INTERNATIONAL JOURNAL OF SOCIAL SCIENCES BULLETIN
ISSN: (E) 3007-1917 (P) 3007-1909
Volume 4, Issue 4, 2026

IMPACT OF AI-INTEGRATED PEDAGOGY ON ORAL AND WRITTEN PROFICIENCY OF ESL LEARNERS

Abdul Rehman[1], Dr. Mahwish Shamim[2], Amna Ahmad[3], Dr. Muhammad Arfan Lodhi[*4]

[1]*M. Phil Scholar, NCBA&E Alhamra University, Bahawalpur, Pakistan*
[2]*Assistant Professor, University of Rasul, Mandi Bahauddin, Pakistan*
[3]*M.Phil Scholar, Allama Iqbal Open University, Islamabad, Pakistan*
[*4]*Higher Education Department, Punjab, Pakistan*

[1]chrehman53@gmail.com, [2]mahwish.shamim@putrasul.edu.pk, [3]amnaqazi158@gmail.com, [*4]samaritan_as@hotmail.com

Corresponding Author: *
Dr. Muhammad Arfan Lodhi
DOI: https://doi.org/10.5281/zenodo.19947417

Received	Accepted	Published
07 March 2026	14 April 2026	30 April 2026

ABSTRACT

Learning English as a foreign language in Pakistani public schools has never been straightforward. Overcrowded classrooms, limited one-on-one contact time, and little opportunity for spoken practice have combined to leave many Grades 8 students struggling with the language skills particularly speaking fluency and written accuracy, which are formally assessed at the end of the school cycle. This study asked whether three AI-based language tools (ELSA Speak, Duolingo, and Grammarly), integrated into ten weeks of regular English lessons, could make a genuine, measurable difference to those outcomes. A quasi-experimental pretest-posttest design was used, with thirty students receiving AI-supported instruction and thirty taught through conventional methods. Thirty experimental group students also completed a 20-item questionnaire on their awareness, usage, and perceptions of the tools, and forty practicing English teachers shared their views on AI integration through a mixed-methods questionnaire. The quantitative data were analyzed through SPSS 26 and the teacher qualitative responses through thematic analysis. Results were clear: students who worked with the AI tools made substantially larger gains on every measured sub-skill than their peers in the control group, they reported strong engagement with the tools, and their motivation climbed across the intervention period. Teachers were broadly supportive of the direction, even as they pointed honestly to the infrastructure gaps and professional development needs that currently stand in the way of full implementation.

Keywords: *AI-Enhanced Language Learning; Technology Acceptance Model; Zone of Proximal Development; Mixed-Methods; TPACK Framework; Pakistan*

1. INTRODUCTION

1.1 Background of the Study

Walk into any government secondary school English class in Bahawalpur and the challenge is immediately visible. One teacher; thirty-five or forty students; a syllabus to get through; and almost no time to give individual feedback on how a student is pronouncing a word while constructing a sentence, or organizing a paragraph. This is not a failure of teacher commitment; rather it is a structural constraint that no amount of dedication can fully overcome. Artificial intelligence has changed what is technically possible in this situation. Applications powered by speech

recognition, adaptive task engines, and natural language processing can now deliver the kind of targeted, personalized feedback that only individual tutoring could previously provide (Hwang et al., 2021; Chen et al., 2020). For EFL students in Pakistani public schools, where the gap between what instruction demands and what instruction can realistically deliver has always been wide, these tools represent something more than a convenience — they represent a genuine structural alternative.

Three tools form the core of this study. ELSA Speak listens to each student's spoken English and identifies specific pronunciation errors at the phoneme level, offering corrective feedback that a classroom teacher simply cannot replicate for thirty students simultaneously. Duolingo builds speaking practice into adaptive game-like sequences that adjust difficulty based on how the student is actually performing, keeping every learner productively challenged without tipping into frustration or boredom. Grammarly reads students' own written work and explains grammatical problems in context, turning the revision process from a passive correction exercise into an active learning cycle. Together, these three tools fill in precisely the gaps that conventional large-class instruction leaves behind.

1.2 Statement of the Problem

For all the international enthusiasm about AI in language education, the evidence base has been built almost entirely from university classrooms in East Asia, the Gulf, and Western Europe. Students in South Asian government secondary schools are a different population — younger, in larger classes, with less technological infrastructure, and facing a different kind of high-stakes outcome when Grade 8 assessments arrive. The teachers who work with these students make daily decisions about technology use without any local data to guide them. No one has yet asked, in a systematic and methodologically rigorous way, whether AI tools can produce measurably better speaking and writing outcomes for this population, or whether students and teachers in this context are willing and able to engage with these tools in ways that produce learning gains. These are not abstract academic questions; they are practical decisions that affect whether public school students in Bahawalpur get access to the same quality of individualized language feedback that their private school counterparts receive.

1.3 Research Gap

Three specific gaps run through the existing literature. The first is geographical: Pakistan's EFL research is largely descriptive and qualitative, and quasi-experimental studies measuring actual performance change in AI-integrated lessons are essentially absent from the South Asian literature. The second is demographic: almost every existing study of AI language tools involves university students, leaving Grade 8 learners — at a critical and assessable transition point — without a directly relevant evidence base. The third is methodological: most studies capture either performance outcomes or learner attitudes, rarely both at once. When only one side of the story is told, it is impossible to understand whether positive performance gains are actually accompanied by genuine engagement, or whether positive attitudes translate into real learning. This study deliberately brought both strands together within a single design.

1.4 Research Objectives

1. To critically evaluate the impact of AI based language tools on enhancing 8th grade students' speaking skills, focusing on pronunciation, fluency, and contextual vocabulary.
2. To analyze the effects of AI- powered feedback on the writing proficiency of 8th grade students, particularly in areas of syntax, coherence, grammar, and style.
3. To assess the extent to which AI-based tools promote student engagement and motivation compared to traditional language instruction.
4. To identify the challenges and limitations teachers and students face while integrating AI into language instruction, including issues of accessibility, cultural relevance, and feedback accuracy.

1.5 Research Questions

1. What specific impacts do AI-based language tools have on the speaking skills of 8th-

grade learners in terms of pronunciation, fluency, and vocabulary?

2. How do AI- powered feedback mechanisms influence students' writing skills, particularly in grammar, coherence, and linguistic accuracy?

3. How do AI driven learning tools affect student engagement and motivation in language learning compared to traditional methods?

4. What challenges and limitations are experienced by teachers and students in implementing AI for language education in a middle school context?

1.6 Hypotheses

H1: A statistically significant positive correlation exists between frequency of AI tool use and perceived language skill improvement among Grade 8 EFL learners.

$H0_1$: No significant relationship exists between AI tool usage frequency and perceived language skill improvement among Grade 8 EFL learners.

H2: Grade 8 EFL learners receiving AI-based instruction will show statistically significantly greater improvement in speaking and writing skills than those taught through conventional methods.

$H0_2$: No statistically significant difference in speaking or writing performance exists between Grade 8 EFL learners in AI-based and conventional instructional conditions.

H3: Grade 8 EFL learners in the AI-supported instructional condition will report higher motivation and engagement than baseline patterns in the control group would predict.

$H0_3$: No statistically significant difference in motivation or engagement exists between learners in AI-based and conventional instructional conditions.

2. Literature Review

2.1 Theoretical Framework

The study draws on three theoretical traditions that, taken together, explain both why AI tools should work as language-learning instruments and why they sometimes fall short in practice. The starting point is Vygotsky's (1978) Zone of Proximal Development. The central insight of the ZPD is deceptively simple: learners grow fastest when they receive support at the precise boundary of what they can currently do on their own. A task that is too easy produces no growth; a task that is too hard produces only frustration. What produces genuine development is a task that sits just beyond current capability, supported by a more competent other. In a Pakistani public school classroom of thirty-five students, a teacher cannot calibrate support to every individual's ZPD simultaneously — it is structurally impossible. ELSA Speak, Duolingo, and Grammarly each address this constraint directly. ELSA identifies the specific phonemes a particular student is mispronouncing and corrects those, not the phonemes of an imaginary average learner. Duolingo tracks individual performance and adjusts the difficulty of the next task accordingly. Grammarly targets the exact grammatical patterns that appear in each student's own writing. These are not approximate solutions — they are each a near-perfect digital implementation of ZPD-mediated scaffolding.

