ISSN 2411-7390

JOURNAL OF LANGUAGE & EDUCATION

Volume 11 Issue 2, 2025

HIGHER SCHOOL OF ECONOMICS
NATIONAL RESEARCH UNIVERSITY

HIGHER SCHOOL OF ECONOMICS
NATIONAL RESEARCH UNIVERSITY

EDITOR-IN-CHIEF

EDITORIAL BOARD

EDITORIAL TEAM

CONTENT

EDITORIAL

RESEARCH PAPERS

REVIEW PAPERS

BOOK REVIEWS

https://doi.org/10.17323/jle.2025. 27387

Enhancing Critical Thinking Skills in ChatGPT-Human Interaction: A Scoping Review

Lilia Raitskaya [1], **Elena Tikhonova** [2, 3]

[1] MGIMO University, Moscow, Russia
[2] HSE University, Moscow, Russia
[3] RUDN University, Moscow, Russia

ABSTRACT

Introduction: The rapid integration of generative artificial intelligence (GenAI) technologies, including ChatGPT, into educational environments has introduced both opportunities and challenges for learners and educators. While GenAI can support advanced learning practices, it also raises concerns about critical engagement and the accuracy of generated content. Previous systematic reviews have explored GenAI's relationship with critical thinking (CT) and self-regulated learning, but a focused synthesis of recent empirical evidence on GenAI's impact on university students' CT skills remains lacking.

Method: This scoping review followed the PRISMA-ScR guidelines and applied the Arksey and O'Malley framework alongside the Population – Concept – Context (PCC) model. Studies were identified via the Scopus database, using inclusion criteria limited to the years 2024–2025, English language, and the Social Sciences subject area. Thirty eligible empirical studies were analysed and visualised using VOSviewer to identify thematic clusters and categories in the literature.

Results: The reviewed studies were grouped into seventeen thematic clusters by the VOSviewer and then manually synthesized into six categories based on semantic interpretation: cognitive and metacognitive development, pedagogical innovation and learning design, academic writing and language learning, AI literacy and learner perception, evaluation and assessment technologies, global and ethical dimensions of GenAI use. The findings were analysed as (1) direct enhancement of CT, (2) metacognitive and reflective gains, (3) contextual factors shaping CT, (4) risks of cognitive offloading, and (5) instructional strategies mediating AI's effect. 21 publications showed predominantly positive impact of GenAI on CT (idea generation, conceptual understanding, construction of arguments, literature review, academic writing, etc.) whereas reported found mixed impact.

Conclusion: The review concludes that GenAI holds substantial potential to support CT development, particularly when pedagogically integrated to promote active reasoning, metacognitive monitoring, and critical autonomy. However, the evidence base is still emerging and is limited by its short temporal scope, narrow database coverage, and reliance on self-reported data. Future research should focus on long-term effects, discipline-specific instructional models, and robust theoretical frameworks linking AI use to cognitive outcomes.

KEYWORDS

development of critical thinking skills, generative AI impact on higher education, ChatGPT and student cognition, AI-supported academic writing, metacognitive engagement with GenAI, GenAI and argumentation, prompt engineering and critical thinking

Citation: Raitskaya, L., & Tikhonova, E. (2025). Enhancing critical thinking skills in chatgpt-human interaction: A scoping review. *Journal of Language and Education, 11*(2), 5-19. https://doi.org/10.17323/jle.2025.27387

Correspondence:
Elena Tikhonova,
etihonova@hse.ru

Received: May 23, 2025
Accepted: June 10, 2025
Published: June 30, 2025

INTRODUCTION

The recent advancements in generative artificial intelligence (GenAI), particularly the release of ChatGPT 3.5, have catalyzed a surge of interest in the educational potential and cognitive implications of human-AI interaction (Fabio et al., 2025). As educational institutions across the globe begin integrating GenAI technologies into teaching and learning processes, researchers have turned to examining how these tools shape students' skills and learning behaviours. Among these

skills, critical thinking (CT) has received considerable attention, as it is widely regarded as both a prerequisite for and an outcome of meaningful engagement with AI (Tikhonova & Raitskaya, 2023; Jiang et al., 2024).

Critical thinking plays a central role in the digital age, not only as a measure of academic development but also as a safeguard against the epistemic risks posed by algorithmically generated content. When students engage in tasks such as refining prompts, interpreting AI-generated responses, verifying information accuracy, or reflecting on the ethical implications of machine-produced language, they are actively employing critical thinking strategies (Babin et al., 2024; Gonsalves, 2024). As Darwin et al. (2024) note, CT becomes a mediating function in student-AI interaction, supporting both metacognitive awareness and epistemological vigilance.

While the concept of critical thinking is well-established, with over 36,000 documents indexed in Scopus as of May 2025 and roots stretching back to the early 20th century, the arrival of GenAI technologies has redefined its operationalization in education. Since the 1990s, interest in CT has grown steadily, with significant acceleration in the 2000s. The foundational framework for CT in educational research remains Bloom's taxonomy and its revisions, which distinguish between levels of cognitive complexity, from basic recall to synthesis and creative problem-solving (Krathwohl, 2002).

The proliferation of GenAI tools challenges educators and researchers to reconsider how critical thinking is taught, assessed, and applied. On one hand, GenAI offers powerful means of enhancing CT through interactive, scaffolded, and personalized learning environments. On the other, there are growing concerns about cognitive offloading, information naivety, and uncritical dependence on AI systems (Risko & Gilbert, 2016; Gerlich, 2025). These tensions underscore the need for a more precise understanding of how GenAI influences different dimensions of critical thinking, especially in higher education contexts where autonomy, reflection, and analytical skills are paramount.

To date, only a limited number of systematic reviews have addressed the intersection of GenAI and critical thinking in higher education. Notably, Melisa et al. (2025) explored how ChatGPT use correlates with students' ability to evaluate information and develop independent judgments, while Sardi et al. (2025) focused on the interplay between CT and self-regulated learning in AI-supported environments. Both reviews emphasized generally positive or neutral outcomes, though they also pointed to a lack of consensus regarding the mechanisms through which GenAI supports or inhibits higher-order thinking.

This scoping review builds on that emerging body of scholarship by focusing specifically on empirical studies published between 2024 and 2025, a period that coincides with a marked expansion in the use of GenAI in academic settings. By concentrating on recent peer-reviewed research indexed in the Scopus database, the review aims to synthesize current knowledge on how generative AI influences the development of critical thinking skills among university students. Unlike earlier reviews that cast a broad net, the present study narrows its focus to uncover the patterns, strategies, and pedagogical conditions under which CT is most effectively cultivated through GenAI interaction.

Accordingly, this review addresses the following research questions:

RQ1. What are the key directions of empirical research on generative AI impact on critical thinking?
RQ2. How are critical thinking skills influenced in human-ChatGPT interaction?

By answering these questions, the review seeks to provide educators, policymakers, and researchers with a more nuanced understanding of the opportunities and limitations associated with GenAI in higher education. It also aims to identify promising avenues for further inquiry and to support the development of evidence-based pedagogical frameworks that foreground critical thinking in the age of artificial intelligence.

METHOD

Protocol

Getting down to the present scoping review, we meticulously developed a research protocol. The reviewers hereby certify that this scoping review report constitutes a faithful, precise, and transparent description of the conducted review. No deviations from the protocol were registered. Any departures from the original study design have been duly elucidated. This scoping review stick to the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) extension for Scoping Reviews (Tricco et al., 2018) and the framework proposed by Arksey and O'Malley (2005).

Eligibility Criteria

In this review, the population, concept, and context (PCC) framework was applied to devise an effective search strategy where each criterion was justified (Table 1).

Information Sources and Search Strategies

The Scopus database was searched on May 25, 2025. The search string used was: TITLE-ABS-KEY("critical thinking") AND TITLE-ABS-KEY("ChatGPT" OR "generative AI" OR "generative artificial intelligence"). Only documents published in

Table 1
Eligibility Criteria

Criterion	Inclusion	Exclusion	Rationale
Population	Students, EFL learners and teachers in the higher education institutions	Other users of GenAI and population at other educational levels	The introduction of generative AI into higher education requires research of the challenges and barriers that may be overcome via fostering a complex of competencies, including higher order thinkings skills at that educational level
Concept	Critical thinking, including Generative AI impact on critical thinking	Other concepts	The aim of the review is to identify the scope and recent trends of research on critical thinking in the context of appliances of GenAI
Context	Higher education	Other contexts	The review dwells upon studies in the higher education environment
Language	English	Other languages	The object of all research in focus is scholarly publications in English. The language choice is also identified by its status as a lingua franca of international science.
Time period	2024-2025	Publications before 2024	The period is selected due to the breakthrough in the generative AI, starting from ChatGPT 3.5. Only the recent publications were considered
Types of sources	In the Scopus database: full texts of articles and book chapters	Unavailable sources, unavailable full texts	This review aims to get a comprehensive understanding of the field
Geographical location	Any location	N/A	Getting international perspective
Database	Scopus	Other bases	Scopus was selected as a comprehensive database indexing top-ranking publications on higher education and technological innovations in the society
Areas of Research	Social Sciences	Other areas	The research in Social Sciences were chosen as the context is limited to higher education

2024 or 2025 were included. Pre-protocol test searches were conducted to assess alternative keywords, but no additional relevant studies were identified.

Document Identification and Screening

Both reviewers independently identified documents (empirical articles and book chapters) subject to the eligibility criteria enumerated in Table 1. The searches base in the Scopus data brought 566 titles. After the Scopus filters (research field, language and types of sources) had been applied, the total decreased to 290. Each author independently screened the titles and abstracts of the 290 documents. Only empirical research papers were selected. Special focus was made on the context (higher education). 207 documents were subsequently eliminated after the screening. Full texts of 36 documents out of 83 were retrieved from the journal sites and the publishers' sites for further screening. Each full text was downloaded, thoroughly read and independently analysed by each author. After they had been thoroughly analysed, six documents were found irrelevant as they were not based on empirical research, or the research was placed in a school environment. When occasional disagreements arose, they were settled by mutual consent. Thirty documents were ultimately included into the review (Figure 1).

Data Charting

A data-charting form was cooperatively developed. Both authors independently charted the data extracted from five identified documents as a pilot calibration. The data-charting form was discussed in an iterative process. The ultimate data included in the form are enumerated in Table 2. All the extracted data were double-checked by the authors.

Data Analysis

Following the data charting process, we employed a combination of descriptive synthesis and thematic analysis to interpret the extracted information and answer the research questions. The analysis was structured in alignment with the review's two guiding questions: (1) the identification of key directions in empirical research on the impact of generative AI (GenAI) on critical thinking; and (2) the examination of how specific aspects of human - ChatGPT interaction influence the development of critical thinking skills among university students.

Initially, we conducted within-case analysis of each study, focusing on contextual variables such as research design, population, mode of GenAI use, and specific indicators of

Figure 1

Selection of Publications for the Review

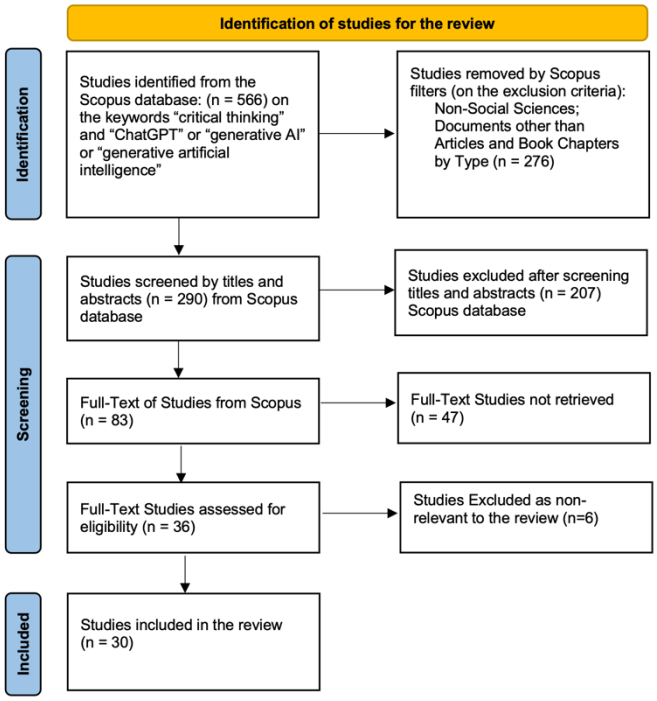


Table 2

Data-Charting Form

Data to be extracted	Population	Main findings related to CT
Title of study	Sample size	Type of GenAI impact
Author(s)	Data collection methods	Reported benefits of GenAI
Country	Main findings related to CT	Reported drawbacks
Type of publication	Type of GenAI impact	Limitations noted by authors
Study design	Sample size	Keywords
Context of GenAI use	Data collection methods	Discipline

critical thinking. This was followed by a cross-study comparison, during which patterns, convergences, and divergences across the publications were systematically coded. The coding was carried out independently by both reviewers and then iteratively refined through discussion to ensure analytical consistency.

Studies were grouped thematically based on the nature of GenAI application (e.g., academic writing support, feedback provision, argument development, or reflection tasks), and the type of critical thinking engagement reported (e.g., evaluation, reasoning, synthesis, or metacognitive awareness). These groupings were not predefined but emerged inductively from the data through repeated reading and clustering of findings. In addition, we paid close attention to how critical thinking was operationalised and measured across studies. Where applicable, we noted whether frameworks such as Bloom's taxonomy, self-regulated learning models, or other cognitive scaffolds were employed to define and assess critical thinking.

To support visual interpretation of trends, thematic clusters were validated using VOSviewer for co-occurrence analysis of keywords. The integration of keyword mapping allowed us to triangulate inductively identified trends with bibliometric signals from the literature set. The final synthesis was structured around the emergent thematic clusters, each supported by illustrative examples from the reviewed studies. These clusters served as the basis for answering the two research questions and outlining research directions in the concluding section of the review.

Data Validation

To ensure the trustworthiness and consistency of the review findings, we implemented a multi-step data validation process at both the extraction and synthesis stages. This process was designed to minimise bias, enhance inter-reviewer reliability, and support analytical transparency.

During the data charting phase, both reviewers independently extracted data from a randomly selected subset of five studies to establish a common interpretive framework and calibrate the data-charting form. Discrepancies were discussed and resolved through consensus, resulting in refinements to the extraction categories and clarification of ambiguous fields. Once alignment was reached, the remaining studies were divided equally between the reviewers, with a second round of cross-checking performed to confirm the consistency of coding and categorisation decisions.

To validate the thematic analysis, each emergent category or cluster was reviewed against the original full texts to confirm that the assigned labels accurately reflected the underlying content and empirical focus of the studies. Particular attention was paid to studies that overlapped thematically or contained findings relevant to multiple categories; these were flagged and jointly reviewed to ensure appropriate placement without duplication or over-interpretation. Additionally, to reduce the risk of interpretive bias, we employed a form of analyst triangulation: each reviewer independently interpreted the findings within each thematic group and then collaboratively compared interpretations. This process helped ensure that the identified trends and conclusions were not the result of individual analytical preferences but rather emerged through iterative dialogue and critical discussion.

Finally, keyword co-occurrence analysis conducted via VOSviewer served as a supplementary validation tool, offering a bibliometric lens to reinforce or question patterns observed in the thematic synthesis. Convergences between manually identified clusters and automatically generated keyword networks strengthened the internal coherence of the review's analytical framework.

RESULTS

Search and Selection Results

The searches in the Scopus database were made as of April 24, 2025. A total of 566 documents were found on a combination of the previously identified keywords: "critical thinking" AND "ChatGPT" OR "generative AI" OR "generative artificial intelligence".

After the Scopus filters, including the eligibility criteria of language, types of publications, and research field had been applied, the total decreased to 290. The authors screened the titles and abstracts of 290 documents. 207 documents were eliminated after the screening. Full texts of 36 documents out of 83 were retrieved for further screening. After they had been thoroughly analysed, six documents were found irrelevant as they were not based on empirical research, or the research was placed in a school environment. Thirty documents were ultimately included into the review (Figure 2).

A Bibliometric Analysis

The 30 documents ultimately included in the review were analysed. The publications indexed in the Scopus database included 29 research articles and one book chapter. The review entailed 19 and 11 documents published in 2024 and in 2025 respectively. Three publications appeared in *Thinking Skills and Creativity*. Two articles were published in each of the following four journals: *Cogent Education, Education and Information Technologies, Frontiers in Education, and International Journal of Educational Technology in Higher Education.* The remaining 18 articles came out in another 18 journals. The sampling entailed one book chapter. It was part of the book entitled "Studies in Systems, Decision and Control" (Emran et al., 2024).

Geographically, the reviewed documents were distributed as follows: China, Indonesia, the United Kingdom, and Taiwan accounted for four publications each. Two documents came from Educator and Malaysia each. The other countries included Italy (2 documents), Spain (2 documents), Switzerland (2 documents), and Bahrain (1 document).

The thirty publications were authored by 91 researchers, nearly 3 authors per record on average. The authors had 63 affiliations, with King's College London, Università degli Studi di Messina, National Yunlin University of Science and Technology, and National Cheng Kung University as the frontrunners (two authors from each). The other 59 affiliations were represented by one researcher each.

Included Studies

After identifying and screening the relevant documents, the ultimate 30 documents were included into the review. The total population in the reviewed documents randomly, purposively or otherwise selected for 29 out of the 30 studies amounted to 4785 participants, including university students, EFL learners, and university teachers. One of the studies that did not report population dwells upon the transformative impact of AI in educational settings, focusing on the necessity for AI literacy, prompt engineering proficiency, and enhanced critical thinking skills (Walter, 2024).

