2019). Exploring students' competence, autonomy, and relatedness in the flipped classroom pedagogical model. *Journal of Further and Higher Education, 43*(1), 115–126. https://doi.org/10.1080/0309877X.2017.1356916

Zandi, H., & Khezrlou, S. (2023). The effect of AI-enhanced feedback on EFL academic writing: Iranian postgraduate students' perspectives. *System, 114,* 102957. https://doi.org/10.1016/j.system.2023.102957

Zarei, G. R., & Aghaei, K. (2019). The effect of blended learning on EFL learners' reading comprehension: An Iranian context. *Cogent Education, 6*(1), 1562032. https://doi.org/10.1080/2331186X.2019.1562032

Zawacki-Richter, O., Marín, V. I., Bond, M., & Gouverneur, F. (2019). Systematic review of research on artificial intelligence applications in higher education – where are the educators? *International Journal of Educational Technology in Higher Education*, *16*, 39. https://doi.org/10.1186/s41239-019-0171-0

Zhai, X., Chu, X., & Wang, M. (2023). AI in education: A systematic review of empirical research from 2011 to 2021. *Computers & Education: Artificial Intelligence*, *4*, 100110. https://doi.org/10.1016/j.caeai.2023.100110

Zimmerman, B. J. (2000). Attaining self-regulation: A social cognitive perspective. In M. Boekaerts, P. R. Pintrich, & M. Zeidner (Eds.), *Handbook of self-regulation* (pp. 13–39). Academic Press. https://doi.org/10.1016/B978-012109890-2/50031-7