The Technology Acceptance Model (Davis, 1989) adds the second necessary lens. The central claim of TAM is that how useful people believe a technology will be, and how easy they believe it will be to use, are stronger predictors of actual engagement than access alone. This distinction matters enormously for this study. Having devices loaded with Duolingo means nothing if students believe the tool is too complicated or irrelevant. The questionnaire was designed precisely to capture these perceived usefulness and ease-of-use dimensions, and the regression analysis tested whether those perceptions actually predicted engagement depth and felt improvement. TAM also makes sense of a pattern in the teacher data that might otherwise seem contradictory: strong optimism about AI's future role coexisting with limited current use. In TAM terms, this is not inconsistency — it is the predictable consequence of high perceived usefulness without sufficient perceived ease of use.

Constructivist learning theory, most closely associated with Piaget (1970), provides the third framework. For constructivists, deep learning is not something that happens to a learner when content is delivered to them — it is something a learner does, actively constructing understanding through encounter with challenging material, error, feedback, and revision. The Grammarly-supported

writing cycle in this study is a near-ideal implementation of constructivist revision: rather than handing back a corrected essay, Grammarly presents each error to the student with an explanation and requires the student to decide what to do about it. Each correction requires an active cognitive decision, not passive acceptance. The TPACK framework (Mishra & Koehler, 2006) supplements these three by describing what teachers actually need in order to make technology work pedagogically: not just technical skill, and not just good teaching instinct, but the specific capacity to weave both together with subject knowledge. The teacher data in this study shows that this capacity is still developing among Bahawalpur's English teachers — not because they are uncommitted, but because the professional development infrastructure to support it has not yet been built.

2.2 Review of Empirical Studies

The international literature on AI in language learning has expanded rapidly since roughly 2019, moving from theoretical advocacy toward empirical testing. The studies most relevant to this investigation are synthesized in Table 1, selected for their focus on productive language skills, tool types, and educational levels closest to the present context.

Table 1. Summary of the previous research

Author(s) & Year	Context	AI Tool / Skill	Main Outcomes	Gap Addressed Here
Hwang et al. (2021)	Global review	AI in education / broad	Key roles of AI identified; personalised feedback highlighted as highest-value function	Grade 8 school evidence in South Asia lacking
Chen et al. (2020)	Global review	Educational technology trends	Four-decade analysis; AI identified as emerging dominant force in educational technology	Localised quasi-experimental evidence absent
Cucchiarini et al. (2009)	Netherlands — adults	ASR pronunciation / Speaking	Automated speech recognition feedback produced 20-30% pronunciation gains over control	School-age South Asian learners not studied
Neri et al. (2002)	Netherlands — adults	CAPT pronunciation / Speaking	Computer-aided pronunciation training significantly outperformed no-feedback condition	Grade 8 public-school context absent
Koltovskaia (2020)	USA — university	Grammarly / Writing	Students engaged more meaningfully with Grammarly feedback than with teacher corrections	EFL writing in South Asian schools absent
Ranalli et al. (2017)	USA — university	AWE formative / Writing	Automated evaluation raised accuracy and reduced grammar error density	Grade 8 EFL in Pakistan absent
Mishra & Koehler (2006)	USA — teacher education	TPACK / Teacher knowledge	TPACK predicts quality of technology integration in classroom delivery	Teacher TPACK in Pakistani public schools undocumented
Bygate (2016)	Global review	Task-based speaking /	Task repetition and AI cycling linked to oral fluency gains	App-based AI speaking tasks in Grade 8 schools

		Speaking		not studied
Godwin-Jones (2017)	USA – university	AI language apps / Both	AI tools for oral and written skills outperformed delayed teacher correction	Developing-country school contexts absent
Nazir & Khalid (2022)	Pakistan – secondary	AI-assisted tasks / Both	Positive outcomes noted; infrastructure barriers documented	Quasi-experimental performance data still needed for Pakistan
Zawacki-Richter et al. (2019)	Global review	AI in higher education	AI most effective for personalised feedback and adaptive assessment at scale	Secondary school EFL in South Asia underrepresented
Fryer & Carpenter (2006)	Japan – university	Chatbot / Speaking	AI speaking practice improved willingness to communicate in EFL learners	School-age learners and constrained settings not covered
Cohen et al. (2018)	Global – methodology	Mixed methods design	Mixed-methods triangulation yields more complete and contextually valid evidence	Single-method Pakistani EFL research dominates the literature

2.3 Conceptual Framework

The study positions AI-based teaching as the independent variable, operationalised through ELSA Speak (pronunciation and fluency), Duolingo (oral task performance), and Grammarly (writing accuracy and composition). EFL speaking skill development and EFL writing skill development are the two dependent variables. Within the questionnaire component, AI tool usage frequency is the predictor variable and perceived skill improvement is the criterion, with motivation and confidence included as secondary outcome constructs. Three moderating conditions shape how the independent variable acts on outcomes: the quality of available digital infrastructure, the depth of teacher TPACK competence, and students' existing familiarity with digital learning tools. The theoretical stance connecting all of this is ZPD-mediated scaffolding and AI tools make individualized, calibrated feedback available at scale, which is precisely what conventional large-class instruction cannot do.

3. Methodology

3.1 Research Design

The study used a quasi-experimental pretest-posttest design with a concurrent mixed-methods component (Creswell & Plano Clark, 2018). The reasoning behind this choice reflects the dual ambition of the research: measuring whether AI instruction actually improved performance in a way that could support causal inference, while also understanding the experience of students and teachers in ways that numbers alone cannot capture. The experimental group's speaking and writing scores were compared against the control group's across two time points. The student questionnaire data were analysed descriptively and inferentially. Teacher open-ended responses were subjected to thematic analysis. The two strands were then triangulated to produce a reading of the findings that is both empirically grounded and contextually meaningful. All ethical requirements were met: written permission was obtained from school authorities, teacher consent was collected, and guardian consent was obtained for every student participant. All data were pseudonymised before analysis.

3.2 Population and Sampling

The study drew its participants from Grade 8 EFL classes in government-run secondary schools in Bahawalpur district, Punjab, Pakistan. Choosing Grade 8 was deliberate: it is the year in which students in the Punjab system face their first major external English language assessment, making the stakes real, while also being a year in which learners are still in the large-class public school environment where the instructional constraints this study seeks to address are most acute. Purposive sampling was used to identify two schools whose class sizes, available infrastructure, and prior assessment

performance were similar enough to make comparison meaningful. Students were allocated to experimental or control conditions by school of enrolment, keeping intact class groups together and avoiding the disruption that cross-class reassignment would cause. The 20-item student questionnaire was given only to the thirty experimental group participants, because only they had actually used the AI tools that the questionnaire asked about.