The data on the population and sample size are included in Appendix 1.

The study designs entail qualitative, quantitative, and mixed. Though, in describing study designs, we focused on the authors' wordings of the type of design in the reviewed documents. In Table 3, we also outlined quasi-experimental design, randomised controlled tests, and other designs, covering "a naturalistic inquiry approach" (Gonsalves, 2024), "a cross-sectional survey design" (Lijie et al., 2025), "an exploratory study" (Panit, 2024), "a case-study" (Walter, 2024), and "a true-experimental design" (Xu & Liu, 2025).

Thematic Clusters

The co-occurrence analysis of author keywords from the 30 included publications was conducted using VOSviewer, based on full counting and minimum threshold set to two occurrences per keyword. The resulting visualization (Figure 2) reveals 17 initial clusters, automatically generated via VOSviewer's clustering algorithm based on keyword prox-

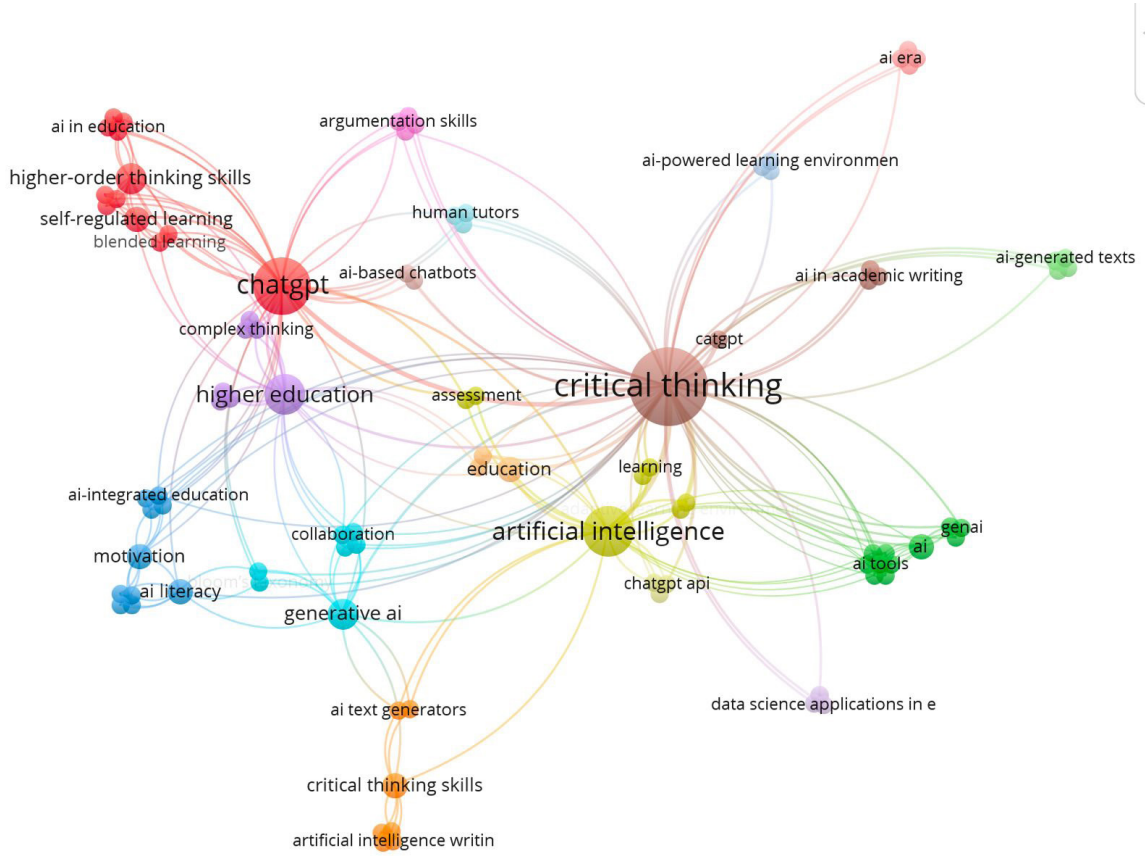
Table 3

Study Designs in the Reviewed Documents

Study design	Documents in the review
Qualitative design	Darwin et al., 2024; Octaberlina et al., 2024; Santamaria-Velasco et al., 2025
Quantitative design	Fadillah et al., 2024; Naatonis et al., 2024; Zhou et al., 2024
Mixed design	Essien et al., 2024; Fakour & Imani, 2025; Gerlich, 2025; Hwang et al., 2025;
Quasi-experimental design	Emran et al., 2024; George-Reyes et al., 2024; Liu & Tu, 2024;
Randomized controlled trial	Lee et al., 2024; Wang et al., 2025
Other designs	Boers et al., 2025; de la Puente ae al., 2024; Fabio et al., 2025; Chaparro-Banegas et al., 2024; Gonsalves, 2024; Lijie et al., 2025; Oates & Johnson, 2025; Panit, 2024; Shen & Teng, 2024; Suriano et al., 2025; Tang et al., 2024; Walter, 2024; Xiao et al., 2025; Xu & Liu, 2025; Yasuf et al., 2024

Figure 2

VOSviewer Visualization of Keyword Co-occurrence and Thematic Clusters



Note. Circle size indicates keyword frequency; spatial distance reflects co-occurrence strength; colors represent cluster groupings by VOSviewer's modularity-based algorithm. Thematic categories were manually synthesized based on semantic interpretation.

imity and frequency. However, due to thematic overlaps and conceptual affinities among clusters, we consolidated them into six broader thematic categories, described below.

Cognitive and Metacognitive Development (Clusters 1, 2, 5, 6, 13, 17)

This group includes studies exploring how GenAI tools support critical thinking, self-regulation, cognitive complexity, dual-process reasoning, and reflective thinking. Bloom's taxonomy, problem-based gamification learning (PBGL), and dual-process theory frequently appear in this set. These studies investigate how AI interaction fosters higher-order thinking skills (HOTS), metacognitive strategies, and deep learning.

Pedagogical Innovation and Learning Design (Clusters 4, 12, 15)

This thematic area focuses on instructional models integrating GenAI, such as guidance-based ChatGPT scaffolding, dialogic learning, and AI-mediated tutoring. Emphasis is placed on human - AI collaboration, Socratic methods, and blended feedback strategies that enhance reasoning, argumentation, and learner autonomy.

Academic Writing and Language Learning (Clusters 7, 8, 10)

Here, the focus is on academic discourse and EFL learners, examining how ChatGPT and other GenAI tools support writing pedagogy, improve argumentative skills, and challenge originality. Concerns about overreliance and intellectual autonomy are also raised, especially for students in postgraduate and EFL contexts.

AI Literacy and Learner Perception (Clusters 3, 9, 11)

This category examines students' and educators' perceptions of GenAI tools, their motivation, trust, and AI-literacy levels. Key concepts include attitudes toward AI, ethical reasoning, critical assessment of AI-generated content**,** and equity of access. A sub-focus is placed on learners' engagement across geographic and socio-economic contexts.

Evaluation and Assessment Technologies (Clusters 14, 16)

Studies in this group explore automated feedback systems, peer review tools, and educational data mining techniques**.** Emphasis is on using GenAI to evaluate critical thinking through scalable models (e.g., BERT, LLM4HA), epistemic network analysis, and evaluation metrics for learning analytics.

Global and Ethical Dimensions of GenAI Use (Clusters 9, 16)

This group overlaps with AI-literacy research but uniquely addresses macro-level implications: global disparities, digital inequality, responsible innovation, and sustainable development goals (SDGs). These studies treat AI not only as a pedagogical tool but as a social phenomenon shaping educational ethics and access.

Taken together, the thematic clusters highlight not only research directions but also the wide spectrum of competencies being shaped through human - ChatGPT interaction. Studies grouped under cognitive and metacognitive clusters emphasize the development of critical reasoning, reflective thinking, and self-regulation - core components of higher-order thinking skills. Clusters focused on academic writing, argumentation, and AI literacy foreground academic and ethical competencies, particularly among EFL learners, with attention to intellectual autonomy and authorship.

The clusters related to pedagogical innovation and evaluation technologies address more instrumental and analytical competencies, such as data analysis, gamification, and feedback design. Finally, clusters engaging with sustainability, equity, and global challenges point to the formation of value-oriented and civic competencies, including ethical judgment, teamwork, and global citizenship.

This competency-oriented lens reinforces the idea that generative AI tools do more than support skill acquisition: they participate in reshaping the educational paradigm, where flexible, integrative, and critically aware thinking becomes central to student preparedness for uncertain, technology-rich futures.

Generative AI Influence on Critical Thinking in Human-ChatGPT Interaction

This section synthesizes the findings of 30 empirical studies addressing how generative AI tools, especially ChatGPT, influence students' critical thinking (CT) skills in higher education. To respond to RQ2, the studies were analysed thematically and categorised into five analytical groups: (1) direct enhancement of CT, (2) metacognitive and reflective gains, (3) contextual factors shaping CT, (4) risks of cognitive offloading, and (5) instructional strategies mediating AI's effect. A structured summary of the reported benefits and drawbacks of AI impact on CT is provided in Appendix 1. 21 publications reported positive influence whereas the remaining nine showed mixed influence.

Direct Enhancement of CT Through AI-Supported Learning Tasks

A growing body of empirical studies has shown that generative AI enhances critical thinking (CT) when integrated into authentic, cognitively demanding tasks such as debate, argumentation, and analytical writing. These interventions appear most effective when AI tools are positioned not as information providers but as reasoning partners. For instance, de la Puente et al. (2024) demonstrated that incorporating

ChatGPT into structured debate sessions significantly improved both argumentation and CT skills, as confirmed by structural equation modelling. Emran et al. (2024) similarly found that students used ChatGPT to explore diverse perspectives and to practise logic-based evaluation, which in turn stimulated fact-checking behaviour and deeper cognitive engagement. Fabio et al. (2025) reported statistically significant gains across multiple CT dimensions (cognitive complexity, reasoning style, and openness) particularly among students who approached ChatGPT with caution and intent. Other studies support these findings across various disciplinary contexts: Santamaria-Velasco et al. (2025) linked ChatGPT use to improved evaluative judgement in historical and ethical analysis; Yusuf et al. (2024) showed that a scaffolded framework for AI-generated texts fostered synthesis and critique in academic writing; and Wang et al. (2025) observed CT improvement via AI-supported feedback mechanisms. Oates & Johnson (2025) further reported increases in students' critical evaluation scores during AI-assisted assessment tasks. Together, these findings highlight the potential of ChatGPT to reinforce CT when its use is grounded in well-designed learning scenarios that prioritize interpretation, justification, and reasoning.

Metacognitive and Reflective Engagement

Beyond direct skill acquisition, generative AI also plays a notable role in cultivating learners' metacognitive awareness and reflective thinking. This refers to students' ability to monitor, regulate, and evaluate their own cognitive processes in interaction with AI tools. Studies included in the review show that when learners are prompted not just to consume AI outputs, but to critically interrogate them, deeper layers of cognition are activated.

A useful conceptual distinction is offered between two complementary forms of critical engagement: one that focuses on the AI itself and another oriented toward the task at hand. Gonsalves (2024) refers to these as "critical thinking toward the AI", including practices such as refining prompts and evaluating potential biases, and "critical thinking for the assignment," which involves the application of AI-generated content to solve real academic problems. Such dual engagement fosters both reflective scepticism and problem-solving reasoning.

Reflective thinking is further supported when students become aware of AI's limitations. Darwin et al. (2024) reported that learners demonstrated increased awareness and caution when confronting the inability of AI systems to recognize irony or emotional nuance. Other studies revealed that this awareness becomes more acute when students encounter the misinformation produced by AI (plausible but incorrect outputs) which stimulate the need for verification and information control. Essien et al. (2024) and Emran et al. (2024) observed that these moments of AI failure led students to fact-check, triangulate sources, and engage more cautiously with content-behaviours closely tied to metacognitive regulation.

In addition to these spontaneous reflective responses, several pedagogical interventions were designed to scaffold metacognition intentionally. For example, Lee et al. (2024) introduced a guidance-based ChatGPT-assisted learning tool that provided indirect prompts rather than direct answers, thereby encouraging learners to articulate their own reasoning. Similarly, Hwang et al. (2025) demonstrated that a prompt-based learning model increased students' ability to generate questions and reflect on their learning paths. Both studies emphasize that structured interaction with AI (especially when mediated through carefully designed instruction) can enhance the quality and depth of students' reflective thought.

These findings suggest that generative AI, when embedded in learning environments that foreground interpretation and self-regulation, has the capacity to develop metacognitive dimensions of critical thinking, rather than merely facilitating content delivery.

Individual and Contextual Moderators of AI's Influence on Critical Thinking

While generative AI offers various opportunities to enhance critical thinking (CT), its actual impact is not uniform across learners or learning environments. Multiple studies suggest that individual learner characteristics and contextual factors substantially shape how AI tools affect cognitive engagement. Motivation, prior digital competence, and self-regulatory behaviours emerged as key variables influencing the effectiveness of AI-assisted learning. Lijie et al. (2025) demonstrated that motivation, AI literacy, and perceived usefulness of generative tools positively predicted students' disposition toward CT. However, ease of use was paradoxically associated with weaker CT development, likely due to reduced cognitive investment. This suggests that overly intuitive tools may inadvertently discourage deeper engagement. Similarly, Shen and Teng (2024) found that students with stronger self-directed learning (SDL) skills benefited more from AI-assisted writing tasks in terms of CT growth. SDL functioned as a moderator in the relationship between AI-supported writing and critical reasoning, indicating that autonomous learning behaviours amplify AI's educational potential.

Platform type also appeared less decisive than quality of engagement. Xu and Liu (2025) compared ChatGPT with Duolingo and found no significant differences in outcomes for CT and learner autonomy, implying that interactive depth, rather than platform architecture, drives cognitive gains. In a related vein, Suriano et al. (2025) showed that students' level of engagement and trust in ChatGPT were more powerful predictors of CT development than tool availability itself.

Other studies highlighted how design features and perceived reliability influence higher-order thinking skills (HOTS). Fadillah et al. (2024) emphasized that students' perceptions of convenience, responsiveness, and output accuracy significantly affected their cognitive engagement, with these factors positively correlating with HOTS development. Context-specific practices also matter: Chaparro-Banegas et al. (2024) in science education, and George-Reyes et al. (2024) in entrepreneurship training, reported notable CT improvements when AI was integrated into active, student-centred learning formats.

AI's educational impact is mediated not only by its functionality but also by learners' motivation, autonomy, and instructional context. These factors must be strategically addressed to unlock the full potential of AI-enhanced critical thinking (Table 4).

Risks of Over-Reliance and Cognitive Offloading

While many studies acknowledge the potential of generative AI to support critical thinking, growing evidence points to notable risks associated with its unmoderated use. A recurring concern across several studies is cognitive offloading - the tendency of students to delegate analytical tasks to AI systems instead of engaging with them independently. Gerlich (2025) identified a significant negative correlation between frequent use of AI tools and critical thinking performance, showing that reliance on automated outputs reduced students' cognitive effort and led to poorer outcomes. Octaberlina et al. (2024) and Panit (2024) reinforced this concern, noting that habitual dependence on AI tools can diminish active engagement and foster a passive learning stance.

More worryingly, this behavior was shown to compromise ethical judgment, particularly in academic writing, where the boundary between support and substitution becomes blurred. In several cases, students were found to accept AI-generated information without sufficient scrutiny, bypassing the deeper cognitive processes necessary for synthesis and evaluation (Panit, 2024). Gerlich (2025) further demonstrated that such patterns were associated with lower critical thinking scores and reduced long-term retention. These findings collectively suggest that, without careful scaffolding and pedagogical oversight, the use of GenAI may undermine the very cognitive capacities it is expected to cultivate.

Instructional Strategies Mediating the Impact of GenAI on Critical Thinking

A growing body of evidence suggests that the pedagogical design of GenAI integration plays a decisive role in shaping its influence on students' critical thinking (CT). Rather than functioning as a neutral tool, GenAI becomes pedagogically meaningful when embedded within instructional frameworks that prioritize active engagement, cognitive challenge, and reflective processing.

Table 4

Moderating Variables Influencing the Effect of GenAI on Critical Thinking Development

Moderator Type	Specific Factor	Effect on Critical Thinking (CT)	Representative Sources
Individual	Motivation (MO)	Strong positive predictor of critical thinking disposition	Lijie et al., 2025
	AI Literacy (AIL)	Promotes metacognitive awareness and more deliberate engagement with GenAI	Lijie et al., 2025
	Self-Directed Learning (SDL)	Enhances the impact of AI-assisted writing on CT development	Shen & Teng, 2024
	Trust and Engagement	High engagement levels correlate with greater CT improvement	Suriano et al., 2025
	Educational Attainment	Higher education levels associated with stronger CT outcomes	Gerlich, 2025
Technological	Perceived Ease of Use (PEOU)	Paradoxically lowers CT due to reduced cognitive effort	Lijie et al., 2025
	Output Reliability and Accuracy	Reinforces analytical behavior and user trust	Fadillah et al., 2024
	Platform Structure (e.g., ChatGPT vs Duolingo)	No significant difference if user engagement is comparable	Xu & Liu, 2025
Pedagogical	Active Learning Strategies (e.g., debates, inquiry tasks)	Amplify CT when integrated with GenAI-supported activities	Chaparro-Banegas et al., 2024; George-Reyes et al., 2024
	Interaction Format (e.g., scaffolding, feedback, mentorship)	Strengthens meaningful application of GenAI and metacognitive growth	Lee et al., 2024; Walter, 2024

One of the most promising instructional strategies involves prompt engineering[1]. As Walter (2024) argues, this practice not only requires students to think analytically about how they interact with GenAI but also fosters deeper understanding of the technology's limitations and ethical dimensions. His findings also emphasize the value of instructional scaffolding in AI literacy courses, where guided questioning, explicit feedback, and modeling of critical evaluation help students move from surface-level interaction to meaningful reflection and ethical reasoning.

Other studies underline the effectiveness of problem-based and gamified learning formats**.** For example, Naatonis et al. (2024) demonstrated that embedding ChatGPT into inquiry-driven modules enhanced students' CT by activating iterative feedback loops and engaging learners in self-directed problem-solving. Similarly, Xiao et al. (2025) evaluated the performance of large language models (LLMs), such as BERT and LLM4HA, in automatically assessing higher-order thinking skills (HOTS). Their results suggest that such tools can serve as efficient complements to human-led instruction by providing scalable, formative evaluation aligned with cognitive complexity.

Instructional strategies that promote self-regulation have been shown to significantly strengthen the impact of GenAI on critical thinking. Tools that prompt learners to plan, monitor, and reflect on their thinking processes encourage deeper engagement with complex problems and foster higher-order reasoning. When these self-regulatory mechanisms are embedded into digital learning environments, students not only solve tasks more effectively but also develop a stronger disposition toward critical inquiry. This approach was demonstrated to be particularly effective in a recent study involving AI-assisted learning design (Zhou et al., 2024). The findings are consistent with broader research emphasizing that autonomy-supportive environments enhance metacognitive awareness and sustain cognitive engagement over time (Ryan & Deci, 2017).

The reviewed evidence affirms that generative AI, particularly in the form of ChatGPT, can meaningfully support the development of critical thinking skills. This contribution is achieved not by substituting instructional efforts, but by enhancing them through intentional and well-structured pedagogical integration. When situated within guided, dialogic, or inquiry-based learning environments, generative AI serves as a catalyst for analytical reasoning, metacognitive development, and reflective judgment. Nevertheless, the effectiveness of such technologies depends substantially on the presence of appropriate instructional scaffolding. In its absence, learners are more likely to interact with AI tools in superficial or utilitarian ways, thereby limiting their transformative potential for education.

DISCUSSION

This scoping review synthesised empirical evidence on the influence of generative AI, particularly ChatGPT, on the development of critical thinking (CT) among university students. The overall trend across the included studies points to a predominantly positive effect, provided that AI use is pedagogically embedded and critically mediated. Key reported benefits include support for idea generation, conceptual understanding, construction of arguments, literature review, academic writing, and engagement in iterative reasoning processes (Fabio et al., 2025; Yusuf et al., 2024; Wang et al., 2025; Santamaria-Velasco et al., 2025). Students also benefited from time saved on routine academic tasks, which allowed greater cognitive focus on higher-order thinking. The reviewed literature consistently emphasised that engaging critically with AI through practices such as fact-checking, prompt refinement, and bias identification stimulates analytical reasoning and fosters critical thinking skills (Emran et al., 2024; Babin et al., 2024; Gonsalves, 2024).

Although benefits were prominent, several studies reported potential challenges. These include cognitive offloading, over-reliance on AI, superficial engagement with generated content, and diminished awareness of authorship or originality (Gerlich, 2025; Panit, 2024; Octaberlina et al., 2024). Darwin et al. (2024) warned against the formation of "echo chambers" and the inability of AI systems to grasp affective nuances. These risks were particularly salient when AI was used without adequate scaffolding or reflection, highlighting the need for structured instructional environments. Nevertheless, pedagogical frameworks such as scaffolded prompting, guided evaluation, and dialogic feedback demonstrated mitigating effects, enabling learners to maintain agency and reflective engagement (Lee et al., 2024; Walter, 2024; Hwang et al., 2025).

While the present findings generally align with the conclusions of earlier systematic reviews (Sardi et al., 2025; Melisa et al., 2025), which reported predominantly positive or neutral outcomes of GenAI use in education, this review extends prior research by offering a more granular synthesis of how specific instructional conditions, user dispositions, and AI design features interact to mediate the development of critical thinking. Unlike previous studies that focused on general outcomes, the current review maps out nuanced moderators and pedagogical mechanisms that shape the quality of human-GenAI interaction. Although Sardi et al. (2024) reported gains in higher-order CT skills and Erlich

[1] The deliberate crafting of inputs to elicit specific cognitive responses from AI systems.

(2025) identified improvements primarily at the lower levels of Bloom's taxonomy, the present review contributes a more differentiated perspective by systematically integrating these diverging outcomes. It further substantiates the ambivalence observed in broader research on AI's cognitive effects (Royer, 2024; Raitskaya & Lambovska, 2024; Crompton & Burke, 2023; Fuchs, 2023; Tikhonova & Raitskaya, 2023; Ivakhnenko & Nikolskiy, 2023), emphasizing the critical role of instructional context, task design, and learner characteristics in determining whether GenAI use promotes or undermines critical thinking.

Limitations in the Literature

The reviewed studies faced methodological limitations that affect the strength of inferences. Most relied on outcome-based indicators of CT, with few offering process-based or behavioural analyses (Boers et al., 2025; Suratmi & Sopandi, 2022). Instruments often included self-report questionnaires or task-based proxies, rather than longitudinal or mixed-method approaches (Essien et al., 2024; Fakour & Imani, 2025; Pan et al., 2025). Furthermore, the cognitive processes underpinning CT remain difficult to isolate in AI-mediated settings, and the influence of disciplinary differences was not systematically addressed.

The Role of Moderating Factors

Several moderating factors were identified that condition the impact of GenAI on CT. Individual factors such as motivation, AI literacy, self-directed learning, and critical disposition were found to correlate with better CT outcomes (Lijie et al., 2025; Shen & Teng, 2024; Suriano et al., 2025). Technological characteristics such as ease of use, perceived reliability, and platform structure played a nuanced role. For instance, Lijie et al. (2025) observed that while ease of use facilitated adoption, it paradoxically reduced cognitive effort. Pedagogical variables such as gamified learning, scaffolded prompts, and instructor feedback were consistently associated with improved CT (Chaparro-Banegas et al., 2024; George-Reyes et al., 2024; Walter, 2024; Lee et al., 2024).

Practical and Theoretical Implications

Instructional strategies such as prompt engineering, metacognitive scaffolding, and reflective tasks appear essential for maximising the educational value of GenAI. Walter (2024) and Gonsalves (2024) emphasised that teaching students to generate purposeful prompts cultivates both analytical reasoning and epistemic awareness. The reviewed studies highlight that GenAI can serve as a catalyst for critical thinking, provided that its use is guided, intentional, and aligned with pedagogical goals. These findings support a constructivist understanding of AI-enhanced learning and underscore the importance of integrating digital literacy and CT into academic curricula (Babin et al., 2024; Chen et al., 2025; Wong, 2024).

This review highlights that the influence of generative AI on critical thinking is shaped by a combination of pedagogical design, learner characteristics, and the nature of AI-supported tasks. While structured and reflective integration of GenAI tends to support the development of analytical and metacognitive skills, uncritical or instrumental use may lead to cognitive passivity and reduced engagement. The evidence points to the need for more targeted instructional strategies that align AI use with specific cognitive goals. These findings call for a reorientation of future research toward identifying the conditions under which GenAI fosters meaningful learning and for refining pedagogical models accordingly.

CONCLUSION

This scoping review examined how generative artificial intelligence (GenAI), particularly in educational contexts involving tools like ChatGPT, contributes to the development of critical thinking among university students. The analysis of recent empirical studies revealed a broad spectrum of approaches to integrating GenAI into teaching and learning. These included strategies such as prompt engineering, task-based learning, self-regulated and self-directed learning, AI-assisted writing, scaffolded instruction, and reflective evaluation of AI-generated content. The reviewed literature confirmed that, under appropriate pedagogical conditions, GenAI can serve as a meaningful catalyst for the development of critical thinking skills. Students who engage actively and consciously with AI tools tend to demonstrate increased metacognitive awareness, improved argumentation, and more effective reasoning.

At the same time, the review identified factors that may inhibit the positive effects of GenAI. Passive use of AI, lack of instructional support, and over-reliance on automatically generated outputs were associated with reduced cognitive engagement and a decline in students' ability to independently evaluate information. The outcomes of GenAI use were also shaped by individual learner variables such as motivation, AI literacy, and the ability to self-regulate learning activities.

Several limitations of the present review must be acknowledged. The analysis was restricted to publications indexed in the Scopus database and covered a narrow timespan, focusing only on the years 2024 and 2025. This may have resulted in the omission of relevant studies published earlier or indexed in other databases. Moreover, the majority of included research relied on short-term observations or self-reported data, which limits the generalisability of the findings and precludes conclusions about the long-term

impact of GenAI use on cognitive development. The lack of consistent theoretical frameworks across the reviewed studies also complicates efforts to compare results and synthesise evidence in a cumulative way.

Future research should address these gaps in several directions. First, longitudinal studies are needed to investigate the enduring effects of GenAI-supported instruction on critical thinking across diverse educational contexts and learner populations. Second, there is a clear need to develop more robust theoretical models that explicitly connect the dimensions of critical thinking, as defined in Bloom's taxonomy, with the specific learning mechanisms that are activated through the use of generative AI. Third, further work should focus on evaluating how different forms of instructional support, including scaffolded AI use and targeted AI literacy training, influence students' cognitive outcomes and their ability to engage with AI critically and productively. By advancing these research priorities, scholars and educators can better understand how to integrate GenAI tools into higher education in ways that support intellectual autonomy, foster analytical thinking, and reinforce the development of critical competencies in an increasingly digital learning environment.

DECLARATION OF COMPETING INTEREST

None declared.

AUTHORS' CONTRIBUTIONS

Lilia Raitskaya: conceptualization; data curation; formal analysis; investigation; methodology; resources; software; validation; visualization; writing – original draft; writing – review & editing; other contribution.

Elena Tikhonova: conceptualization; data curation; formal analysis; investigation; methodology; resources; software; validation; visualization; writing – original draft; writing – review & editing; other contribution.

REFERENCES

Ajlouni, A. O., Wahba, F. A. A., & Almahaireh, A. S. (2023). Students' attitudes towards using ChatGPT as a learning tool: The case of the university of Jordan. *International Journal of Interactive Mobile Technologies*, *17*(18), 99-117. https://doi.org/10.3991/ijim.v17i18.41753

Al-Mamary, Y.H., & Abubakar, A.A. (2025). Empowering ChatGPT adoption in higher education: A comprehensive analysis of university students' intention to adopt artificial intelligence using self-determination and technology-to-performance chain theories. *The Internet and Higher Education*, *66*, Article 101015. https://doi.org/10.1016/j.iheduc.2025.101015

Arksey, H., & O'Malley, L. (2005). Scoping studies: Towards a methodological framework. *International Journal of Social Research Methodology*, *8*(1), 19–32. https://doi.org/101080/1364557032000119616

Babin, J.I., Raber, H., & Mattingly II, T.J. (2024). Prompt pattern engineering for test question mapping using ChatGPT: A cross-sectional study. *American Journal of Pharmaceutical Education*, *88*, Article 101266. https://doi.org/10/1016/j.ajpe.2024.101116

Bertocchi, F. M., De Oliveira, A. C., Lucchetti, G., & Lucchetti, A.L.G. (2022). Smartphone use, digital addiction and physical and mental health in community-dwelling older adults: A population-based survey. *Journal of Medical Systems, 46*(8), 53. https://doi.org/10.1007/s10916-022-01839-7

Bloom, B. S., Krathwohl, D. R., Engelhart, M.B.., Furst, E.J., & Hill, W.H. (1956). *Taxonomy of educational objectives: The classification of educational goals, by a committee of college and university examiners. Handbook I: Cognitive domain.* Longman.

Boers, J., Etty, T., Baars, M., & van Boekhoven, K. (2025). Exploring cognitive strategies in human-AI interaction: ChatGPT's role in creative tasks. *Journal of Creativity*, *35*(1), Article 100095. http://doi.org/ 10.1016/j.yjoc.2025.100095

Chaparro-Banegas, N., Alicia, M.-T., & Roig-Tierno, N. (2024). Challenging critical thinking in education: New paradigms of artificial intelligence. *Cogent Education*, *11*(1), Article 2437899. https://doi.org/10.59429/esp.v9i11.3141

Chen, B., Zhang, Z., Langere, N., & Zhu, S. (2025). Unleashing the potential of prompt engineering for large language models. *Patterns*, *6*(6), Article 101260. https://doi.org/10.1016/j.patter.2025.101260

Crompton, H., & Burke, D. (2023). Artificial intelligence in higher education: the state of the field. *International Journal of Educational Technology in Higher Education*, *20*(1), 22. https://doi.org/10.1186/s41239-023-00392-8

Darwin, Rusdin, D., Mukminatien, N., Suryati, N., Laksmi, E.D., & Marzuki (2024). Critical thinking in the AI era: An exploration of EFLbstudents' perceptions, benefits, and limitations. *Cogent Education*, *11*(1), Article 2290342. https://doi.org/10.1080/2331186X.2023.2290342

de la Puente, M., Torres, J., Troncoso, A.L.B., Meza, Y.Y.H., & Carrascal, J.X.M. (2024). Investigating the use of ChatGPT as a tool for enhancing critical thinking and argumentation skills in international relations debates among undergraduate students. *Smart Learning Environments*, *11*, 55. https://doi.org/10.1186/s40561-024-00347-0

Deng, R., Jiang, M., Yu, X., Lu, Y., & Liu, S. (2025). Does ChatGPT enhance student learning? A systematic review and meta-analysis of experimental studies. *Computers & Education*, *227*, Article 105224. https://doi.org/10.1016/j.compedu.2024.105224

Emran, A.Q.M., Aldallal, A., & Nadheer, A. (2024). Investigating the impact of ChatGPT on enhancing university students' critical thinking skills. In A. Hamdan & A. Harraf (Eds.), *Business development via AI and digitalization* (vol. 1, pp. 567-574). Springer. https://doi.org/10.1007/978-3-031-62102-4_47

Essien, A., Bukoye, O.T., O'Dea, X., & Kremantzis, M. (2024). The influence of AI text generators on critical thinking skills in UK business schools. *Studies in Higher Education*, *49*, 5, 865-882. https://doi.org/10.1080/03075079.2024.2316881

Fabio, R.A., Plebe, A., & Suriano, R. (2025). AI-based chatbot interactions and critical thinking skills: An explanatory study. *Current Psychology*. https://doi.org/10.1007/s12144-024-06795-8

Fadillah, M.A., Usmeldi, U., & Asrizal, A. The role of ChatGPT and higher-order thinking skills as predictors of physics inquiry. *Journal of Baltic Science Education*, *23*(6), 1178-1192. https://doi.org/10.33225/jbse/24.23.1178

Fakour, H., & Imani, M. (2025). Socratic wisdom in the age of AI: a comparative study of ChatGPT and human tutors in enhancing critical thinking skills. *Frontiers in Education*. https://doi.org/10.3389/feduc.2025.1528603

Firat, M. (2023). What ChatGPT means for universities: Perceptions of scholars and students. *Journal of Applied Learning and Teaching*, *6*(1), 57-63. https://doi.org/10.37074/jalt.2023.6.1.22

Fuchs, K. (2023). Exploring the opportunities and challenges of NLP models in higher education: Is ChatGPT a blessing or a curse? *Frontiers in Education*, *8*, 1166682. https://doi.org/10.3389/feduc.2023.1166682

George-Reyes, C.E., Vilhunen, E., Avello-Martinez, R., & Lopez-Caudana, E. (2025). Developing scientific entrepreneurship and complex thinking skills; Creating narrative scripts using ChatGPT. *Frontiers in Education*, *9,* Article 1378564. https://doi.org/10.3389/feduc.2024.1378564

Gerlich, M. (2025). AI tools in society: impacts on cognitive offloading and the future of critical thinking. *Societies*, *15*, 6. https://doi.org/10.3390/soc15010006

Gonsalves, C. (2024). Generative AI's impact on critical thinking: Revisiting bloom's taxonomy. *Journal of Marketing Education*, 1-16. https://doi.org/10.1177/02734753241305980

Hwang, G.-J., Cheng, P.-Y., Chang, C.Y. (2025). Facilitating students' critical thinking, metacognition and problem-solving tendencies in geriatric nursing class: A mixed-method study. *Nurse Education in Practice*, *83*, Article 104266. https://doi.org/ 10.1016/j.nepr.2025.104266

Ivakhnenko, E. N., Nikolskiy, V. S. (2023). ChatGPT in higher education and science: A threat or a valuable resource? *Higher Education in Russia*, *32*(4), 9-22. https://doi.org/10.31992/0869-3617-2023-32-4-9-22

Jiang, T., Sun, Z., Fu, S., & Lv, Y. (2024). Human-AI interaction research agenda: A user-centered perspective. *Data and Information Management*, *8,* Article 1000078. https://doi.org/10.1016/j.dim.2024.100078

Krathwohl, D.R. (2002). A revision of Bloom's taxonomy: An overview. *Theory in to Practice*, *41*(4), 37-41. https://doi.org/10.1207/s15430421tip4104_2