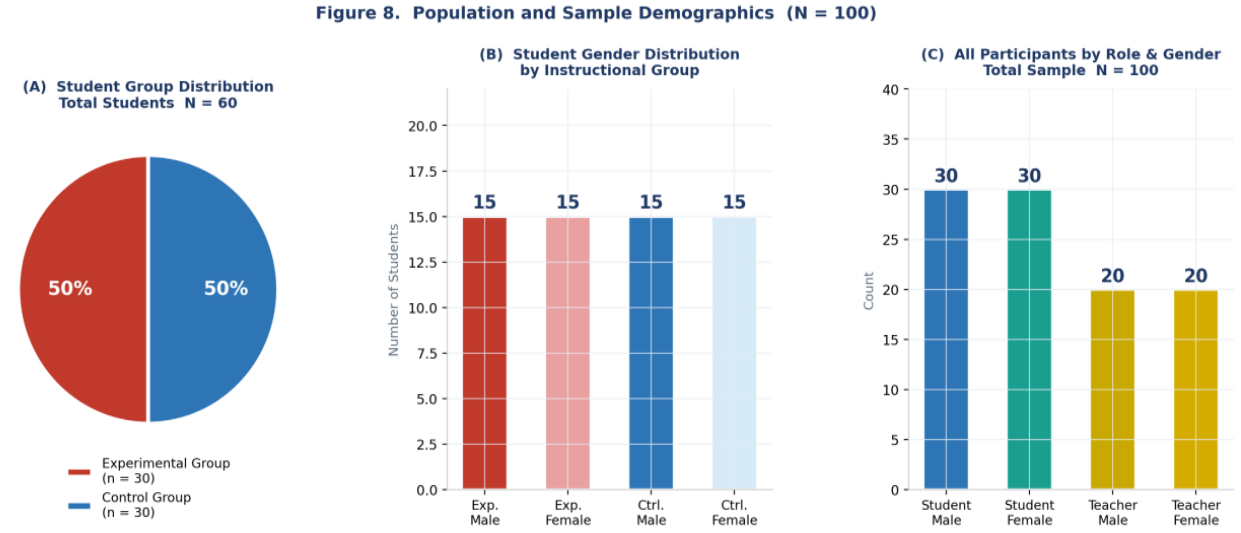

Figure 1. Population and Sample Demographics

Table 2. Sample Description

Participant Category	Condition / Role	Male (n)	Female (n)	Total (N)
Students	Experimental Group – AI-Assisted Instruction	15	15	30
Students	Control Group – Conventional Instruction	15	15	30
Teachers	Questionnaire	20	20	40
Total		50	50	100

3.3 Instrumentation

Data were gathered through five instruments. Two parallel performance tests — a speaking test and a writing test — were administered to all sixty students at pre-test and again at post-test. The speaking test covered ELSA Speak-aligned pronunciation and fluency (10 points) and a Duolingo-style oral task battery including picture description, sentence repetition, and short spoken responses (10 points). The writing test assessed grammar and vocabulary in the Grammar aligned sense (10 points) alongside a creative paragraph writing task (30 points), giving a 50-point composite total. The 20-item student questionnaire, administered only to experimental group participants, asked about awareness of the tools, frequency of use, perceived skill improvement, and motivation. The teacher questionnaire combined Likert-scale ratings with open-ended qualitative prompts about integration experiences and barriers.

Table 3. Description of Student's Questionnaire

Section	Domain	Items	Scale	Sample Statement
A	Awareness of AI Tools	5	0 (Never) to 5 (Always)	I know about AI learning tools like Duolingo or ELSA Speak.
B	Frequency of AI Tool Usage	5	0 (Never) to 5 (Always)	I use AI tools to practise my English speaking.
C	Perceived Skill Improvement	5	0 (Never) to 5 (Always)	AI learning tools make my English lessons more interesting.
D	Motivation and Confidence	5	0 (Never) to 5 (Always)	AI tools make me want to practise English more.
	Total Items	20		

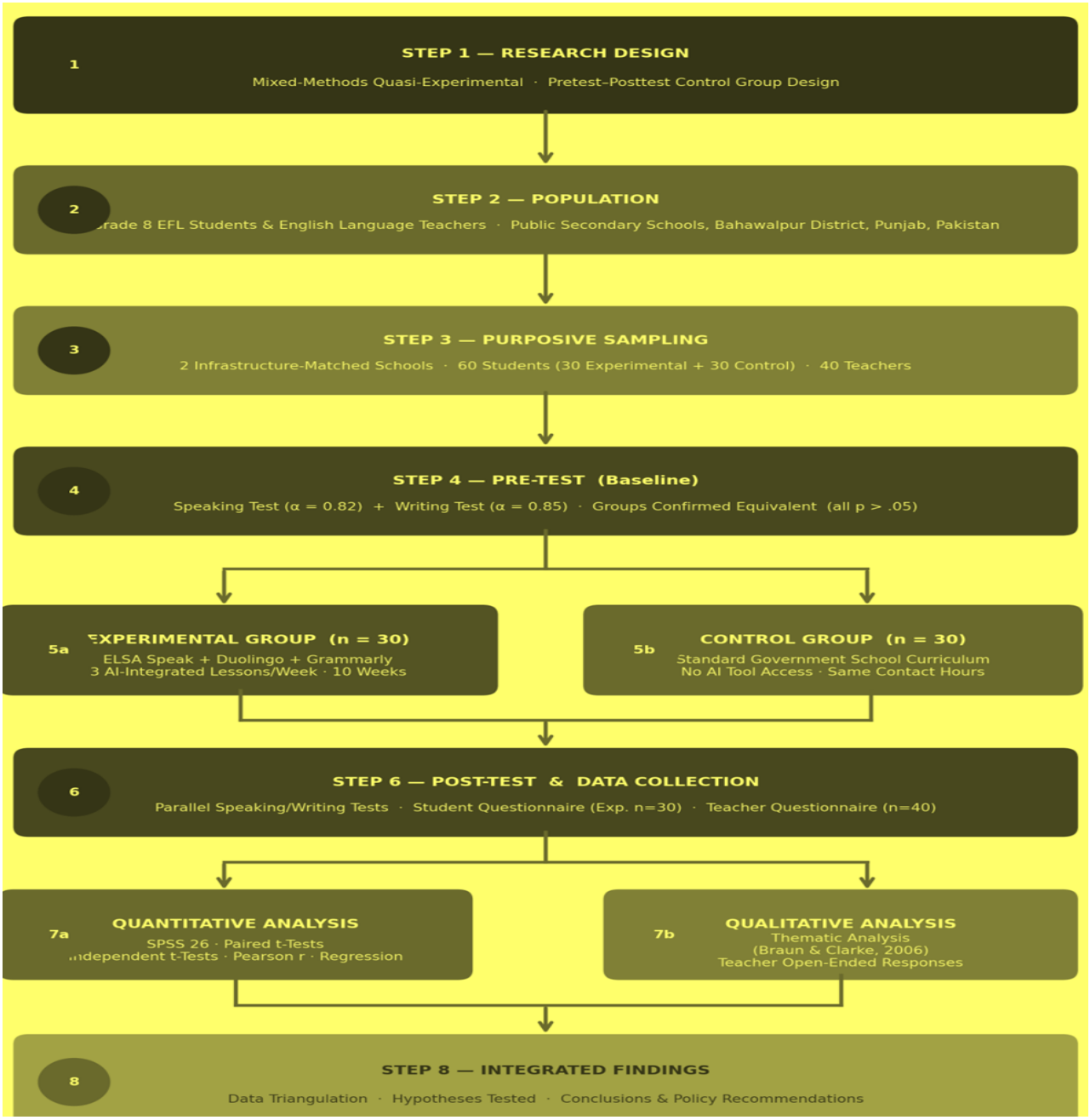

Figure 2. Research Procedure Flowchart

3.4 Validity and Reliability

Getting the instruments right before data collection began was a priority. Three experienced EFL teachers and one AI education specialist worked through every test item and questionnaire statement, checking that each one genuinely corresponded to a specific learning objective in the Punjab Grade 8 English curriculum. Items that could not be anchored to a clear curriculum outcome were revised or cut. The rubrics were then cross-referenced against Punjab Examination Commission performance descriptors to establish criterion validity. A small pilot group of students and teachers worked through the instruments and flagged anything confusing, which led to minor rewording of a handful of items. When all four main instruments were checked for internal consistency using Cronbach's alpha, every one

returned a coefficient above 0.70, confirming that the items within each instrument were measuring the same underlying construct rather than pulling in different directions (Field, 2013).