Lee, H.-Y., Chen, P.-H., Wang, W.-S., Huang, Y.-M., & Wu, T.-T. (2024). Empowering ChatGPT with guidance mechanism in blended learning: Effect of self-regulated learning, higher-order thinking skills, and knowledge construction. *International Journal of Educational Technology in Higher Education*, *21*, 16. https://doi.org/10.org/10.1186/s41239-024-00447-4

Lijie, H., Yusoff, S.M., & Marzaini, A.F.M. (2025). Influence of AI-driven educational tools on critical thinking dispositions among university students in Malaysia: A study of key factors and correlations. *Educational and Information Technologies*, *30*, 8029-8053. https://doi.org/10.1007/s10639-024-13150-8

Lim, W. M., A. Gunasekara, J. L. Pallant, J. I. Pallant, & Pechenkina, E. (2023). Generative AI and the future of education: Ragnarök or Reformation? A paradoxical perspective from management educators*. The International Journal of Management Education*, *21*(2), Article 100790. https://doi.org/10.1016/j.ijme.2023.100790

Liu, Q., & Tu, C.C. (2024). Improving critical thinking through AI-supported socio-scientific issues instruction. *Journal of Logistics, Informatics and Service Science*, *11*(3), 52-65. https://doi.org/10.33168/JLISS.2024.0304

Melisa, R., Ashadi, A., Triastuti, A., Hidayati, S., Salido, A., Ero, P. E. L., Marlini, C., Zefrin., & Fuad, Z. A. (2025). Critical thinking in the age of AI: A systematic review of AI's effects on higher education. *Educational Process: International Journal*, *14*, Article e2025031. https://doi.org/10.22521/edupij.2025.14.31

Naatonis, R.N., Rusijono, Jannah, M., & Malahina, E.A.U. (2024). Evaluation of Problem Based Gamification Learning (PBGL) model on critical thinking ability with Artificial Intelligence approach integrated with ChatGPT API: An experimental study. *Qubahan Academic Journal*, *4*(3), 485-520. https://doi.org/10.58429/qaj.v4n3a919

Oates, A., & Johnson, D. (2024). ChatGPT in the classroom: Evaluating its role fostering critical evaluation skills. *International Journal of Artificial Intelligence in Education*. https://doi.org/10.1007.s40593-024-00452-8

Octaberlina, L.R., Muslimin, A.I., Chamidah, D., Surur, M., & Mustikawan, A. (2024). Exploring the impact of AI threats on originality and critical thinking in academic writing. *Edelweiss Applied Science and Technology*, *8*(6), 8805-8814. https://doi.org/10.55214/25768484.v8i6.3878

Pan, Z., Moore, O.A., Papadimitriou, A., & Zhu, J. (2025). AI literacy and trust: a multi-method study of Human-GAI team collaboration. *Computers in Human Behavior: Artificial Humans, 4*, Article 100162. https://doi.org/10.106/j.chbah.2025.100162

Panit, N.M. (2024). Can critical thinking and AI work together? Observations of science, mathematics, and language instructors. *Environment and Social Psychology, 9*(11), Article 3141. https://doi.org/10.59429/esp.v9i11.3141

Promma, W., Imjai, N., Usman, B., & Aujirapongpan, S. (2025). The influence of AI literacy on complex problem-solving skills through systematic thinking skills and intuition thinking skills: An empirical study in Thai gen Z accounting students. *Artificial Intelligence*, *8*, Article 100382. https://doi.org/10.1016/j.caeai.2025.100382

Raitskaya, L., & Lambovska, M. (2024). Prospects for ChatGPT application in higher education: A scoping review of international research. *Integration of Education*, *28*(1), 10-21. https://doi.org/10.15507/1991-9468.114.028.202401.010-021

Risko, E.F., & Gilbert, S.J. (2016). Cognitive offloading. *Trends in Cognitive Sciences*, *20,* 676-688. htto://doi.org/10.1016/j.tics.2016.07.002

Royer, C. (2024). Outsourcing humanity? ChatGPT, critical thinking, and the crisis in higher education. *Studies in Philosophy and Education*, *43*, 479-497. https://doi.org/10.1007/s11217-024-09946-3

Santamaria-Velasco, J., Nunez-Naranjo, A., & Morales-Urrutia, X. (2025). Critical thinking and AI: Enhancing history teaching through ChatGPT simulations. *Internation Journal of Innovative Research and Scientific Studies*, *8*(1), 564-575. https://doi.org/10.53894/ijirss.v8i1.4403

Sardi, J., Darmansyah, Candra, O., Devi Faizah, Y., Habibullah, Yanto, D.T.P., & Fivia, E. (2025). How generative AI influences students' self-regulated learning and critical thinking skills? A systematic review. *International Journal of Engineering Pedagogy*, *15*(1), 94-108. https://doi.org/10.3991/ijep.v15i1.53379

Shen, X., & Teng, M.F. (2024). Three-wave cross-lagged model on the correlations between critical thinking skills, self-directed learning competency and AI-assisted writing. *Thinking Skills and Creativity*, *52*, Article 101524. https://doi.org/10.1016/j.tsc.2024.101524

Steele, J.L. (2023). To GPT or not GPT? Empowering our students to learn with AI. *Computers and Education: Artificial Intelligence*, *5*, 100160. https://doi.org/10.1016/j.caeai.2023.100160

Suratmi, S., & Sopandi, W. (2022). Knowledge, skills, and attitudes of teachers in training critical thinking of elementary school students. *Journal of Education and Learning (EduLearn), 16*(3), 291–298. https:// doi. org/ 10. 11591/ edule arn.v16i3. 20493

Suriano, R., Plebe, A., Accial, A., & Fabio., R.A. (2025). Student interaction with ChatGPT can promote complex critical thinking skills. *Learning and Instruction*, *95*, Article 102011. https://doi.org/9510.1016/j.learninstruc.2024.102011

Tang, T., Sha, J., Zhao, Y., Wang, S., Wang, Z., & Shen, S. (2024). Unveiling the efficacy of ChatGPT in evaluating critical thinking skills through peer feedback analysis: Leveraging existing classification criteria. *Thinking Skills and Creativity*, *52*, Article 101607. https://doi.org/10.1016/j.tsc.2024.101607

Tikhonova E., & Raitskaya L. (2023). ChatGPT: Where Is a silver lining? Exploring the realm of GPT and large language models. *Journal of Language and Education*, *9*(3), 5-11. https://doi.org/10.17323/jle.2023.18119

Tricco, A.C., Lillie, E., Zarin, W., O'Brien, K.K., Colquhoun, H., Levac, D., Moher, D., Peters, M.D.J., Horseley, T., Weeks, L., Hempel, S., & Akl, E.A. (2018). PRISMA extension for scoping reviews (PRISMA-ScR): Checklist and explanation. *Annals of Internal Medicine*, *169*(7), 467–73. https://doi.org/107326/M18-0850

van Rensburg, J.J. (2024). Artificial human thinking: ChatGPT's capacity to be a model for critical thinking when prompted with problembased writing activities. *Discover Education*, *3*(1), 42. https://doi.org/10.1007/s44217-024-00113-x

Walter, Y. (2024). Embracing the future of Artificial Intelligence in the classroom: The relevance of AI literacy, prompt engineering, and critical thinking in modern education. *International Journal of Technology in Higher Education, 21*, Article 15. https://doi.org/10.1186/s41239-024-00448-3

Wang, W.-S., Lin, C.-J., Lee, H.-Y., Huang, Y.-M., & Wu, T.-T. (2025). Enhancing self-regulated learning and higher-order thinking skills in virtual reality: The impact of ChatGPT-integrated feedback aids. *Education and Information Technologies.* https://doi.org/10.1007/s10639-025-13557-x

Wong, R. S.-Y. (2024). ChatGPT in medical education: Promoting learning or killing critical thinking? *Education in Medicine Journal*, *16*(2), 177-183. https://doi.org/10.21315/eimj2024.16.2.13

Xiao, X., Li, Y., He, X., Fang, J., Yan, Z., & Xie, C. (2025). An assessment framework of higher-order thinking skills based on fine-tuned large language models. *Expert Systems with Applications*, *272*, Article 126531. https://doi.org/10.1016/j.eswa.2025.126531

Xu, J., & Liu, Q. (2025). Uncurtaining windows of motivation, enjoyment, critical thinking, and autonomy in AI-integrated education: Duolingo Vs. ChatGPT. *Learning and Motivation*, *89*, Article 102100. https://doi.org/10.1016/j.lmot.2025.102100

Yusuf, A., Bello, S., Pervin, N., & Tukur, A.K. (2024). Implementing a proposed framework for enhancing critical thinking skills in synthesizing AI-generated texts. *Thinking Skills and Creativity*, *53*, Article 101619. https://doi.org/10.1016/j.tsc.2024.101619

Zhai, X., Nyaaba, M., & Ma, W. (2025). Can generative AI and ChatGPT outperform humans on cognitivedemanding problem-solving tasks in science? *Science & Education*, *34*, 649-670. https://doi.org/10.1007/s11191-024-00496-1

Zhou, X., Teng, D., & Al-Samarraie, H. (2024). The mediating role of generative ai self-regulation on students' critical thinking and problem-solving. *Education Sciences*, *14*, Article 1302. https://doi.org/10.3390/educsci14121302

https://doi.org/10.17323/jle.2025.17642

Boosting Punctuation Proficiency: The Power of an Interactive Chatbot for EFL Learners

Ali Al Ghaithi [1], **Behnam Behforouz** [1,2]

[1] Sohar University, Oman

[2] University of Technology and Applied Sciences, Shinas, Oman

ABSTRACT

Background: While technology is becoming more integrated into education, little research has been done on the contribution that chatbots must play when it comes to helping EFL learners gain some language ability, like grammar. Most current research focuses on vocabulary or learning grammar and thus does not understand how interactive, computer-based technologies can help develop rule-based writing skills. This study fills this gap by identifying the degree to which a WhatsApp chatbot improves punctuation skills, thus gaining insight into the promise of chatbot-enhanced teaching for certain language subskills.

Purpose: To evaluate the effect of a WhatsApp-based chatbot on improving Omani EFL learners' knowledge of English punctuation marks.

Method: Sixty Omani EFL learners from Sohar University in Oman participated in this study, divided into a control and experimental group. Both groups obtained in-class training on punctuation marks. The experimental group received additional explanation and practice regarding punctuation marks using an interactive chatbot. The chatbot bot was developed for the experimental group to enable two-sided interactions with the students. All the tests were piloted to ensure reliability and validity before the data collection and the main study round. A pretest, followed by a posttest and a delayed posttest, was conducted to compare participants' performance on English punctuation marks.

Results: The experimental group's mean scores increased from 8.06 (pretest) to 26.66 (posttest) and 24.16 (delayed posttest), significantly exceeding those of the control group. Kruskal-Wallis tests revealed no prior differences ($p = 1.000$), but statistically significant improvements were observed in the experimental group at both the posttest and delayed posttest ($p < .001$; $\eta^2 > 0.73$), with large effect sizes. These results indicate the lasting impact of the chatbot on punctuation ability.

Conclusion: This study underscores the importance of integrating interactive technologies into language learning environments to foster learner engagement and independence. By prioritising learners' needs and preferences, instructors can develop more efficient, student-centred methods that promote enhanced language acquisition. The study's findings highlight the potential of cutting-edge tools, such as interactive chatbots, to enhance language acquisition and promote ongoing improvement.

KEYWORDS:

Interaction, WhatsApp bot, EFL, Punctuation

Citation: Al Ghaithi, A., & Behforouz, B. (2025). Boosting punctuation proficiency: The power of an interactive chatbot for EFL learners. *Journal of Language and Education, 11*(2), 20-34. https://doi.org/10.17323/jle.2025.17642

Correspondence:
Ali Al Ghaithi,
AGhaithi@su.edu.om

Received: March 14, 2024
Accepted: June 10, 2025
Published: June 30, 2025

INTRODUCTION

The integration of artificial intelligence (AI) into education has become increasingly prevalent in recent years, offering novel ways to individualize instruction and enhance student engagement (Roos, 2018; Schmidt & Strasser, 2022). Defined as a field focused on developing methods for high-level reasoning based on low-level input features without the need for direct human intervention (Healey, 2020), AI is now being applied across various educational domains. Among the most widely adopted AI applications in education are chatbots (automated conversational agents that provide immediate, interactive feedback) (Okonk-

wo & Ade-Ibijola, 2021). These tools have demonstrated potential to foster learner autonomy, increase motivation, and enhance language acquisition outcomes, particularly in vocabulary development and communicative competence (Wei-Xun & Jia-Ying, 2024; Song & Xiong, 2023; Wei, 2023).

In the context of language education, chatbots have been utilised to facilitate learner–system interaction through both textual and spoken input (Kerly et al., 2006) and have proven effective in addressing students' diverse learning needs (Colace et al., 2018). Their adaptability enables personalised pacing, independent exploration, and sustained engagement, while their integration into virtual classrooms has been linked to improved feedback quality and increased learner involvement (Chou et al., 2021; Vázquez-Cano et al., 2021; Essel et al., 2022). The versatility of available chatbot platforms (from low-code solutions like Flow XO and Botsify to more advanced environments like DialogFlow) makes their adoption accessible to a broad range of users, regardless of technical background (Satam et al., 2020; Kumar, 2021).

Despite growing interest in AI-supported language learning, a notable gap remains in the use of such tools to support writing-related subskills, particularly punctuation. Punctuation serves not only as a system of standardised marks that enhances textual coherence (Jan, 2009), but also as a crucial discourse marker that facilitates the interpretation of meaning (Daffern & Mackenzie, 2015; Scull & Mackenzie, 2018). Its correct use is considered a key indicator of linguistic and communicative competence (Vázquez-Cano et al., 2018). Nevertheless, traditional approaches to teaching punctuation, such as rule memorization and sentence correction, often fail to foster a deep understanding of its communicative function (Fang & Wang, 2011; Macken-Horarik & Sandiford, 2016). Learners benefit more from experiential and meaning-based engagement with punctuation, rather than from mechanical drills (Scull & Mackenzie, 2018). Yet, providing timely and individualized feedback (a critical component of such instruction) can be logistically challenging in conventional classroom settings.

Given these challenges, chatbots have been proposed as promising tools for delivering individualized writing practice and immediate feedback outside the classroom (Shail, 2019; Subramaniam, 2019; Kurup et al., 2021). However, most existing chatbots are designed for general conversational use in the learner's first language (L1), and there remains a lack of readily available tools for second language (L2) learners targeting specific writing competencies, such as punctuation (Kwon et al., 2023). This oversight is particularly relevant in contexts where writing accuracy is essential for academic success in English as a Foreign Language (EFL) settings.

To address this gap, the present study examines the use of an interactive WhatsApp-based chatbot designed to support punctuation training among Omani English as a Foreign Language (EFL) learners. While prior research has recognized the educational potential of WhatsApp bots (Al Ghaithi et al., 2024; Behforouz & Al Ghaithi, 2024), their application for targeted subskill development (specifically in punctuation) has yet to be examined. This study contributes to the literature by evaluating the extent to which chatbot-supported instruction can enhance learners' punctuation proficiency, offering insights into how interactive digital tools may be strategically employed to improve a frequently neglected aspect of language learning.WhatsApp was chosen because it is widely used in Oman, especially among university students. This makes it a user-friendly, well-known, and cheap way to teach languages. Other systems, such Elbot or Kuki, are less common in the area and may be harder to use or integrate. The emphasis on punctuation as a subskill rectifies a recognised deficiency in EFL instruction, where punctuation is frequently undervalued despite its crucial role in maintaining clarity and coherence in writing (Crystal, 2012; Truss, 2003). Moreover, focusing on Omani EFL learners addresses a particular contextual requirement: prior research has primarily emphasised grammar and vocabulary, resulting in an inadequate analysis of punctuation within this field. The present study addresses this deficiency by utilising a frequently used communication tool, thereby providing a pragmatic and contextually relevant enhancement to conventional research methodologies. Accordingly, this research investigates the following question: To what extent can a WhatsApp Bot improve Omani EFL students' proficiency in using punctuation marks?

LITERATURE REVIEW

Benefits of Chatbots in Education

Chatbots have quickly emerged as revolutionary tools in educational environments, particularly in optimising and individualising the learning and teaching process. Chatbots can be practical tools for guiding teaching-learning processes by assessing student work, enhancing materials, and tailoring training schedules based on individual needs (Bii, 2013; Ghose & Barua, 2013), and they are associated with microlearning tasks that give learners more authority over the procedure of education instruction and allow them to choose the pace at which they complete the task. Using this educational technique, students can more easily create learning environments and experiences where they can practice, study, or engage within a limited timeframe (Bruck et al., 2012; Vázquez-Cano, 2012, 2014). Research has shown that utilizing chatbots or virtual agents can enhance students' academic achievement and motivation in learning environments (Liew et al., 2017). Kamita et al. (2019) coupled chatbots with online education and discovered that chatbots aided self-learning, boosted motivation, and reduced stress. These studies together demonstrate the pedagogical effectiveness of chatbots; nonetheless, they predominantly highlight general academic enhancements without assess-

ing the technology's performance in particular language subskills or diverse educational contexts.

The benefits of employing AI chatbots in English education, together with the essential pedagogical components, have been thoroughly examined in another study. A total of 28 prospective instructors specializing in English education were instructed to utilize Kuki chatbots for one week, after which they were required to provide an evaluation of their experiences. The qualitative data from surveys and interviews were subjected to thematic analysis, identifying six main themes: teacher perspectives, student perspectives, communication, linguistic factors, affective factors, and assessment. The comprehensive results indicate that preservice teachers recognize AI chatbots as valuable tools for facilitating teaching and learning, benefiting both educators and students. Applying learner data to chatbot technology requires a strategic approach to ensure optimal chat interactions. Additionally, the research found that chatbots have the potential to enhance the confidence and motivation of English as a Foreign Language (EFL) learners in speaking the English language (Yang, 2022). This qualitative dimension adds significant complexity to predominantly quantitative research, suggesting that the perceptions of both learners and instructors on the usefulness of chatbots are vital for assessing long-term adoption. Yang's research enhances our understanding of teacher cognition; however, it does not directly link perceptions to quantifiable classroom outcomes. This indicates a significant disparity between theoretical possibilities and practical realities in this area.

The integration of AI chatbot help with collaborative notetaking (CNT) has been examined for its effect on the acquisition of semantic knowledge in EFL learners. Chen (2024) employed a quasi-experimental design and included 60 students from the English language and literature program. It aimed to compare the group that received support from an AI chatbot (AI-CNT) with the group that received conventional support (CNT). The findings demonstrated that the AI-CNT group exhibited superior performance compared to the CNT group in terms of learning performance, achievement, self-efficacy, metacognition, and reduced learning anxiety. The study highlights the potential of utilizing AI chatbot interaction in conjunction with the CNT technique to enhance the EFL semantic learning experience. This approach offers tailored language practices that involve interaction and are enhanced with feedback and emotional support. Nonetheless, it remains unclear whether these advantages stem from the chatbot's linguistic support, emotional reinforcement, or feedback system and how these components may vary across different educational contexts. The studies together provide compelling evidence for the effectiveness of chatbots in EFL teaching, particularly in enhancing motivation, vocabulary, and speech proficiency. However, the existing work frequently overlooks longitudinal designs, cross-skill integration, and comparative evaluations with alternative digital tools. Moreover, several studies focus on emotional or perceptual outcomes while failing to effectively link them to language development benchmarks. This study aims to address these deficiencies by examining the immediate linguistic outcomes of chatbot-assisted learning, as well as the cognitive and affective variables that influence these outcomes, particularly within underrepresented English as a Foreign Language (EFL) populations. These findings confirm the significant educational potential of chatbots; however, a comprehensive analysis of their impact on specific language acquisition subskills, particularly speaking, listening, and spelling, is essential for elucidating their functional scope and pedagogical precision.

Chatbots and Langauge Learning Subskills

Speaking and listening skills, essential for communicative competency, have been positively impacted by chatbot-assisted learning. In a study using the LINE Chatbot messaging app to examine exercises for English conversations, one of the educational activities in which 73 students participated was a four-week English conversation exercise that included both speaking and listening components. Participants established the experimental and control groups. The two groups engaged in learning activities after school using LINE Chatbot. The experimental group signed up for competitive discussion with the LINE Chatbot, whereas the control group signed up for unrivaled chat exercises. After school, while the experimental group used the LINE Chatbot to engage in competitive discussion activities, the control group did the same activities without assistance. According to the results, the LINE Chatbot enhanced each group's pupils' speaking and listening abilities (Chien et al., 2022). The results demonstrate that the level of chatbot interaction, specifically if it is competitive and dynamic, is central to improving vocal language abilities. This study did not examine the long-term retention of improvements or the impact of learner characteristics, such as anxiety levels and past exposure, on the outcomes. This distinction warrants further examination regarding how the integration of chatbots may affect specific cognitive and affective aspects of language acquisition.

The integration of voice chatbots with AI into language education has shown promise in increasing the proficiency of Vietnamese undergraduate students in speaking English through innovative and interactive means of developing fluency and communication skills. Duong and Suppasetseree (2024) conducted a quasi-experimental study involving 30 Vietnamese undergraduate students over an eight-week period. The participants engaged in English-speaking exercises with an AI voice chatbot over two weekly class sessions. They also took a speaking exam before and after the experiment, answered a questionnaire, and participated in a semi-structured interview. The findings indicated a notable improvement in students' English-speaking proficiency after using the AI voice chatbot, characterised by an enhanced use of hedging phrases, grammatical structures, and vocabulary.

The study proposes the incorporation of AI voice chatbots in teaching and learning activities to improve the speaking abilities of Vietnamese EFL undergraduate students. Furthermore, there is no consensus in current research about the emotional aspects that most consistently benefit from chatbot use, particularly across various cultures and levels of competency.

Supplementary research supports the efficacy of chatbots in enhancing receptive and productive abilities over extended instructional periods. In another study, Kim (2018) assessed the impact of chatbots on the listening and reading comprehension of English among 46 college students. Participants were randomly distributed between the experimental and control groups. For 16 weeks, members of the control group did not communicate with Elbot, whereas members of the experimental group conversed with the chatbot. Pretests and posttests were administered to measure the effectiveness of the treatment. The findings revealed that reading and listening comprehension increased significantly for both groups. In the post-listening assessment, however, the improvements in the experimental group were more noticeable. Furthermore, after being encouraged to interact with the chatbot, the experimental group showed improvement in listening comprehension, progressing from the intermediate to the advanced stage. The longitudinal gains have encouraged more studies into utilizing advanced AI technologies in speech learning, particularly those with voice recognition and natural language generation capabilities. Nonetheless, the study did not investigate whether learners employed different strategies while interacting with chatbots compared to traditional training, nor did it identify which chatbot features, such as personalization, adaptivity, or repetition, were more impactful.

Using chatbots in English learning settings has shown encouraging outcomes, especially in fundamental language competencies such as spelling. In a study aimed at developing an interactive chatbot to facilitate spelling improvement and gather students' perspectives on using chatbots in the language learning process, 60 Omani English as a Foreign Language (EFL) students were evenly divided into two groups: a control group and an experimental group. The results revealed that using interactive chatbots in an educational setting enhanced spelling proficiency among participants in the experimental group (Al Ghaithi et al., 2025). These findings demonstrate the key advantage of chatbot interventions: measurable improvements in academic achievement. This study emphasizes cognitive and affective benefits while presenting challenges related to scalability, long-term effects, and relevance to more complex linguistic domains, such as discourse-level speaking. Moreover, the absence of comparative data with other digital solutions leads to a methodological shortcoming. Chatbot-enhanced instruction has proven effective for fundamental receptive and productive language skills; however, its impact on writing subcomponents such as punctuation remains underexplored, warranting more rigorous empirical investigation and targeted instructional innovation.

Technology and Punctuation

Experimental research has examined the impact of digital aids on learners' punctuation proficiency in academic writing. Ivanova et al. (2022) attempted to measure the effect of digital support on learners' appreciation and mastery of punctuation in academic writing, investigating this with 42 students. Standard face-to-face practices based on textbooks were implemented in the control group, whereas the experimental group underwent training using digital resources and a simplified textbook. Three final examinations served as a gauge of punctuation knowledge and proficiency. The findings showed that learning academic writing from standard textbooks alone was insufficient for improving punctuation. Digital assistance for teaching punctuation had considerable pedagogical potential. In the final assessments, the experimental group outperformed the control group. As a result, digital assistance should be a key component of academic writing instruction and included in training curricula. The study primarily focused on the number of punctuation marks, while additional studies have explored the types and frequencies of punctuation marks. This research highlighted the frequency of punctuation used, suggesting future studies analyze both the frequency and types of marks used, indicating a need for further scrutiny in future research. This study yields substantial quantitative data; nonetheless, it needs to examine how learners cognitively process punctuation usage using digital tools or whether these tools augment metalinguistic awareness. A cross-analysis of learners' proficiency levels and writing genres might yield more targeted pedagogical insights.

Interactive multimedia systems have demonstrated significant efficacy in teaching learners with disabilities about punctuation at various educational levels. To assist middle and high school students with various learning difficulties in applying punctuation techniques, a multimedia software tool was developed. This approach aims to determine whether a multimedia tool is effective in enhancing punctuation instruction for students who experience academic difficulties. Every student at every school level was randomly allocated to the experimental or control group. According to the results, experimental pupils at all school levels outperformed the control group, scoring noticeably better on the punctuation usage exam. Furthermore, the experimental students used less erroneous punctuation compared to the controls in the two educational groups. Additionally, the experimental students used the skills in a written test in which they composed original statements. Therefore, this study demonstrates that when taught using interactive multimedia, individuals can acquire punctuation procedures with high competence and apply their skills to sentence modification, as well as phrase-building problems. However, boundaries remain, particularly on the program's effects

on students' independent writing after instruction, highlighting the need for future research focused on long-term retention and transferability (Schumaker et al., 2019). This study provides significant evidence for the effectiveness of multimedia-assisted learning in students with impairments; however, it does not examine whether mainstream learners can achieve analogous advantages. Moreover, there is a lack of examination into whether the multimedia tool promotes strategic self-regulation or learner reflection, both of which are essential for sustainable writing improvement.

The efficacy of WhatsApp Messenger has been tested in various educational settings to improve students' punctuation marks knowledge. In one such study, Abdul Fattah (2015) examined an experimental group comprising undergraduate students who received instructional materials through WhatsApp to determine whether it could enhance the student's use of punctuation marks. The results revealed that students' proper use of punctuation marks significantly increased in the presence of WhatsApp Messenger. The study scope was limited to one skill area, implying the need to extend mobile treatments to encompass more facets of writing development components.

Self-editing programs have helped analyze the role of smartphone usage in developing EFL learners' self-editing skills. Eighteen Saudi Arabian students were assigned White Smoke, a self-editing application, to evaluate how technology could assist them in enhancing their writing skills. In week three, the learners were instructed to analyse and correct punctuation mark mistakes. The results revealed significant progress in the experimental group's ability to distinguish and correct punctuation-related errors within their learning context. Notwithstanding these advancements, the researchers acknowledged the limitations of their scope and sample size, advocating for more comprehensive investigations into the effects of mobile learning on various linguistic characteristics (Al-Wasy & Mahdi, 2016). This intervention's structured progression offers valuable insights into phased editing growth; however, the study overlooks the consideration of student viewpoints on the editing process and the advancement of their metacognitive editing strategies. These elements necessitate additional examination to understand not just the alterations that occur but also the underlying mechanisms and rationale behind them.

It has also become a source of interest in developing a new educational tool for teaching punctuation marks to university students using chatbots. The development of one to support Spanish university students in learning punctuation is a testament to the potential of AI-driven education in higher learning. In this case, the control group received instructions via the traditional method of writing exercises on paper, while the experimental group followed the instructions through a chatbot. The results revealed that students who used chatbots to interact with the materials performed more significantly than their counterparts in the control group. The chatbot provided a means for interacting with course materials, providing instant feedback and tailored responses that increased interaction and comprehension. This study suggests that chatbots may offer focused microinstruction in writing mechanics, indicating a need for further research into the application of chatbots in comprehensive writing instruction. As technology continues to develop, ongoing research will be necessary to assess its long-term effects and potential applications in other learning environments (Vázquez-Cano et al., 2021). This study highlights the potential of chatbot-assisted writing education; however, it does not compare it with human-facilitated formative feedback or examine the impact of chatbot characteristics, such as adaptivity and feedback type, on learning outcomes. These factors remain unexamined and are essential for the advancement of pedagogically sound AI systems.

AI-assisted training has demonstrated significant potential in several language abilities, including speaking, listening, reading, and spelling, yet punctuation remains a remarkably underexplored domain in EFL chatbot-based learning. Current research either focuses on general writing improvement or uses digital techniques unrelated to chatbot technology. There is a notable deficiency of focused studies on mobile-based chatbot instruction concerning punctuation, particularly in the context of WhatsApp, a widely accessible and familiar platform for learners in many English as a Foreign Language (EFL) environments. This project fills the gap by developing and utilising a WhatsApp-based chatbot that provides targeted microinstruction on punctuation. This study evaluates measurable language enhancements, offering a comprehensive understanding of chatbot efficacy in a sometimes overlooked facet of writing instruction.

METHOD

Participants

To collect the required data, 60 Omani students of English as a Foreign Language (EFL) were randomly selected as the study population. These students studied English in the General Foundation Program (GFP), a preparatory program that allows students to move to higher education studies. The participants' ages ranged from 18 to 19 years old. The students, who were a combination of both genders, were native Arabic speakers, and their English proficiency level was intermediate, as determined by the institution's placement test.

Research Instruments

English Punctuation Tests

The researcher created an English punctuation pretest, posttest, and delayed posttest to track participants' per-

formance before and after implementing the WhatsApp Bot, aiming to determine its effectiveness in helping Omani students use English punctuation marks correctly. Each test consisted of three sections: Section 1, which included 10 questions for correcting incorrect punctuation marks within sentences. In Section 2, another 10 multiple-choice questions were related to selecting correct punctuation marks, and Section 3 included 10 questions on True/False statements based on the punctuation marks used within the statements. The logic behind using these types of questions was that learners' regular exams and quizzes within the semester were all in the same format as in this study; therefore, students were familiar with the format of the questions. The pretest is provided in as an example (Appendix 1).

To ensure the reliability of the punctuation mark tests, a pilot study was conducted with 20 Omani intermediate English as a Foreign Language (EFL) students within the same institution. Two Omani PhD holders in Applied Linguistics were selected to review and refine the questions to validate those of the three tests. Table 1 shows the reliability results for the pretest, posttest, and delayed posttest. As can be seen, the tests had high-reliability indices of 0.860, 0.780, and 0.885, respectively.

Table 1

The Reliability of all Sets of the Tests

Tests	Cronbach's Alpha	N of Items
Pretest	.860	30
Posttest	.780	30
Delayed posttest	.885	30

Whatsapp Bot

The WhatsApp bot employed a decision-tree algorithm to determine the appropriate punctuation exercises, taking into account the student's prior responses and development. For example, if a learner provided a valid answer, the bot would present increasingly intricate exercises. In contrast, if errors were identified, the bot provided fewer complex exercises to strengthen understanding. The bot promptly delivered feedback in various formats: for questions with True or False options, it verified the correctness of the student's selection, and for exercises involving sentence correction, it supplied the appropriate punctuation when mistakes were made. The rapid feedback approach enabled pupils to learn from their errors promptly.

The developer pre-programmed and meticulously examined the responses to guarantee that the bot's input was high quality and accurate. The bot's database provided accurate responses for all questions, ensuring that students received dependable and pedagogically effective feedback. The bot's stability was further enhanced through regular monitoring and adjustments based on user interactions.

Punctuation Marks

The WhatsApp bot was developed using Python programming and linked to a nearby mobile device, facilitating smooth and uninterrupted interaction between the students and the bot. This configuration enabled students to engage with the bot via their WhatsApp program, providing them with the convenience of receiving and responding to punctuation exercises. The bot can send various questions, respond instantly, and distinguish between accurate and inaccurate answers, providing an interactive and engaging learning experience.

Although there are various punctuation marks in English, the primary focus of this study was on the full stop, apostrophe, comma, parenthesis, and question mark as the targets of the current study, which aimed to interact with the experimental group through the WhatsApp bot. The reason for selecting these types of punctuation marks was their commonalities among the writing papers of students, as determined through a detailed evaluation of their writing tasks within regular training. To elicit accurate details on the punctuation marks, a book by Noah Lukeman (2007) called A Dash of Style: The Art and Mastery of Punctuation was selected as the primary source of providing information.

Three different types of questions were developed for each English punctuation mark (True or False, multiple-choice, and sentence correction by writing the correct punctuation mark in six questions for each punctuation mark). These test formats were regularly used in the exams and continuous assessments of learners; therefore, there was no need for extra training or explanation on how to select the correct responses.

Procedures

This research was conducted in the fall semester of 2023 at Sohar University in the North Al Batinah region. Initially, learners were informed that their participation was voluntary and would not affect their future assessments. A pretest was conducted before implementing the treatment to ensure the homogeneity of the participants' knowledge of punctuation marks. Both groups were provided in-person instruction during their lessons to maintain a uniform teaching environment. Nevertheless, the experimental group employed the WhatsApp bot as a mediator, with the teacher utilizing the bot's interactive capabilities to send supplementary materials and exercises. This facilitated supplementary practice and consolidation beyond the designated classroom time. The researcher guided the experimental group on troubleshooting procedures and conducted a

training session to ensure comprehensive comprehension of the process among all students. The treatment process lasted for three weeks. Two sessions were specified each week to teach and practice the punctuation marks.

The questions and exercises were shared through the interactive chatbot for the experimental group. In True or False questions, students were given sentences with incorrect punctuation marks and had to decide whether the sentences were punctuated correctly or incorrectly by selecting True or False. In the second format, learners were given sentences with punctuation problems. The sentences must be rewritten with proper punctuation. Finally, students were given statements and sentences to be completed with the best option between the two others. In all these techniques, the bot provided immediate feedback to students on whether their answer was correct or incorrect.

It was possible to practice the questions indefinitely using WhatsApp Bot. In addition to demonstrating the use of the WhatsApp bot to practice English punctuation, the researcher advised the experimental group on what to do in the case of issues. Four questions were created as a training exercise for the colon and slash, among other punctuation marks, and the researcher held a training session to ensure that every student understood the process well.