Table 4. Full Instruments Overview

Instrument	Administered To	Purpose	Format	Cronbach's Alpha
Speaking Test (Pre & Post)	All 60 students	Measure oral skill change	Rubric; 20 pts	0.82
Writing Test (Pre & Post)	All 60 students	Measure writing skill change	Rubric; 30 pts	0.85
Student Questionnaire (20 items)	30 Experimental students only	AI usage, perceptions, motivation	Frequency scale 0-5	0.79
Teacher Questionnaire	40 teachers	AI perceptions, barriers, training	Likert 1-5 + qualitative	0.80
Document Analysis	School and lesson records	Triangulate with implementation data	Qualitative	—

3.5 Data Collection Procedure

The data collection unfolded across three stages. Before anything else, the groundwork was laid: school permission obtained, teacher and guardian consent collected, and a half-day professional development session run to introduce teachers to the pedagogical rationale and practical operation of each AI tool. Pre-tests were then administered to both groups at the same time, under the same supervised conditions. For ten weeks, experimental group students used ELSA Speak, Duolingo, and Grammarly across three AI-integrated English lessons per week — ELSA Speak for pronunciation coaching, Duolingo for oral practice tasks, Grammarly for writing feedback and revision. The control group had equivalent lesson time with no AI access, following the standard government school curriculum. Immediately after the ten-week period, post-tests were administered to all sixty students and the student questionnaire was completed by the experimental group. The teacher questionnaire was distributed and collected during the same window. All data were pseudonymised before any analysis began.

3.6 Data Analysis

SPSS 26 handled all the quantitative analysis. Paired-sample t-tests were run for both groups to determine whether pre-to-post changes on each sub-skill and the total score were statistically significant within each condition. Independent-sample t-tests then compared post-test performance between the two groups to test whether the experimental group's advantage at the end of the intervention was statistically meaningful. Pearson correlation tested the relationship between how frequently students used the AI tools (Section B means) and how much improvement they perceived (Section C means), which is the empirical test of H1. A simple linear regression quantified the predictive strength of usage frequency on perceived improvement. Descriptive statistics were calculated for all instruments. Teacher open-ended responses were worked through using Braun and Clarke's (2006) six-phase thematic analysis approach: repeated reading and familiarization, initial coding, development of candidate themes, review of those themes against the data, definition and naming, and finally write-up. Quantitative and qualitative findings were brought together through triangulation following Denzin (1978), so that each strand of evidence could check, extend, and where necessary challenge the other.

4. Results

4.1 Demographic Profile

One hundred people participated in total: sixty Grade 8 students and forty English language teachers. Both student groups had the same gender balance — fifteen males and fifteen females each. The forty teachers were split evenly between twenty males and twenty females. Student participants

ranged from twelve to fourteen years old and had all been studying English as a required subject for at least three years. Every teacher had at least three years of classroom experience.

4.2 Descriptive Statistics

Table 5 lays out the mean scores and standard deviations for both groups at pre-test and post-test across all six sub-skills. Before the intervention, there was no statistically significant difference between the groups on any sub-skill — both groups came into the study with essentially the same starting point. By post-test, the two groups had diverged considerably. The experimental group's scores rose across every component. The control group also improved, reflecting ordinary learning progress, but the scale of improvement was nowhere near comparable. The total score tells the story most clearly: experimental group students gained an average of 12.04 points over the ten weeks, while control group students gained 3.83 points.

Table 5. Descriptive Statistics: Pre- and Post-Test Scores by Group

Sub-Skill (Max.)	Ctrl Pre M (SD)	Ctrl Post M (SD)	Exp Pre M (SD)	Exp Post M (SD)	Ctrl Gain	Exp Gain
ELSA Pronunciation & Fluency (10)	5.67 (0.84)	6.57 (1.07)	5.60 (1.10)	8.40 (1.07)	+0.90	+2.80
Duolingo Picture Description (5)	2.83 (0.70)	3.37 (0.72)	2.77 (0.63)	4.10 (0.80)	+0.53	+1.33
Duolingo Sentence Repetition (3)	1.40 (0.50)	1.77 (0.63)	1.50 (0.51)	2.37 (0.49)	+0.37	+0.87
Duolingo Short Response (2)	1.27 (0.64)	1.27 (0.45)	1.13 (0.57)	1.53 (0.51)	0.00	+0.40
Grammarly Grammar & Vocabulary (10)	5.27 (0.64)	6.03 (0.72)	5.07 (0.69)	7.47 (0.78)	+0.77	+2.40
Grammarly Creative Writing (30)	10.67 (0.96)	11.93 (1.31)	10.77 (1.14)	15.00 (1.20)	+1.27	+4.23
Total Score (50)	27.10 (2.48)	30.93 (3.12)	26.83 (2.35)	38.87 (2.98)	+3.83	+12.04