For learners to learn from their mistakes, the bot informed them if their responses were correct or incorrect. Additionally, if the students' responses were inaccurate, the bot corrected them by providing the learners with the correct ones, helping them immediately recognize their mistakes. In contrast, when the instructions over some punctuation marks ended in the control group, similar types of exercises, including true/false, sentence correction, and multiple-choice, were administered using paper-based materials. To provide feedback to this group, the teacher identified errors in the students' assignments and offered general feedback to the class as a whole. Unlike using a chatbot, which provided personalised automated feedback, the control group received no individualised feedback. Immediately after the last week of treatment, the researchers conducted a posttest for both groups. Moreover, three weeks later, a delayed posttest was conducted.

RESULTS

This section presents the procedures and findings of the data analysis, including the assessment of normality, and systematically addresses the research questions using appropriate and interpretive methods. The primary research question of this study aims to investigate the effect of using an interactive chatbot on learners' proficiency in using English punctuation marks. Therefore, the following statistical procedures and tests were conducted comprehensively:

Measuring the Normality of Data in all the Tests

Initially, conducting a test of normality was necessary to select the appropriate parametric or nonparametric test for analyzing the data from punctuation tests in depth. Therefore, a Kolmogorov-Smirnov test was done, and the results are presented in Table 2.

Table 2

The Results of the Kolmogorov-Smirnov Test of Normality for all the Tests in two Groups

	Statistic	df	Sig.
Control pre	.136	30	.164
Control post	.166	30	.034
Control delayed	.121	30	.200
Experiment pre	.184	30	.011
Experiment post	.151	30	.080
Experiment delayed	.184	30	.011

Table 2 shows that the control pretest and delayed posttest data do not substantially deviate from a normal distribution (p-values 0.164 and 0.200, respectively), indicating that they conform to a normal distribution. However, the control posttest, experimental pretest, and experimental delayed posttest data show substantial deviations from normality, with p-values of 0.034, 0.011, and 0.011, respectively. The experimental posttest data, with a p-value of 0.080, is on the threshold and may not exhibit a substantial departure from normality. These findings suggest that nonparametric tests may be suitable for comparing the groups across all tests.

Performance of Control Group in Punctuation Tests

To monitor the performance of students in the control group based on the punctuation tests, a Wilcoxon test was performed, and the results are presented in Table 3.

As shown in Table 3, the statistical tests comparing different test times indicate significant variations in scores throughout each testing period. The comparisons between posttest and pretest ($Z = -4.789$, $p = 0.000$), delayed posttest and pretest ($Z = -4.237$, $p = 0.000$), and delayed posttest and posttest ($Z = -4.800$, $p = 0.000$) all demonstrate p-values below 0.05, showing statistically significant enhancements. The results indicate that the intervention led to notable improvements in performance from the first assessment to the final assessment, and these enhancements were sustained throughout the follow-up assessment. Additionally, extra progress was observed between the posttest and the delayed posttest. The effect size of all sets was measured accordingly to gain more detailed information and determine the exact differences and significance. The results are presented in Table 4.

Table 4, based on the effect sizes for the various test times, reveals significant disparities in performance. The pretest yielded a Cohen's d value of 2.158, indicating a substantial effect size. The 95% confidence interval for this effect size is from 1.494 to 2.810. The posttest demonstrates a significantly larger effect size, with a Cohen's d value of 9.816 and a confidence interval of 7.274 to 12.349, indicating a significant improvement. The delayed posttest demonstrates a substantial effect size, with a Cohen's d value of 5.312 and a confidence range of 3.902 to 6.713. The substantial effect sizes indicate that the intervention had a notable and enduring influence on performance overall assessment periods.

Performance of Experimental Group in Punctuation Tests

After the thorough analysis of the control group, the next step was to measure the progress of the experimental group from pretest to posttest and to delayed posttest; therefore, another Wilcoxon test was run to measure the effects of the treatment (i.e., using a chatbot to receive extra explanations and practice), and the results are shown in Table 5.

Table 5 indicates the presence of statistically significant differences in all the sets. The Z values for the comparisons between the posttest and pretest (Z = -4.787), the delayed posttest and pretest (Z = -4.788), and the delayed posttest and posttest (Z = -4.498) are all very significant, with p-values of 0.000. The findings suggest notable improvements in scores between the pretest and posttest, as well as between the pretest and delayed posttest. Additionally, there was a significant disparity between the posttest and delayed posttest scores. This suggests that the intervention led to significant and lasting improvements in performance over time. To further analyse the amount of difference within the experimental group in each set, Cohen's d was calculated, and Table 6 below presents the effect size results.

Table 6 illustrates substantial improvements in the experimental group across all testing intervals. The pretest yielded a Cohen's d value of 2.158, indicating a substantial initial effect. The 95% confidence interval for this effect is from 1.494 to 2.810. The posttest demonstrates a significantly greater effect size, with a Cohen's d value of 14.564 and a confidence interval ranging from 10.816 to 18.304, indicating a substantial and noteworthy improvement. The delayed posttest demonstrates a significant effect size, with a Cohen's d value of 13.482 and a confidence interval ranging from 10.009 to 16.946. The results underscore the substantial and enduring impact of the intervention on the experimental group's performance, surpassing the levels observed in the control group.

Comparison of Experimental and Control Groups in Punctuation Tests

Table 3

The Results of the Wilcoxon Test for the Control Group

	posttest - pretest	Delayed posttest - pretest	Delayed posttest - posttest
Z	-4.789	-4.237	-4.800
Asymp. Sig. (2-tailed)	.000	.000	.000

Table 4

The Effect Size in all the Tests in the Control Group

		Standardizer	Point Estimate	95% Confidence Interval	
				Lower	Upper
Pretest	Cohen's d	3.80018	2.158	1.494	2.810
posttest	Cohen's d	1.80325	9.816	7.274	12.349
delayed posttest	Cohen's d	2.10227	5.312	3.902	6.713

Table 5

The Results of the Wilcoxon Test for the Experimental Group

	posttest - pretest	Delayed posttest - pretest	Delayed posttest - posttest
Z	-4.787	-4.788	-4.498
Asymp. Sig. (2-tailed)	.000	.000	.000

After measuring learners' performance within their groups, it was necessary to compare the performance of each group across all sets of tests. Table 7 below presents the descriptive statistics for the control and experimental groups in the pretests, posttests, and delayed posttest.

Table 7 shows that the mean scores for the control group's pretest, posttest, and delayed posttest were 8.20, 17.70, and 11.16, respectively, while the mean scores for the experimental group's pretest, posttest, and delayed test were 8.06, 26.66, and 24.16, respectively. In both groups, as indicated by their mean scores, the average score of students on the posttest was higher than on the pretest and delayed posttest, and this comparison favours the experimental group. However, to better understand the differences between the groups, a Kruskal-Wallis Test was conducted, and the effect sizes of each set of tests are presented in Table 8.

Table 8 results reveal significant differences between the control and experimental groups' scores at different testing stages. For the pretest, the Kruskal-Wallis H value is 0.000 with a p-value of 1.000, indicating no significant difference between the groups and an epsilon-squared (ε^2) of -0.017, suggesting no effect size. However, for the posttest, the Kruskal-Wallis H value is 43.882 with a highly significant p-value ($p < 0.001$) and an ε^2 of 0.739, indicating a large effect size. Similarly, the delayed posttest has an H value of 44.628, with a highly significant p-value ($p < 0.001$) and an ε^2 of 0.752, signifying a large effect size. These results demonstrate that, although the groups started with similar performance levels, the intervention had a substantial and lasting positive impact on the experimental group's performance in both the posttest and the delayed posttest.

DISCUSSION

This study investigated the effectiveness of an interactive WhatsApp chatbot in enhancing the proficiency of Omani English as a Foreign Language (EFL) learners in using English punctuation marks. The findings demonstrated that the experimental group, which received supplementary training

Table 6

The Results of Effect Size within the Experimental Group in all Sets

		Standardizer	Point Estimate	95% Confidence Interval	
				Lower	Upper
pretest	Cohen's d	3.80018	2.158	1.494	2.810
posttest	Cohen's d	1.61352	14.564	10.816	18.304
delayed posttest	Cohen's d	1.54771	13.482	10.009	16.946

Table 7

The Descriptive Statistics for all the Tests in Two Groups

	N	Min	Max	Mean	Std. Deviation
Con_Pre	30	1.00	15.00	8.200	3.800
Con_Post	30	14.00	21.00	17.700	1.803
Con_Delayed	30	7.00	15.00	11.166	2.102
Exp_Pre	30	1.00	13.00	8.066	3.741
Exp_Post	30	23.00	30.00	26.666	1.971
Exp_Delayed	30	20.00	27.00	24.166	1.315
Valid N (listwise)	30				

Table 8

The Results of the Kruskal-Wallis Test for the Comparison of both Groups

	pretest	posttest	Delayed posttest
Kruskal-Wallis H	.000	43.882	44.628
df	1	1	1
Asymp. Sig.	1.000	.000	.000
ε2	-0.017	0.739	0.752

via the chatbot, significantly outperformed the control group across all testing phases, including the posttest and delayed posttest. This suggests that incorporating chatbot-assisted instruction into language education can yield measurable and lasting improvements in learners' mastery of specific linguistic features, such as punctuation. These results expand the breadth of current research by demonstrating the efficacy of chatbots in the specialised domain of punctuation, which is currently underrepresented in the AI-driven writing instruction literature. This finding strengthens the theoretical framework of form-focused education by validating that precise, targeted feedback delivered through technology may increase the cultivation of micro-level writing skills. Additionally, it aligns with sociocultural theory since the chatbot functioned as a scaffolding medium that promoted learners' development through interaction.

The observed progress within the control group can be attributed to both groups' classroom-based instruction. This finding aligns with the established pedagogical understanding that explicit instruction and consistent exposure lead to gradual improvement in form-focused language features such as punctuation (Fang & Wang, 2011; Macken-Horarik & Sandiford, 2016). However, the notably greater gains in the experimental group underscore the added value of chatbot integration as a mode of technology-enhanced learning. This result reinforces prior findings from studies highlighting the educational potential of mobile-based interventions, particularly in contexts that require frequent individualized practice and feedback (Abdul Fattah, 2015; Ivanova et al., 2022). The relationship between chatbot feedback types and learner engagement in form-focused training remains unclear. This study indirectly investigated this connection through its design. The chatbot intervention extends beyond traditional methods by employing constructivist ideas that enable learners to create their meaning through interaction, feedback, and adaptation. These are all essential parts of the personalized learning model.

The WhatsApp bot contributed to enhanced performance through several mechanisms. First, it facilitated immediate corrective feedback and individualised learning pathways based on learner performance, unlike the delayed and generalised feedback provided to the control group. Such responsiveness likely promoted greater learner engagement and more efficient error correction, factors emphasized in prior research on intelligent tutoring systems (Essel et al., 2022; Duong & Suppasetseree, 2024). Moreover, the chatbot's adaptive feedback and structured decision-tree logic correspond with key principles of practical formative assessment and scaffolding in second language acquisition. This demonstrates that mobile chatbots can serve as cost-effective and scalable alternatives to more complex AI teaching systems. This is an area that needs additional research in schools with limited resources. The effectiveness of these formative feedback loops reinforces the necessity of integrating AI-driven instructional designs into conventional English as a Foreign Language (EFL) teaching, especially in contexts where teacher-student ratios limit individualised attention.

Second, the chatbot afforded repeated and self-paced practice opportunities that are rarely possible in traditional classroom environments. The benefits of spaced repetition and self-regulated learning, both of which were operationalized in this study's intervention design, are well-documented in language pedagogy literature (Chen, 2024; Kwon et al., 2023). This automated interaction model enabled learners to gradually internalise complex punctuation rules through contextualised repetition. A subsequent study might benefit from a more comprehensive analysis of learner interaction logs to enhance comprehension of the cognitive mechanisms involved during chatbot usage. This strengthens the metacognitive dimension of second language acquisition, wherein learners cultivate awareness of their linguistic production and progressively improve their accuracy via autonomous engagement. This process is less achievable with teacher-centered instruction alone.

The findings further echo those of previous studies involving similar mobile-assisted writing interventions. For instance, Al-Wasy and Mahdi (2016) demonstrated that Saudi EFL learners using self-editing applications on mobile devices significantly improved in identifying punctuation errors. Likewise, the present study corroborates Vázquez-Cano et al. (2021), who showed that a Spanish-language chatbot significantly enhanced university students' punctuation performance. Additionally, Abdul Fattah (2015) reported positive outcomes from using WhatsApp-based instruction to enhance paragraph writing, with a particular focus on punctuation accuracy. Thus, the chatbot intervention employed here appears effective and consistent with international trends in mobile-mediated language instruction. Even so, the global studies mentioned show promising outcomes, but they differ in terms of technology interface, language focus, and assessment criteria. This makes it difficult to compare them and underscores the need for a standardised approach to evaluating AI-assisted writing therapies. This study fills a gap in the literature by providing data on a widely used platform (WhatsApp) and a well-defined linguistic subskill (punctuation), which improves the consistency of the evidence base for cross-study comparison.

In another study with similar findings, Črček and Patekar (2023) evaluated the frequency with which Croatian university students use ChatGPT for writing assignments, their performance, and their ethical perceptions. An online questionnaire completed by 201 university students from both state and private institutions gathered data. Evidence showed that more than half of the respondents use ChatGPT to create ideas, summarize, paraphrase, and proofread text; a high percentage also use it to write parts of their projects. The study identified the immediate need for institutions to establish clear ethical standards for the use of artificial in-

telligence software in scholarly activities. The focus on usage patterns and ethical considerations, while important, reveals a conceptual gap in the research on the influence of AI tools on learners' development of writing skills, an aspect our study addresses by emphasising educational outcomes rather than mere tool adoption. This study prioritizes pedagogical performance over utilitarian access, shifting the focus toward outcomes-based evaluation of AI integration in language education.

Additionally, the study's findings are consistent with those of Enamorado and Angel (2025), who found that the development of ChatGPT has generated significant excitement about its use in learning situations. The study investigated how 91 students of University of Valencia's English 2 course integrated the use of ChatGPT with English for Tourism Purposes (ETP) teaching. Targeted exercises, such as reading, vocabulary building, and role-plays, were conducted by students under the guidance of electronic devices and ChatGPT. The study centered on ChatGPT's learning efficacy, specifically its precision and ability to improve language skills. Through the presentation of interactive, feedback-driven exercises, ChatGPT clearly improves students' writing, communication, and vocabulary gains. Although output accuracy needs to be monitored by the teacher, the program's capability for emulating real-life communication situations in a short preparation time makes it an effective tool for directed and independent learning. Nonetheless, preliminary research in this emerging domain has pinpointed the specific linguistic or emotional attributes most affected by AI interaction, particularly in resource-constrained or high-stakes academic environments. This study fills that gap by giving controlled, skill-specific data on how chatbots might be used to help people learn English as a foreign language. Thus, it contributes to the current discussion by providing a context-sensitive, skill-specific framework that may aid future research on utilizing AI in EFL instruction in specific areas.

Importantly, the pedagogical value of such interventions lies in their performance outcomes and in promoting learner autonomy. The experimental group had access to learning tools outside the classroom, allowing for practice during otherwise idle moments, an approach in line with microlearning principles (Mohammed et al., 2018; Bruck et al., 2012). This kind of out-of-class support may be especially crucial for students in EFL contexts who lack exposure to English outside the academic setting. This study highlights the importance of low-barrier interventions that align with contemporary informal learning trends by situating chatbot utilization within learners' everyday digital practices. This has implications for the development of more ecologically valid language learning systems. This underscores that mobile-centric, feedback-rich microlearning facilitates learners in attaining academic objectives while fostering autonomous language development, a vital aim in modern language teaching.

Limitations

First, although the overall sample size (n = 60) was adequate for quantitative analysis, the absence of qualitative data limits its insight into learners' perceptions and experiences with the chatbot. Second, the study was conducted within a single institutional and cultural context, which may constrain generalizability. Lastly, while the chatbot focused on five punctuation marks selected based on curricular analysis, future studies may wish to include a broader range of punctuation features to further validate the tool's comprehensiveness.

Future research should consider expanding the participant pool to include learners from multiple institutions, varying levels of language proficiency, and diverse demographic backgrounds. It is also recommended that future studies employ mixed methods designs to capture both learning outcomes and user experiences in greater depth. Investigating the chatbot's efficacy across other domains of language learning (such as syntactic accuracy, paragraph cohesion, or pragmatic competence) would also be valuable. Additionally, integrating voice-based chatbot interaction could further personalize instruction and support learners with limited literacy skills.

CONCLUSION

This study has provided empirical evidence supporting the effectiveness of an interactive WhatsApp chatbot's effectiveness in improving English punctuation among Omani EFL learners. By integrating a process-oriented, learner-responsive tool into the learning environment, the study demonstrates how mobile-assisted technologies can enhance writing accuracy in a measurable and pedagogically meaningful way.

The experimental group, which received supplemental instruction through the chatbot, outperformed the control group in immediate and delayed posttests. These results indicate short-term learning gains and sustained retention of punctuation rules, suggesting that mobile-based interventions can support long-term language development. The chatbot's design (featuring individualized feedback, self-paced engagement, and decision-tree logic) proved an effective mechanism for reinforcing language conventions often neglected in traditional instruction.

The study contributes to the growing literature on mobile-assisted language learning (MALL) and offers a scalable solution for improving writing mechanics in resource-con-

strained or exam-oriented educational contexts. Nevertheless, certain limitations remain. The relatively small sample and lack of qualitative data limit the generalizability and explanatory depth of the findings. Future studies should consider larger, more diverse participant groups and incorporate learners' reflections and interaction logs to gain a deeper understanding of user experience and engagement patterns.