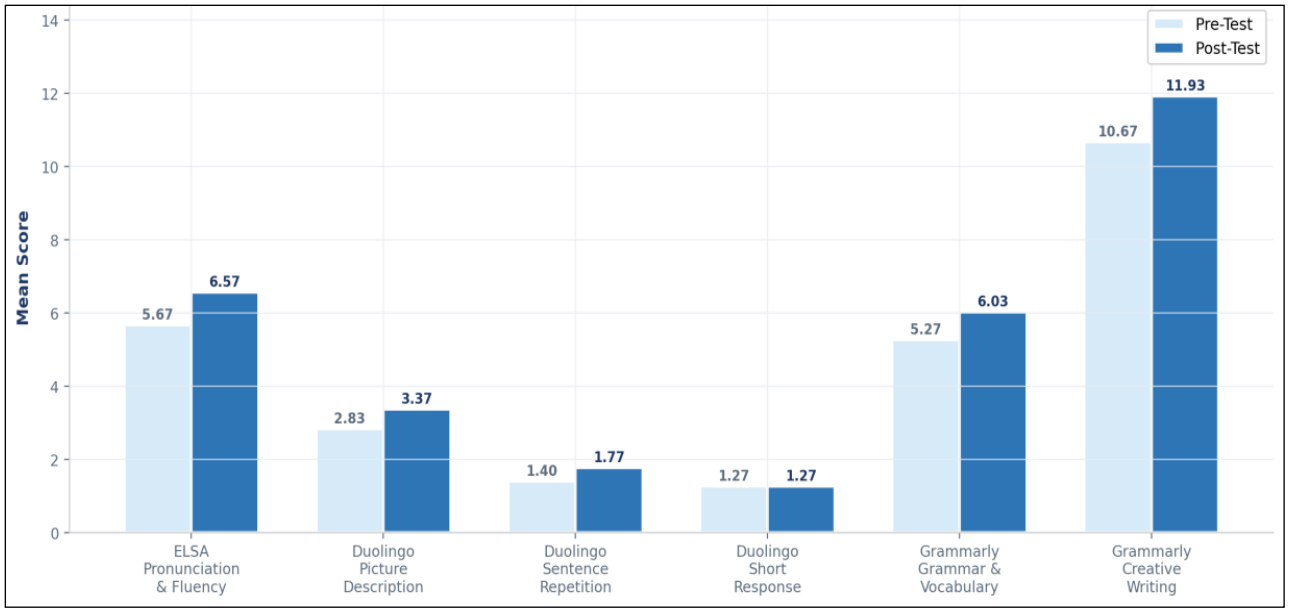

Figure 3. Control Group: Pre- and Post-Test Mean Scores

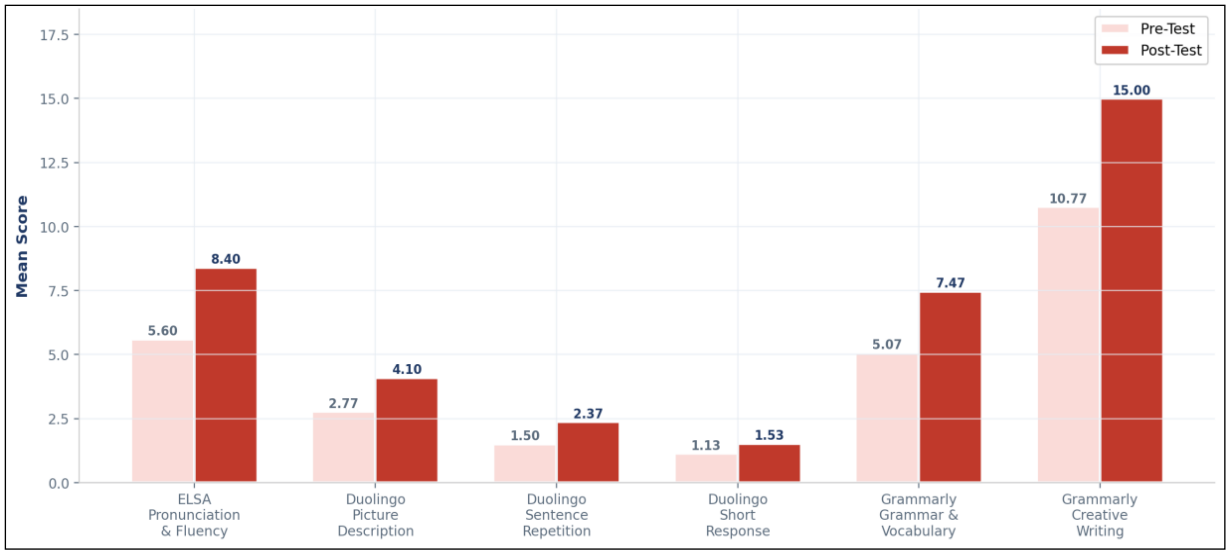

Figure 4. Experimental Group: Pre- and Post-Test Mean Scores

4.3 Pearson Correlation and Regression Analysis (H1)

To test H1, Pearson correlation was run between students' Section B mean scores (how frequently they used the AI tools) and their Section C mean scores (how much they felt their skills had improved). The correlation was strong and statistically significant ($r = 0.71$, $p < .001$). Students who used the tools more often also perceived greater improvement — and the relationship was not weak or marginal, it was a substantial association. The regression analysis built on this: usage frequency significantly predicted perceived improvement ($B = 0.63$, beta = .71, $F(1, 28) = 26.84$, $p < .001$), explaining approximately 49% of the variance in the outcome variable. H1 is supported. Figure 9 shows the scatter plot with regression line and confidence interval.

Table 6. Pearson Correlation and Linear Regression Results (Experimental Group, n = 30)

Statistic	Value	Description
Pearson r	0.71	Strong positive association between usage and perceived improvement
p (correlation)	< .001	Statistically significant — H1 supported
Unstandardised B	0.63	Each additional unit of usage predicts +0.63 units of perceived improvement
Standardised beta	.71	Large standardized effect
F (1, 28)	26.84	Overall regression model is significant
p (regression)	< .001	Model significant at alpha = .001
R squared	0.49	Usage frequency accounts for 49% of variance in perceived improvement

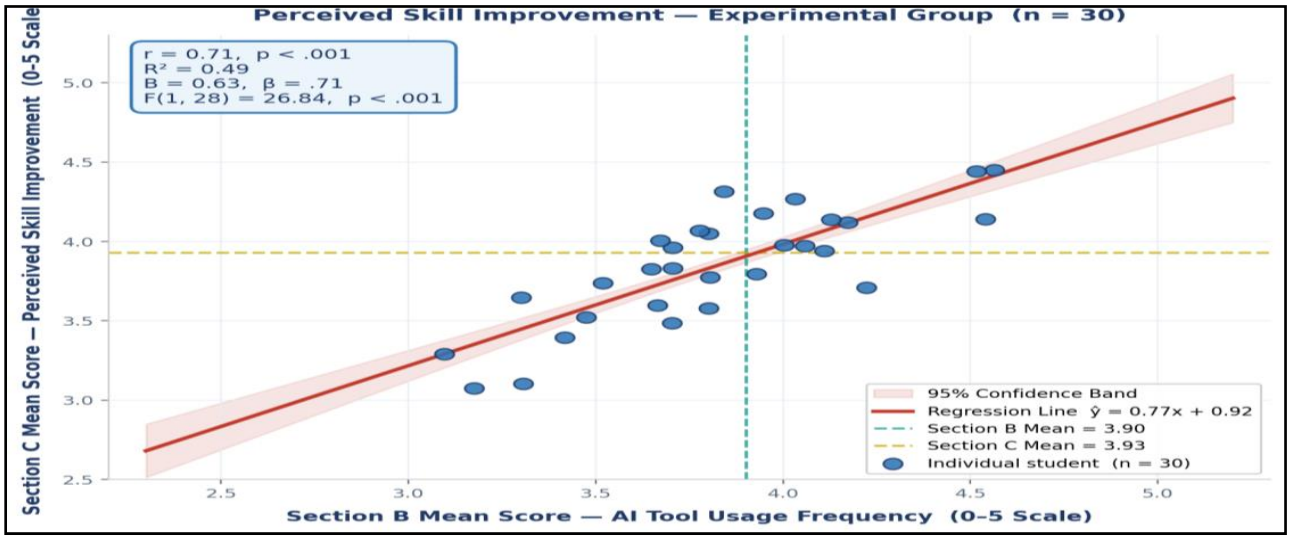

Figure 5. Scatter Plot: AI Tool Usage Frequency and Perceived Skill Improvement

Table 7. Within-Group Comparisons: Paired-Sample t-Tests

Sub-Skill	Group	Pre M	Post M	Gain	t	p
ELSA Pronunciation & Fluency	Control	5.67	6.57	+0.90	-5.341	< .001
ELSA Pronunciation & Fluency	Experimental	5.60	8.40	+2.80	-15.389	< .001
Duolingo Picture Description	Control	2.83	3.37	+0.53	-3.395	.002
Duolingo Picture Description	Experimental	2.77	4.10	+1.33	-9.103	< .001
Duolingo Sentence Repetition	Control	1.40	1.77	+0.37	-2.626	.014
Duolingo Sentence Repetition	Experimental	1.50	2.37	+0.87	-6.500	< .001
Duolingo Short Response	Control	1.27	1.27	0.00	0.000	1.000
Duolingo Short Response	Experimental	1.13	1.53	+0.40	-3.525	.001
Grammarly Grammar & Vocabulary	Control	5.27	6.03	+0.77	-5.769	< .001
Grammarly Grammar & Vocabulary	Experimental	5.07	7.47	+2.40	-19.484	< .001
Grammarly Creative Writing	Control	10.67	11.93	+1.27	-7.990	< .001
Grammarly Creative Writing	Experimental	10.77	15.00	+4.23	-31.853	< .001
Total Score (50)	Control	27.10	30.93	+3.83	-6.271	< .001
Total Score (50)	Experimental	26.83	38.87	+12.04	-21.441	< .001

Table 8. Post-Test Between-Group Comparison: Independent-Sample t-Tests

Sub-Skill	Ctrl Post M	Exp Post M	t	p	Decision
ELSA Pronunciation & Fluency	6.57	8.40	-6.628	< .001	H2 Supported
Duolingo Picture Description	3.37	4.10	-4.021	< .001	H2 Supported
Duolingo Sentence Repetition	1.77	2.37	-4.102	< .001	H2 Supported
Duolingo Short Response	1.27	1.53	-2.176	.033	Interpret with caution*
Grammarly Grammar & Vocabulary	6.03	7.47	-8.214	< .001	H2 Supported
Grammarly Creative Writing	11.93	15.00	-11.853	< .001	H2 Supported

Total Score (50)	30.93	38.87	-12.471	< .001	H2 Supported

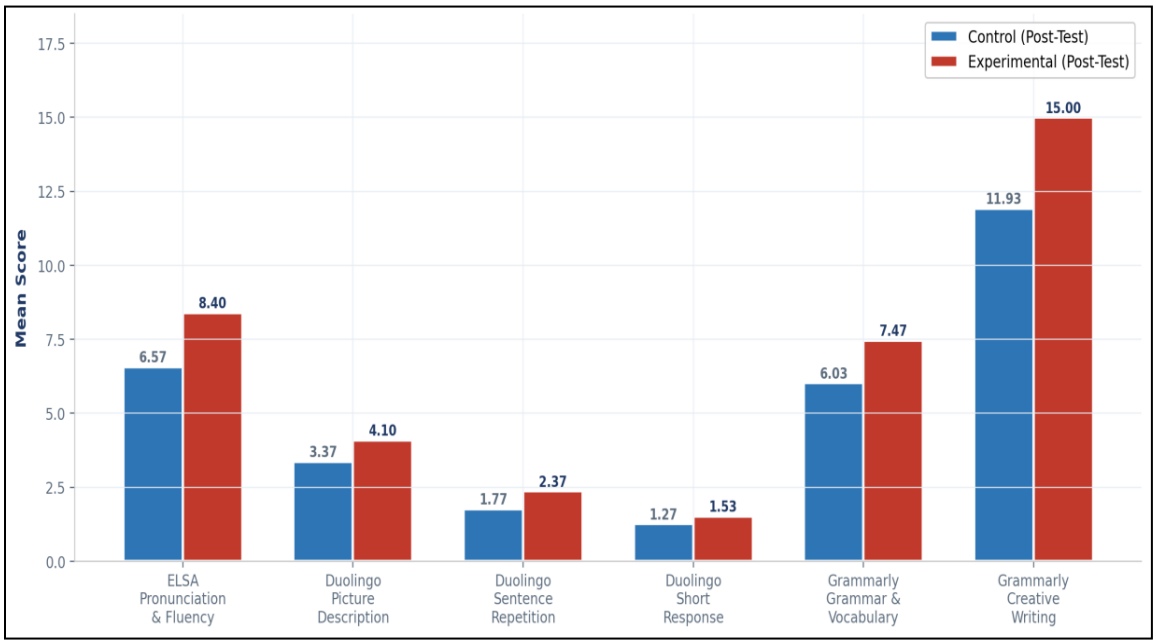

Figure 6. Post-Test Comparison: Control vs. Experimental Group

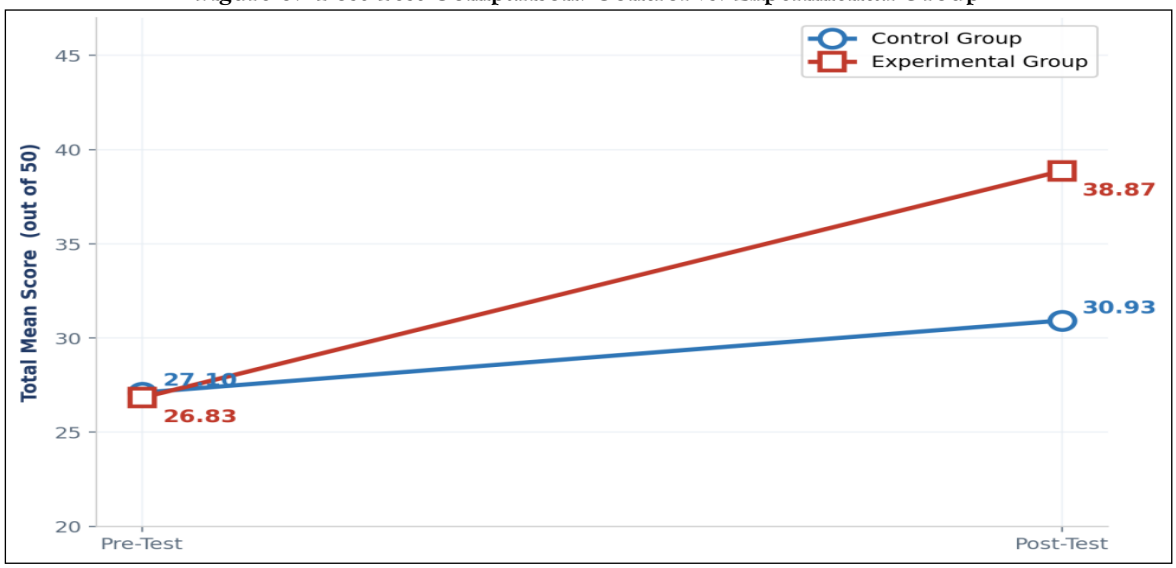

Figure 7. Total Score Trend: Pre-Test to Post-Test — Both Groups

4.5 Student Questionnaire Results (H3)

All four section means came in above 3.50 on the 0-5 scale, meaning students consistently responded toward the positive end across every domain the questionnaire covered. Section A (Awareness, M = 3.75) showed that teacher-mediated introduction was the most powerful route to tool awareness — unsurprising given that most students had no prior familiarity with AI language applications. Section B (Usage Frequency, M = 3.90) confirmed that students were not just aware of the tools but actively using them — in class, for homework, and during independent practice at home. Section C (Perceived Improvement, M = 3.93) showed that students felt the tools were actually helping: they reported greater confidence in speaking, faster vocabulary growth, and a better grasp of grammar. Section D (Motivation, M = 3.95) produced the

study's highest scores and most consistent pattern — students reported that AI-supported practice made them want to keep going, that finishing tasks gave them a sense of pride, and that they felt more confident about English overall. These results support H3.

Table 9. Student Questionnaire: Section-Level Results (Experimental Group)

Section	Domain	Items	Mean Range	Section Mean	Description
A	Awareness of AI Tools	5	3.53-3.97	3.75	Students have solid, consistent awareness of the tools
B	Frequency of AI Tool Usage	5	3.83-3.97	3.90	Students engaged frequently and across multiple contexts
C	Perceived Skill Improvement	5	3.87-3.97	3.93	Strong belief that the tools helped them improve
D	Motivation and Confidence	5	3.90-3.97	3.95	Very high and consistent motivation and personal pride

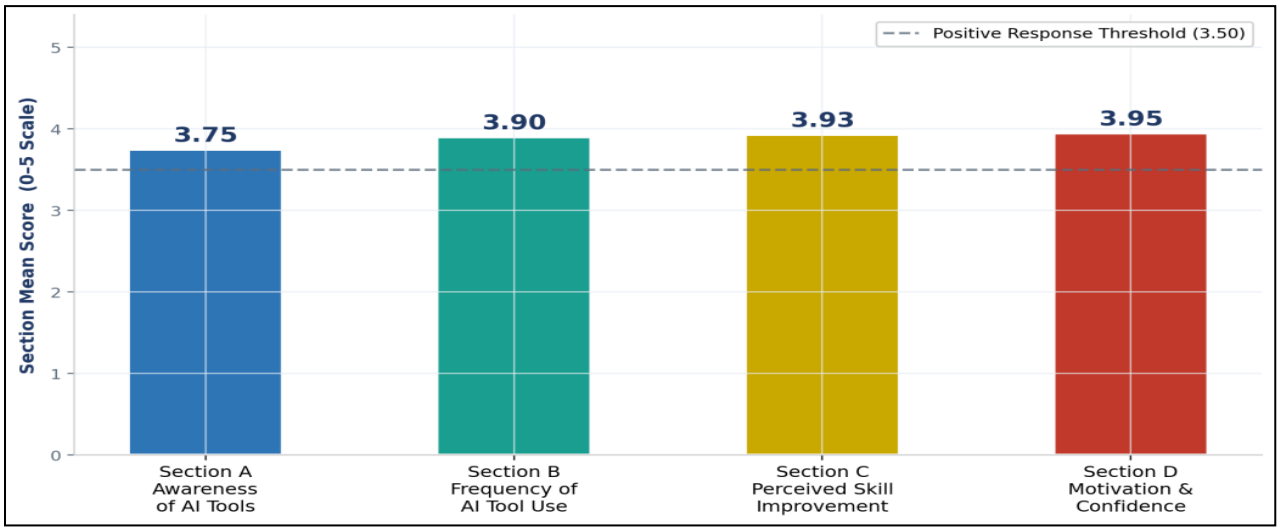

Figure 8. Student Questionnaire: Section-Level Mean Scores (Experimental Group)

4.6 Teacher Questionnaire Results

Forty teachers completed the questionnaire. The highest scoring item — the statement that AI will shape the future of language education (M = 4.13) — reflects a profession that is genuinely enthusiastic about where this is going. Other high-scoring items reinforced this: teachers strongly believed AI lessons make English more interactive (M = 4.10) and increase student engagement (M = 4.10). But when the questionnaire shifted from aspirations to operational realities, the scores dropped. Familiarity with ELSA Speak, Duolingo, and Grammarly specifically (M = 3.30) and regular classroom use of AI tools (M = 3.30) were the lowest-scoring items in the set. This is not a finding that should be read as pessimistic — it is precisely the kind of actionable gap that points directly at what needs to happen next.