Ultimately, this study underscores the potential of integrating intelligent chatbot technologies into formal and informal language instruction. As digital platforms continue to expand their role in education, targeted applications like the one developed here offer promising avenues for personalised learning, particularly in areas such as punctuation, which often receive limited instructional attention.

DECLARATION OF COMPETING INTEREST

None declared.

AUTHORS' CONTRIBUTIONS

Ali Al Ghaithi: conceptualization; writing – original draft; software.

Behnam Behforouz: formal analysis, writing – review and editing.

REFERENCES

Abdul Fattah, S. F. E. S. (2015). The effectiveness of using WhatsApp messenger as a mobile learning technique to develop students` writing skills. *Journal of Education and Practice, 6*(32), 115-127. http://dx.doi.org/10.13140/RG.2.2.11492.83846

Al Ghaithi, A., Behforouz, B., & Isyaku, H. (2024). The effect of using WhatsApp bot on English vocabulary learning. *Turkish Online Journal of Distance Education (TOJDE), 25*(2), 13. https://doi.org/10.17718/tojde.1297285

Al Ghaithi, A., Behforouz, B., & Isyaku, H. (2025). Enhancing learners' spelling skills with an interactive chatbot facilitator. *International Journal of Interactive Mobile Technologies*, *19*(02), 79–93. https://doi.org/10.3991/ijim.v19i02.50531

Al-Wasay, B. Q., & Mahdi, H. S. (2016). The effect of mobile phone applications on improving EFL learners` self-editing. *Journal of Education and Human Development, 5*(3), 149-157. https://doi.org/10.15640/jehd.v5n3a16

Angelillo, J. (2002). *A fresh approach to teaching punctuation*. Scholastic Inc.

Behforouz, B., & Al Ghaithi, A. (2024). The effectiveness of an interactive WhatsApp bot on listening skills. *International Journal of Interactive Mobile Technologies*, *18*(02), 82–95. https://doi.org/10.3991/ijim.v18i02.44327

Bii, K. (2013). Chatbot technology: A possible means of unlocking student potential to learn how to learn. *Educational Research*, *4*(2), 218-221.

Bram, B. (1995). *Write well, improving writing skills*. Penerbit Kanisius.

Bruck, P. A., Motiwalla, L., & Foerster, F. (2012). Mobile learning with micro-content: A framework and evaluation. *BLED eConference, 25*, 527-543. https://aisel.aisnet.org/bled2012/2

Chen, M. R. A. (2024). The AI chatbot interaction for semantic learning: A collaborative notetaking approach with EFL students. *Language Learning & Technology, 28*(1), 1–25. https://hdl.handle.net/10125/73586

Chien, Y. C., Wu, T. T., Lai, C. H., & Huang, Y. M. (2022). Investigation of the influence of artificial intelligence markup language-based LINE Chatbot in contextual English learning. *Frontiers in Psychology, 13*, 1-8. https://doi.org/10.3389/fpsyg.2022.785752

Colace, F., De Santo, M., Lombardi, M., Pascale, F., Pietrosanto, A., & Lemma, S. (2018). Chatbot for e-learning: A case of study. *International Journal of Mechanical Engineering and Robotics Research*, *7*(5), 528-533. https://doi.org/10.18178/ijmerr.7.5.528-533

Črček, N., & Patekar, J. (2023). Writing with AI: University students' use of ChatGPT. *Journal of Language and Education, 9*(4), 128–138. https://doi.org/10.17323/jle.2023.17379

Crystal, D. (2013). *Spell it out: The curious, enthralling and extraordinary story of English spelling*. PROFILE Books.

Daffern, T., & Mackenzie, N. (2015). Building strong writers: Creating a balance between the authorial and secretarial elements of writing. *Literacy Learning: The Middle Years*, *23*(1), 23-32. https://doi.org/10.3316/aeipt.206520

Duong, T., & Suppasetseree, S. (2024). The effects of an artificial intelligence voice chatbot on improving Vietnamese undergraduate students' English speaking skills. I*nternational Journal of Learning, Teaching and Educational Research, 23*(3), 293-321. https://doi.org/10.26803/ijlter.23.3.15

Enamorado, M., & Ángel, J. (2025). Development of linguistic competence in English for specific purposes through ChatGPT: A case study. *Journal of Language and Education, 11*(1), 85–100. https://doi.org/10.17323/jle.2025.23745

Essel, H. B., Vlachopoulos, D., Tachie-Menson, A., Johnson, E. E., & Baah, P. K. (2022). The impact of a virtual teaching assistant (chatbot) on students' learning in Ghanaian higher education. *International Journal of Educational Technology in Higher Education*, *19*(57), 1-19. https://doi.org/10.1186/s41239-022-00362-6

Fang, Z., & Wang, Z. (2011). Beyond rubrics: Using functional language analysis to evaluate student writing. *The Australian Journal of Language and Literacy*, *34*(2), 147-165. https://doi.org/10.3316/ielapa.112929734219836

Ghose, S., & Barua, J. J. (2013, May 17-18). Toward the implementation of a topic-specific dialogue-based natural language chatbot as an undergraduate advisor. In *Proceedings of the 2013 International Conference on Informatics, Electronics and Vision (ICIEV)* (pp. 1–5). Dhaka, Bangladesh. https://doi.org/10.1109/ICIEV.2013.6572650

Han, S., & Lee, M. K. (2022). FAQ chatbot and inclusive learning in massive open online courses. *Computers & Education*, *179*, 1-13. https://doi.org/10.1016/J.COMPEDU.2021.104395

Healey, J. (2020). *Artificial intelligence*. The Spinney Press.

Hien, H. T., Cuong, P. N., Nam, L. N. H., Nhung, H. L. T. K., & Thang, L. D. (2018). Intelligent assistants in higher-education environments: The FIT-EBot, a chatbot for administrative and learning support. In *Proceedings of the 9th International Symposium on Information and Communication Technology* (pp. 69–76). Association for Computing Machinery. https://doi.org/10.1145/3287921.3287937

Ivanova, M., Arupova, N., & Mekeko, N. (2022). Digital support for teaching punctuation in academic writing in English. *Journal of Language and Education, 8*(3), 82-97. https://doi.org/10.17323/jle.2022.13608

Jan, L. W. (2009). *Write ways: Modelling writing forms* (3rd ed.). Oxford University Press.

Kamita, T., Ito, T., Matsumoto, A., Munakata, T., & Inoue, T. (2019). A chatbot system for mental healthcare based on SAT counseling method. *Mobile Information Systems,* 1-11. https://doi.org/10.1155/2019/9517321

Kerlyl, A., Hall, P., & Bull, S. (2006). Bringing chatbots into education: Towards natural language negotiation of open learner models. In R. Ellis, T. Allen, & A. Tuson (Eds.), *Applications and innovations in Intelligent Systems XIV* (pp. 179-192). Springer London. https://doi.org/10.1007/978-1-84628-666-7-14

Kim, N.-Y. (2018). A study on chatbots for developing Korean college students' English listening and reading skills. *Journal of Digital Convergence, 16*(8), 19-26. https://doi.org/10.14400/JDC.2018.16.8.019

Kwon, S. K., Shin, D., & Lee, Y. (2023). The application of chatbot as an L2 writing practice tool. *Language Learning & Technology, 27*(1), 1–19. https://doi.org/10.10125/73541

Kumar, J. A. (2021). Educational chatbots for project-based learning: Investigating learning outcomes for a team-based design course. *International Journal of Educational Technology in Higher Education*, *18*(1), 1-28. https://doi.org/10.1186/s41239-021-00302-w

Kurup, L., Narvekar, M., & Sasikumar, M. (2021). AI based tutoring system for English punctuation: E-VAKYA. *International Journal of Mechanical Engineering, 6*(3), 3708-3713. https://doi.org/10.5281/zenodo.5491655

Liew, T. W., Mat Zin, N. A., & Sahari, N. (2017). Exploring the affective, motivational and cognitive effects of pedagogical agent enthusiasm in a multimedia learning environment. *Human-Centric Computing and Information Sciences*, *7*(1), 1-21. https://doi.org/10.1186/s13673-017-0089-2

Lukeman, N. (2007). *A dash of style: The art and mastery of punctuation.* W. W. Northon & Company.

Macken-Horarik, M., & Sandiford, C. (2016). Diagnosing development: A grammatics for tracking student progress in narrative composition. *International Journal of Language Studies*, *10*(3), 61-94.

Mohammed, G. S., Wakil, K., & Nawroly, S. S. (2018). The effectiveness of microlearning to improve students' learning ability. *International Journal of Educational Research Review*, *3*(3), 32-38. https://doi.org/10.24331/ijere.415824

Okonkwo, C. W., & Ade-Ibijola, A. (2021). Chatbots applications in education: A systematic review. *Computers and Education: Artificial Intelligence*, *2*, 1-10. https://doi.org/10.1016/j.caeai.2021.100033

Pérez, J. Q., Daradoumis, T., & Puig, J. (2020). Rediscovering the use of chatbots in education: A systematic literature review. *Computer Applications in Engineering Education*, *28*(6), 1549-1565. https://doi.org/10.1002/CAE.22326

Roos, S. (2018). *Chatbots in education: A passing trend or a valuable pedagogical tool?* [Master's thesis, Uppsala University]. http://www.diva portal.org/smash/record.jsf?pid=diva2%3A1223692&dswid=-2674

Satam, S., Nimje, T., Shetty, S., & Kurle, S. (2020). Mentoring chatbot using artificial intelligence framework. *SAMRIDDHI: A Journal of Physical Sciences, Engineering and Technology*, *12*(1), 151-156.

Scull, J., & Mackenzie, N. (2018). Developing authorial skills: Child language leading to text construction, sentence construction and vocabulary development. In N. M. Mackenzie, & J. Scull (Eds.), Understanding and supporting young writers from birth to 8 (1 ed., pp. 89-115). Routledge. https://doi.org/10.4324/9781315561301-6

Shail, M. S. (2019). Using micro-learning on mobile applications to increase knowledge retention and work performance: A review of literature. *Cureus*, *11*(8), 1-7. https://doi.org/10.7759/cureus.5307

Schmidt, T., & Strasser, T. (2022). Artificial intelligence in foreign language learning and teaching: A CALL for intelligent practice. *Anglistik: International Journal of English Studies*, *33*(1), 165-184. https://doi.org/10.33675/ANGL/2022/1/14

Schumaker, J. B., Fisher, J. B., & Walsh, L. D. (2019). Effects of computerized instruction on the use of punctuation strategies by students with LD. *Learning Disabilities Research & Practice, 34*(3), 158–170. https://doi.org/10.1111/ldrp.12203

Song, B., & Xiong, D. (2023). A comparative study of the effects of social media and language learning apps on learners' vocabulary performance. Asia Pacific Education Review. https://doi.org/10.1007/s12564-023-09871-z

Subramaniam, N. K. (2019). Teaching & learning via chatbots with immersive and machine learning capabilities. In *Proceedings of the ICE 2019 Conference* (pp. 145–156). Jyväskylä, Finland.

Truss, L. (2003). *Eats, shoots & leaves: The zero tolerance approach to punctuation*. Gotham Books.

Vázquez-Cano, E. (2012). Mobile learning with Twitter to improve linguistic competence at secondary schools. *New Educational Review*, *29*(3), 134-147.

Vazquez-Cano, E. (2014). Mobile distance learning with smartphones and apps in higher education. *Educational Sciences: Theory and Practice*, *14*(4), 1505-1520. https://doi.org/ 10.12738/estp.2014.4.2012

Vázquez-Cano, E., González, A. I. H., & Sáez-López, J. M. (2018). An analysis of the orthographic errors found in university students' asynchronous digital writing. *Journal of Computing in Higher Education*, *31*(1), 1-20. https://doi.org/10.1007/s12528-018-9189-x

Vázquez-Cano, E., Mengual-Andrés, S., & López-Meneses, E. (2021). Chatbot to improve learning punctuation in Spanish and to enhance open and flexible learning environments. *International Journal of Educational Technology in Higher Education*, *18*(1), 1-20. https://doi.org/10.1186/s41239-021-00269-8

Wei, L. (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.2023.1261955

Wei-Xun, L., & Jia-Ying, Z. (2024). Impact of AI-driven language learning apps on vocabulary acquisition among English learners. *Research Studies in English Language Teaching and Learning, 2*(1), 1–11. https://doi.org/10.62583/rseltl.v2i1.32

Yang, J. (2022). Perceptions of preservice teachers on AI chatbots in English education. *International Journal of Internet, Broadcasting and Communication, 14*(1), 44-52. http://dx.doi.org/10.7236/IJIBC.2022.14.1.44

APPENDIX 1

The Pretest of Puncuation Marks

English Punctuation Pretest

General Instructions:

You are going to answer 30 questions in total, divided into three sections. Read each instruction carefully. You may not use a dictionary or mobile phone. Write clearly and review your answers before submitting the test.

Section 1: Sentence Correction (10 marks)

Instruction: Each sentence below contains one or more punctuation errors. Rewrite the sentence using correct punctuation.

Example:

✗ he said i am going to the store now

✓ He said, “I am going to the store now.”

1. what time is the class starting
2. I bought apples bananas oranges and grapes
3. its raining outside take your umbrella
4. My friend who lives in Muscat is visiting me today
5. I cant believe it she really passed the exam
6. I read the article titled the effects of climate change
7. Yes I agree with you
8. She said she would arrive by 5 pm
9. The movie was good however the ending was disappointing
10. Dr Smith will join us tomorrow for the lecture

Section 2: Multiple Choice (10 marks)

Instruction: Choose the punctuation mark that correctly completes the sentence. Circle the correct option (A, B, or C).

Example:

What did she say___

A. . B. ? C. !

✓ Correct answer: B

1. Do you like coffee ___
A. . B. ? C. !
2. I have three pets ___ a dog, a cat, and a parrot.
A. : B. ; C. ,
3. It’s raining outside ___ don’t forget your coat.
A. , B. . C. ;
4. She shouted ___ “Watch out!”
A. , B. : C. ;
5. The meeting is on Monday ___ May 5.
A. , B. - C. :
6. Mr ___ Al Saidi is our new teacher.
A. , B. . C. :
7. I have visited Muscat ___ Nizwa ___ and Salalah.
A. , / , B. ; / , C. , / and
8. “That’s amazing” ___ he said.
A. , B. ! C. .
9. She is a hard-working ___ honest student.
A. , B. ; C. -
10. They brought everything ___ plates, spoons, and cups.
A. : B. ; C. ,

Section 3: True/False Statements (10 marks)

Instruction: Read each sentence carefully. Decide whether the punctuation used is correct. Write 'True' if it is correct and 'False' if it is incorrect.

Example:

I can’t wait to see you!

✓ True

1. My uncle, who lives in Salalah is coming to visit.
2. Please bring the following items: pencils, erasers and rulers.
3. What a beautiful day it is.
4. “Are you ready to go” she asked.
5. He said, that he would be late.
6. Let’s meet at 6:00 pm.
7. No I haven’t seen your notebook.
8. He asked me if I liked pizza?
9. Tomorrow, we’ll visit the museum.
10. This is Johns book.

https://doi.org/10.17323/jle.2025.22731

The Influence of Multimodal Visual Methodologies on EFL University Students' Audio-Visual Comprehension, Verbal and Nonverbal Communication

Ibrahim Hassan Ali Al-jumaily [1], **Istabraq Tariq Jawaad Alazzawi** [2]

[1] University Of Anbar, Ramadi, Iraq
[2] University of Tikrit, Tikrit, Iraq

ABSTRACT

Background: In the last decades, there have been an increased use of multimodal teaching (text, images, audio, video) in EFL education, offering learners diverse input to suit different learning styles. However, how specific input types (e.g., dialects, registers, multimedia) affect learners with varying cognitive or cultural backgrounds remains unclear. Addressing this gap is essential for effective EFL instruction in today's multimodal learning environments.

Purpose: This study aims to examine how multimodal visual methodologies influence EFL students' development in audio-visual comprehension as well as verbal and nonverbal communication. By investigating these dimensions, the research seeks to fill a critical gap in the understanding of how integrated sensory modalities shape communicative competence in technologically enhanced learning environments. Furthermore, the study explores underlying psychological, social, and pedagogical factors that facilitate or hinder these outcomes, thereby providing tangible insights for designing more cognitively effective EFL curricula.

Method: The study's sample consists of 214 EFL university students. A mixed-mode descriptive research design was used. Audio-visual comprehension, verbal and nonverbal communication tests served as the quantitative data collection instruments. Repeated-measures ANOVA and t-test paired samples were conducted on the set of three test scores over time. In the qualitative phase, data were collected from 20 purposively selected students by using a semi-structured focus group interview and was analysed qualitatively based on thematic analysis.

Results: The findings demonstrated statistically significant improvements in students' audio-visual comprehension, verbal communication, and nonverbal communication across three repeated-measures assessments. For example, mean scores for audio-visual comprehension increased from 46.22 to 67.59 ($p < .001$, $\eta^2 = .535$). Similar gains were observed in verbal ($\eta^2 = .561$) and nonverbal communication ($\eta^2 = .559$). Qualitative data confirmed that students perceived the multimodal learning environment as psychologically engaging and socially supportive, highlighting increased mental readiness, reduced listening anxiety, and improved interpretation of body language.