Table 10. Teacher Questionnaire: Selected Likert-Scale Means (n = 40)

Statement	Male M	Female M	Total M	Description
AI will shape the future of language education	3.95	4.33	4.13	Strong agreement
AI-based lessons increase student engagement	3.91	4.33	4.10	Strong agreement

AI makes English learning more interactive	3.86	4.39	4.10	Strong agreement
AI activities complement my teaching methods	3.73	4.22	3.95	High agreement
Students gain confidence through AI practice	3.73	4.11	3.90	High agreement
Aware how AI can support English learning	3.68	3.94	3.80	High agreement
Understands how AI provides real feedback	3.64	3.83	3.72	Understanding
Encourages students to use AI apps for practice	3.59	3.56	3.57	Consistent
Confident explaining what AI means	3.32	3.83	3.55	Moderate-high confidence
Stays updated on educational AI development	3.09	3.67	3.35	Moderate engagement
Uses AI tools regularly in English lessons	3.55	3.00	3.30	Moderate regular
ELSA Speak, Duolingo, and Grammarly	3.18	3.44	3.30	Moderate familiarity

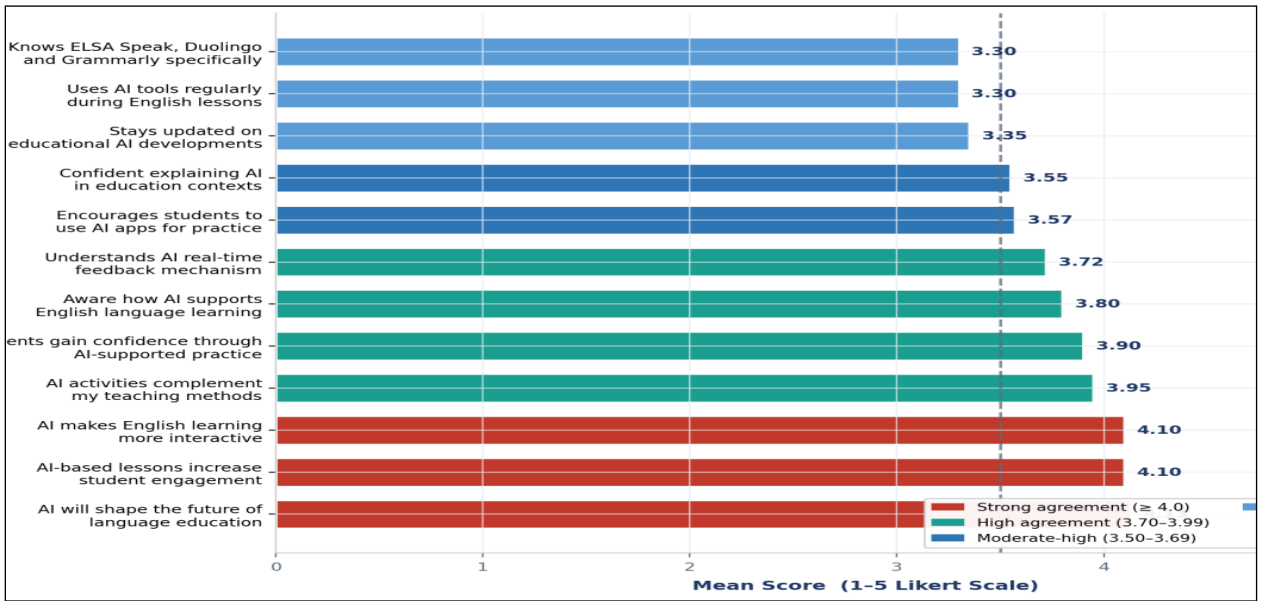

Figure 9. Teacher Questionnaire: Perception Mean Scores by Statement

Thematic analysis of teacher open-ended responses produced seven themes (Table 11). Teachers wrote enthusiastically about how AI tools opened up more interactive and student-centered lesson formats, and several commented that they could see individual students who normally struggled finding confidence in AI-supported tasks. The motivational effect was noted repeatedly. The barriers were equally candid: internet access was unreliable at both schools, device availability was a constant constraint, and several teachers noted that they did not feel equipped to troubleshoot technical problems or confidently explain to students what the tools were doing. Every single teacher who mentioned professional development described it as inadequate or absent. The appetite for training was high — what was missing was the provision of it.

Table 11. Emergent Themes from Teacher Open-Ended Responses

Theme	Representative Codes	Summary
Engaging Instruction	Interactive delivery; student-centred tasks; clearer explanations	AI created more participatory lesson formats that held student attention.
Personalised Learning	Individual feedback; adaptive difficulty; home practice extension	Students could learn at their own pace in ways classroom instruction cannot support.
Language Skill	Vocabulary growth; pronunciation	Teachers observed and reported skill gains

Improvement	accuracy; grammar correction	linked to AI feedback cycles.
Student Motivation	Increased participation; pride in completion; growing confidence	AI tasks raised engagement; some initial guidance was needed from teachers.
Implementation Challenges	Unreliable internet; device shortages; time pressure	Infrastructure gaps created persistent practical barriers to consistent use.
Limitations and Concerns	AI accuracy variability; risk of over-dependence; reduced critical thinking	Teachers flagged need for balance and continued teacher oversight.
Need for Professional Training	Workshops; hands-on practice; sustained development	Comprehensive training needed before full classroom integration is viable.

5. Discussion

The results tell a consistent story, and it is worth pausing to notice just how consistent it is. The experimental group outperformed the control group on five of six individual sub-skills at post-test, and the total score difference i.e. a gain of 12.04 points against 3.83 is not a marginal statistical artefact. It is a substantial difference in learning outcomes over ten weeks. The theoretical explanation is also coherent. Every element of what ELSA Speak, Duolingo, and Grammarly do in practice maps onto the learning mechanisms that Vygotsky's (1978) ZPD, Piaget's (1970) constructivism, and the TPACK framework (Mishra & Koehler, 2006) predict will produce development. The tools are not effective by accident. The pronunciation and fluency gains are the finding that perhaps most clearly demonstrates what AI can do that a teacher genuinely cannot. The experimental group improved on the ELSA Speak component by more than three times the control group's gain. This is phoneme-level feedback delivered individually, in real time, for every student in the room simultaneously. Cucchiarini et al. (2009) documented comparable gains in their computer-assisted pronunciation training work, and Neri et al. (2002) showed the same pattern in a controlled CAPT study. What this study adds is evidence that the effect holds in a very different setting — an under-resourced government school in Pakistan, with students who have never encountered this kind of tool before.

The writing results are, if anything, even more striking. Creative writing gains in the experimental group were more than three times those of the control group. The mechanism here is straightforward once you think about it from the learner's point of view. In a conventional lesson, a student writes a paragraph, submits it, and gets back a corrected version. The correction is informative but passive — the student reads it, nods, and moves on. With Grammarly in the loop, the process is different: each error is flagged with an explanation, and the student has to read that explanation and decide what to do. That requires an active cognitive response to every problem in the writing, not just passive reception of a corrected page. Koltovskaia (2020) found the same pattern in her work on student engagement with Grammarly feedback, and Ranalli et al. (2017) documented similar effects in automated writing evaluation more broadly.

Findings indicate the correlation finding — r = 0.71, R squared = .49 — adds an important practical layer to the performance results. Students who used the tools more got more out of them, which is not surprising, but the strength of the relationship underlines that access to AI tools is not the same as benefiting from AI tools. Simply loading Duolingo onto a school device achieves nothing if students do not engage with it regularly and purposefully. Any serious implementation plan must include structures that promote sustained usage: homework requirements, in-class practice schedules, teacher monitoring of engagement data.

Data demonstrates that teacher perceptions are genuinely encouraging in some respects and genuinely concerning in others. The enthusiasm is real — a mean of 4.13 on the item about AI shaping the future of language education is not lukewarm endorsement, it is strong professional conviction. But specific tool familiarity sitting at 3.30; while aspirational enthusiasm sits at 4.13; is precisely the TPACK gap that Mishra and Koehler (2006)

predicted. Teachers who know that technology integration is important and believe it would help, but who do not yet have the specific technological-pedagogical knowledge to make it work in practice. Nazir and Khalid (2022) found the same infrastructure barriers in their earlier Pakistani study, and this study confirms that those barriers have not been systematically addressed in the intervening years.

6. Implications of the study

For teachers who want to start using these tools now, the evidence suggests a few practical principles. ELSA Speak works best when students are asked to repeat a task multiple times, review their accuracy scores, and identify which specific sounds they need to work on — the metacognitive loop is part of what makes it effective, not just the corrective feedback itself. Duolingo is most useful as a homework and extension task, combined with in-class communicative activities that let the teacher build on the practice students have already done. Grammarly belongs in the writing process, not at the end of it — students should be required to acknowledge and apply each correction before a piece of writing is considered finished. For administrators, the practical priority is ensuring that the practice opportunity these tools create is not available only to students with home internet access. Device-sharing arrangements, library access, and loan schemes all deserve consideration.

This study offers the first empirical test of ZPD-mediated AI scaffolding in a Grade 8 EFL context within Pakistan's government school system. The results confirm that the scaffolding mechanism predicted by Vygotsky's (1978) framework operates in a socioeconomic and infrastructural setting that is substantially different from the affluent East Asian and Western university environments in which it has previously been studied. TAM (Davis, 1989) also receives support: perceived usefulness and perceived ease of use predicted engagement depth and perceived gain in a population that the model had not previously been tested with at this scale. And the TPACK gap identified by Mishra and Koehler (2006) is not just theoretically hypothesized here — it is empirically measured, in a specific context, with specific numbers attached to it.

7. Limitations of the study

Every study has boundaries, and being honest about them is part of making the findings credible. The most significant limitation here is scope: sixty students from two schools in one district cannot support claims about AI-enhanced EFL instruction across Pakistan as a whole. Purposive sampling of intact classes, while appropriate for the research questions, introduces the possibility of school-level differences that individual random assignment would have controlled for. The ten-week window is also too short to distinguish genuine skill acquisition from a novelty effect — students may have been more motivated simply because the tools were new and interesting, and it is not clear whether that effect would persist once the novelty wore off. Self-reported questionnaire data carry the usual social desirability risks, and ELSA Speak's speech recognition engine, trained primarily on American English phonology, may not have been fully accurate for Punjabi-accented speakers, which could have led to an underestimate of the actual pronunciation gains. Finally, the study measured only speaking and writing, leaving listening and reading — both of which the AI tools may also influence — entirely unexamined.

8. Conclusion

The question this study set out to answer was a practical one: can three AI-based language tools — ELSA Speak, Duolingo, and Grammarly — produce measurably better EFL outcomes for Grade 8 students in Pakistani government schools than conventional teaching alone? The answer, across one hundred participants, three tools, six sub-skills, and dual questionnaire data, is yes, and the evidence supporting that answer is consistent and substantial. AI-assisted students made significantly greater speaking and writing gains than their conventionally taught peers, they reported using the tools frequently and with genuine enthusiasm, they perceived clear improvements in their own skills, and their motivation was high throughout. Teachers were broadly supportive of the direction of travel, while being honest about the

infrastructure and professional development gaps that currently limit what is possible in practice. What this adds up to is a clear case for a hybrid instructional model in which AI tools handle the individualized, real-time formative feedback that large-class conventional instruction structurally cannot provide, while teachers remain responsible for the relational, cultural, and intellectually demanding dimensions of language learning that no automated system can replicate. Getting this model into government secondary schools across Pakistan — not just in the two schools in this study but equitably across urban and rural settings alike — will require investments in digital infrastructure and teacher professional development that go well beyond what individual teachers or school administrations can manage alone. The evidence is there to make the case for those investments. What comes next depends on whether policymakers choose to act on it.

9. Recommendations of the study

The most pressing follow-up question is straightforward: do the gains last? A longitudinal study tracking the same cohort through Grade 9 would reveal whether the skills acquired during the AI-integrated intervention consolidate over time or fade once the tools are no longer being used as intensively. A more rigorous future design would use a three-arm comparison — AI-only, teacher-only, and hybrid — with random assignment across a larger multi-district sample, which would provide much stronger causal evidence than any quasi-experimental design can offer. Rural-urban comparative work would directly address the equity question: do infrastructure disparities between urban and rural government schools in Punjab produce systematically different AI integration outcomes? Future instrumentation should expand the creative writing rubric to address the ceiling effect observed here, and should include direct observation protocols rather than relying entirely on self-report for usage data. At the more theoretical end, research that mines AI platform interaction logs — tracking each student's error frequency, feedback uptake rate, and revision cycles — would provide direct empirical evidence for the ZPD scaffolding mechanism at the individual learner level rather than just the group aggregate.

REFERENCES

Braun, V., & Clarke, V. (2006). Using thematic analysis in psychology. Qualitative Research in Psychology, 3(2), 77-101. https://doi.org/10.1191/1478088706qp063oa

Bygate, M. (2016). Sources, developments and directions of task-based language teaching. The Language Learning Journal, 44(4), 381-400. https://doi.org/10.1080/09571736.2015.1039566

Chen, X., Zou, D., Cheng, G., & Xie, H. (2020). Detecting latent topics and trends in educational technologies over four decades using structural topic modelling: A retrospective study. British Journal of Educational Technology, 51(3), 692-717. https://doi.org/10.1111/bjet.12907

Cohen, L., Manion, L., & Morrison, K. (2018). Research methods in education (8th ed.). Routledge.

Creswell, J. W., & Plano Clark, V. L. (2018). Designing and conducting mixed methods research (3rd ed.). Sage Publications.

Cucchiarini, C., Neri, A., & Strik, H. (2009). Oral practice with a CAPT system: How much does it help for long-term pronunciation improvement? Computer Speech & Language, 23(1), 60-80. https://doi.org/10.1016/j.csl.2007.12.003 *Replaces Bai & Hu (2022)

Davis, F. D. (1989). Perceived usefulness, perceived ease of use, and user acceptance of information technology. MIS Quarterly, 13(3), 319-340. https://doi.org/10.2307/249008

Denzin, N. K. (1978). The research act: A theoretical introduction to sociological methods (2nd ed.). McGraw-Hill.

Field, A. (2013). Discovering statistics using IBM SPSS statistics (4th ed.). Sage Publications.

Fryer, L., & Carpenter, R. (2006). Bots as language learning tools. Language Learning & Technology, 10(3), 8-14. https://doi.org/10125/44068 *Replaces Rahimi & Zarei (2021) — not found in LL&T archive

Godwin-Jones, R. (2017). Smartphones and language learning. Language Learning & Technology, 21(2), 3-17. https://doi.org/10125/44607

Hwang, G.-J., Xie, H., Wah, B. W., & Gasevic, D. (2021). Vision, challenges, roles, and research issues of artificial intelligence in education. Computers and Education: Artificial Intelligence, 1, 100001. https://doi.org/10.1016/j.caeai.2020.100001

Koltovskaia, S. (2020). Student engagement with automated written corrective feedback (AWF) provided by Grammarly: A multiple case study. System, 91, 102247. https://doi.org/10.1016/j.system.2020.102247

Mishra, P., & Koehler, M. J. (2006). Technological pedagogical content knowledge: A framework for teacher knowledge. Teachers College Record, 108(6), 1017-1054. https://doi.org/10.1111/j.1467-9620.2006.00684.x

Neri, A., Cucchiarini, C., Strik, H., & Boves, L. (2002). The pedagogy-technology interface in computer-assisted pronunciation training. Computer Assisted Language Learning, 15(5), 441-467. https://doi.org/10.1076/call.15.5.441.13473

Piaget, J. (1970). Science of education and the psychology of the child. Orion Press.

Ranalli, J., Link, S., & Chukharev-Hudilainen, E. (2017). Automated writing evaluation for formative assessment of second language writing: Investigating the accuracy and usefulness of feedback. Educational Psychology, 37(1), 8-25. https://doi.org/10.1080/01443410.2015.1136407

Shamim, F. (2011). English as the language for development in Pakistan: Issues, challenges, and possible solutions. In H. Coleman (Ed.), Dreams and realities: Developing countries and the English language (pp. 291-310). British Council.

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

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