Conclusion: The study highlights the critical role of linguistic and psychological factors in EFL, particularly emphasising the significance of audio-visual comprehension and both verbal and nonverbal communication. It also identifies three key variables (i.e., audio-visual comprehension, verbal, and nonverbal communication) that aid EFL students in enhancing their retention, long-term memory and confidence for independent English learning.

KEYWORDS

multimodal learning, EFL university students, audio-visual comprehension, verbal and nonverbal communication, subtitled video input, psychological factors in language learning

Citation: Al-jumaily I. H. A., & Alazzawi I. T. J. (2025). The Influence of multimodal visual methodologies on EFL university students' audio-visual comprehension, verbal and nonverbal communication. *Journal of Language and Education, 11*(2), 35-56. https://doi.org/10.17323/jle.2025.22731

Correspondence:
Ibrahim Hassan Ali Al-jumaily,
ibrahim.hasan@uoanbar.edu.iq

Received: July 19, 2023
Accepted: June 10, 2025
Published: June 30, 2025

INTRODUCTION

In today's rapidly evolving digital landscape, the shift from traditional print-based media to digital platforms has transformed the ways in which discourse is practiced. This evolution reflects the concept of multimodal communication, in which meaning co-constructed based on interaction among different modes comprising text, image, sound, gesture and more (AbdulGhafoor & Challob, 2021). Multimodality, as a theoretical framework, explores how these modes work together under sociocultural influences (Jewitt & Kress, 2003; Norte Fernández-Pacheco, 2018). In education, multimodal ensembles—combinations of different modes—enhance learning by integrating visual, auditory, and kinesthetic elements (Love, 2019; Halliday & Hasan, 1985). In multimodal education, ensembles reflect a combination of modes that improve teaching and learning and articulate visual, aural and kinetic elements (Norte Fernández-Pacheco, 2018; Heimbürger, 2013). Multimodal approaches in EFL education have been revolutionizing language teaching by stepping away from text-based approaches and incorporating visual, auditory, as well as kinesthetic aspects, making intensive learning (Purwaningtyas, 2020; Ghoushchi et al., 2021). This is in accordance with Communicative Language Teaching (CLT) concepts which emphasize real and contextual communication (Richards & Rodgers, 2014). Utilizing images, videos, gestures, spatial arrangements, and appealing to a variety of senses supports learning and the ability to understand and produce language in natural settings (Van Leeuwen, 2020).

One of the key components of multimodal learning in EFL education is audio-visual comprehension, a core aspect of multimodal EFL learning, involves interpreting auditory and visual stimuli like videos, podcasts, and multimedia. It enhances real-life communication skills by exposing learners to authentic pronunciation, intonation, and nonverbal cues (Yi et al., 2024; Yeh, 2022). Such materials improve listening and speaking abilities while boosting engagement through contextualized, immersive scenarios (Chiriac, 2025; Putri et al., 2024). Research underscores their value in aiding comprehension via visual context and nonverbal decoding, crucial for EFL learners navigating cultural and linguistic nuances (Vandergrift & Goh, 2012; ED-DALI, 2024). However, challenges persist with fast-paced audio or lack of visual support (Huang, 2006; Fitria, 2024). Technology's role is crucial but needs to be mediated by the teacher to enhance the potential of multimodal tools for language learning. Furthermore, verbal communication (including the spoken and written word) is an indispensable vehicle through which messages can be delivered, thoughts can be expressed and interaction may be achieved. As pointed out by Ruswandi & Arief (2024) and Nofali & Gasim (2024), oral practice is a must when aiming to gain both fluency and accuracy, since it involves learners in active production of language. This active engagement is further sustained by conversational interaction that elicited feedback and clarification that support a dynamic learning setting (Ruswandi & Arief, 2024; Adhitya & Valiansyah, 2024). These interactions not only improve students' language proficiency but also install a sense of self-assurance, which is an important element in successful communication (Nofali & Gasim, 2024).

However, verbal communication does not function independently; it is deeply intertwined with nonverbal cues that shape meaning. For instance, a teacher's tone of voice can profoundly influence message reception (Abdikarimova et al., 2021; Sutiyatno, 2018; McDuffie et al., 2021). This coordination of verbal and nonverbal elements emphasizes the advantage of multi-modal approaches where various sensory inputs are combined to aid in understanding and comprehension. Studies have shown that video-based tasks (such as role- plays, simulations) introduce students to authentic use of language, assisting them to mimic the pronunciation and discourse patterns of native speakers (Hahl et al., 2025). Furthermore, visual learning tools like infographics and mind maps provide students with more coherence to structure and structure ideas synthesize and organize using a text, which connects the understanding with production.

It also emphasizes the significant role of nonverbal communication, such as gestures, expressions, eye contact, body volume and distance in enriching interaction and in backing verbal communication especially to EFL contexts. Because these features are cultural-bound, multimodal strategies (for example, video analysis and interactive activities) can be used to ecode and practise them (Pratolo, 2019; Ruswandi & Arief, 2024). Teachers may send non-verbal cues to help confirm the message, promote inclusiveness, and increase student engagement (Haneef et al., 2014; Burgoon, 2003; Bambaeeroo & Shokrpour, 2017). Studies have confirmed that facial expressions were one of the most influential nonverbal approaches, and verbal and non-verbal communications have a positive impact on the academic performance (Mikhael et al., 2022; Sutiyatno, 2018). Moreover, adjusting communication styles to cultural situation is important in EFL to avoid misconceptions and enhance the general communicative competence (Tussupova et al., 2017). The incorporation of nonverbal communication into language instruction encourages a more holistic understanding and genuine cross-cultural communication between individuals.

The integration of audio-visual comprehension, verbal, and nonverbal communication is essential because communication is inherently multimodal. In the real world, these modes are interconnected in which nonverbal signs frequently present themselves in support or supplement of verbal messages (Kress & Van Leeuwen, 2001; Ruswandi & Arief, 2024). For EFL learners, a combined use of the three modes is necessary for effective communication because auditory

and visual information presented together facilitates comprehension (Sueyoshi & Hardison, 2005). To concentrate on just one mode would be to get only a partial picture—one which ignores the oscillating relationships between modes, as well as their aggregate, holistic function in language acquisition. Here, of course, multimodal technologies including Smart Boards, projectors, Vocast (for audio-visual comprehension) and visual tools for verbal, nonverbal cues and the like would improve the EFL students' understanding and communication. They offer interactive, standardized and measurable way to evaluate progress made in listening, speaking, and body language which provides data driven evaluation of multimedia language teaching (Olelewe et al., 2023).

Despite the growing body of research supporting the effectiveness of multimodal methodologies in EFL education, there remain gaps in the literature that warrant further investigation. While many studies have addressed individual components of multimodal input, few have examined their integrative effect on communicative competence as a multidimensional construct, such as listening and speaking, there is a need for more comprehensive research that examines the interplay between audio-visual comprehension, verbal, and nonverbal communication. What remains insufficiently studied is how these three domains interact over time within a multimodally enriched EFL environment, and how learners' engagement and psychological readiness mediate this process. Addressing these gaps is essential for developing a more nuanced understanding of how multimodal methodologies can be tailored to meet the diverse needs of EFL learners.

On the basis of the above concepts, this study aims to examine how multimodal visual methodologies influence EFL students' development in audio-visual comprehension as well as verbal and nonverbal communication. By investigating these dimensions, the research seeks to fill a critical gap in the understanding of how integrated sensory modalities shape communicative competence in technologically enhanced learning environments. Furthermore, the study explores underlying psychological, social, and pedagogical factors that facilitate or hinder these outcomes, thereby providing actionable insights for designing more inclusive and cognitively effective EFL curricula. The study seeks to answer the following research questions:

RQ1: What is the influence of multimodal visual methodologies on EFL students' audio-visual comprehension, verbal and nonverbal communication among three rounds of testing?

RQ2: How can multimodal visual methodologies influence EFL students' audio-visual comprehension, verbal communication and nonverbal communication?

LITERATURE REVIEW

Multimodal Approach: The Integration of Diverse Modes in Educational Practices

The underlying theoretical basis for multimodal practices lies in socio-cultural theory (Vygotsky, 1978), which argues that cognitive development, is mediated by social interaction and culture tools. From this view, language development is a socially situated activity, involving shared activities with more knowledgeable others and the manipulation of cultural tools (tools in the broadest sense, such as texts, graphics, and technologies). A multimodal theory This principle is realised in practice as multimodal approaches that combine a range of cultural tools (i.e. visual, auditory and kinesthetic resources) to mediate reflection and learning. However, the possibility that learners may be differentially proficient in combining the two modes is not considered in the current investigation. And for some students processing too much audio, visual and kinesthetic input simultaneously can create cognitive dissonance and detract from rather than facilitate understanding.

Furthermore, these approaches are consistent with cognitive load theory (Sweller, 1988) presenting that when the information load is distributed over several sensory modalities, the efficiency of learning is increased. Multimodal technologies mitigate cognitive overload through multiple modes of representation (visual and auditory input) and learners can process linguistic input more effectively. Taken together, these theoretical perspectives provide an underpinning for the integration of multimedia in EFL teaching by suggesting its double significance in enabling social interaction and cognitive attention. These results support the emphasis in the present study on the ways in which multimodal scaffolding (e.g., subtitles, images, gestures) can support EFL learners' retention and comprehension of auditory and visual input.

According to multimodal academics, modern learners create meaning by applying several modes (written and spoken language, gesture, visuals, sound, and movement) in a fast-changing, diverse era filled with numerous semiotic and digital resources in their everyday lives (Jewitt & Kress, 2003; New London Group, 1996). This perspective underscores that knowledge is inherently multimodal, co-constructed, and performed or represented (Miller, 2007, p. 65). Consequently, students' ability to consume, understand, and produce multimodal literacies is crucial for both academic and social purposes (Jewitt & Kress, 2003; Yi, 2014). Multimodal approaches in education present content through visual, aural, and written formats, which can enhance learning performance. Research suggests that using various presenta-

tion modes makes learning easier and increases attention, particularly for lower-achieving students (Chen & Fu, 2003; Farías et al., 2014; Pintado & Fajardo, 2021). These findings align with the present study's focus on multimodal instruction as a tool for educational equity, as they demonstrate that differentiated presentation methods can reduce performance gaps. However, these studies primarily examine short-term engagement rather than long-term knowledge retention, which is a key concern in the current research.

Based on this idea, Mayer (2014) proposed the *multimedia effect*, which states that students learn more when words are accompanied by pictures than when words appear alone. This confirms the claim that the multichannel input facilitates understanding and long-term retention of the target language, an important factor at one of the EFL settings, where learners need to process information from number of sensory channels. Nevertheless, Mayer's research is based on cognitive processing in mainstream education that does not take into account the kind of language and culture difficulties experienced by EFL learners in their interpretation of modal information. Gilakjani et al. (2011) who claimed that visualizations could support learning by giving students an external representation of information, thus impelling deep cognitive processing and helping to maintain attention. Graphs and charts make information interesting and enticing and help simplify complex topics. These results are consistent with the focus of the current study on the potential of audio-visual stimuli to scaffold EFL learners' comprehension, specifically to lower cognitive load for verbal input processing. However, the research has not directly focused on non-verbal communication (gestures, visual symbolism), which is essential for EFL students to comprehend meaning in multimodal text. Kress (2010) further emphasized the importance of multimodal integration, arguing that it is now impossible to fully understand texts—even their linguistic components—without considering how other features, such as images and sounds, contribute to meaning. This reinforces the centrality of multimodal literacy in contemporary communication, particularly in digitally mediated EFL environments. This interconnectedness is amplified by computer technology, which seamlessly integrates text, audio, video, and images in meaning-making processes (Kress & van Leeuwen, 2006). Kress and van Leeuwen (2006) also introduced a visual design «grammar,» exploring how visual literacy can be integrated into education. This is particularly relevant to the present study's aim of developing visual methodologies for EFL instruction, as it offers tools to deconstruct visual rhetoric in learning materials. That said, their model primarily examines Western visual conventions, potentially marginalizing culturally diverse EFL learners who may interpret imagery differently.

A multimodal approach in the classroom emphasizes the strategic use of multiple modalities in authentic learning environments. Each modality serves as a resource for students' meaning construction, offering unique perspectives on phenomena that can challenge prior conceptions and provide tools for imagining and thinking (Kress, 2009). For instance, teachers often use gestures and voice to highlight images and other references during instruction, creating a dynamic interplay between modes (Kress, 2009). Kress (2010) also argued that gestures, drawings, voice, and physical objects interact in ways that enrich meaning construction. Each mode contributes uniquely: speech provides differentiation, blackboard images offer visual context, object manipulation creates a physical context, actions provide dynamic clarity, and textbook images serve as stable summaries. Repetition, synchronisation, similarity, and contrast further enhance cohesion. The selection of modes is purposeful, as each metaphorical journey is unique, and each mode builds meaning differently. Students must engage in distinct tasks to comprehend each mode, fostering deeper understanding. The dynamic interplay between modes (e.g., voice and gestures reinforcing images) aligns with the present study's focus on how synchronized multimodal input strengthens learners' ability to decode verbal and nonverbal cues in audio-visual materials.

Multimodal approaches utilize both verbal and nonverbal modes to represent content knowledge, thereby enhancing visual and sensory education. This approach not only promotes relevance and innovation but also improves course quality and diversifies academic programs (Maguire, 2005; Moreno & Mayer, 2007). By integrating multiple modes, educators can create richer, more engaging learning experiences that cater to diverse learner needs and preferences. This supports the argument that multimodal methodologies are crucial for fostering deeper comprehension and engagement in EFL contexts, as they leverage both auditory and visual stimuli to reinforce learning. However, these studies overlook the specific challenges EFL learners may face when processing multiple modes simultaneously, such as cognitive overload or cultural differences in interpreting nonverbal cues. Further research is needed to explore how different learner profiles (e.g., proficiency levels, cultural backgrounds) interact with multimodal input.

Multimodal Audio-Visual Comprehension

The progress in science and technology is transforming the landscape of second language (L2) listening teaching, which has evolved from the old-fashioned audio-only based instruction to modern multimodal audio-visual instruction. There's an essential emphasis on technology for both instruction and assessment. There is interactive content, real-time annotation and gesture-based interaction with Smart Boards, multitouches and data show tools (projectors/ visualizers) normalize visual oriented stimuli (multimodality). Taken together, these technologies offer a systematic and comprehensive model to investigate how multimodal approaches improve various communication skills among EFL students (Abdullah et al., 2020). This change has yielded a proliferation of academic literature on the effectiveness

of these techniques[1] (Namaziandost & Nasri, 2019; Arbab, 2020). Audio-visual input, with motion picture, sound, and text on a screen, appeals to vision and hearing, and has three metafunctions: image, writing and doing. Numerous studies have confirmed that information to both auditory and visual modalities is not only better received, but it also is better encoded in memory than that delivered to the auditory modality only (Surguladze et al., 2001; Campbell, 2008). Since listening comprehension is among other complex selection and information processing activities of the brain and higher cortical function is facilitated during the activation of several sensory channels. The onset of multimedia in classrooms has prompted an additional revolution in the teaching industry – the ability to make a simultaneous assessment of both auditive and visual information. These developments have now been supplemented by large-scale standardized high-stakes testing, such as TOEFL's iBT and China's CET-4 and CET-6 Internet-based exams (Wang at al., 2014). This supports the argument that multimodal input enhances cognitive processing by engaging multiple sensory pathways, which is critical for EFL learners who rely on both verbal and visual cues for comprehension. These studies do not account for longitudinal changes in multimodal comprehension. However, these studies overlook the specific role of nonverbal communication (e.g., gestures, facial expressions) in facilitating language acquisition, which is a key component of multimodal learning.

However, a need still persists for empirical research that focuses on the effects that technological tools and computer-based or digital materials of/the visual and auditory input combined resources may have on students' comprehension. Yet, many studies have also emphasized the role of both linguistic and extra-linguistic information in processing. For example, Sueyoshi and Hardison (2005) highlighted the importance of context-specific cues (e.g., facial expressions and gestures) for comprehension. This supports the argument that nonverbal elements in multimodal input are critical for comprehension, aligning with the present study's focus on how visual methodologies augment EFL learners' interpretive skills. However, this study overlooks the specific technological affordances (e.g., interactive digital platforms) that could further optimize these nonverbal cues for language learning.

Similarly, Ramírez and Alonso (2007) demonstrated that Spanish children achieved a deeper understanding of foreign language structures through engagement with English digital stories. Their findings also revealed that such audio-visual materials not only improved listening skills but also fostered other competencies, such as communication, as students were able to provide feedback in the target language after viewing the videos. These findings align with the present study's focus on multimodal input enhancing both comprehension and verbal communication. However, the study does not explore whether certain types of visual stimuli (e.g., animations vs. live-action videos) yield differential effects, a gap this research could address. Wagner (2010) further corroborated these findings through a comparative study, showing that an experimental group exposed to audio-visual input outperformed a control group that relied solely on auditory input in comprehension tests. This highlights the need for multimodal approaches in EFL Learning, notably for its role in retention. However, the study confined attention to comprehension scores and does not include the impact on nonverbal communication (e.g., gesture understanding) which is a central aspect of this study. Taken together, these studies highlight the multi-dimensional contributions of multimodal input to second language learning, and argue that complementary visual and auditory information provides a substantial boost to both understanding and remembering.

Multimodal Approach to Verbal and Nonverbal Communication

Verbal and visual communication are essential components of the learning process, each playing a distinct yet complementary role. In fact, visual communication enriches the active learning by involving learners in auditory and visual representations as well as promotes higher cognitive functions and application of learning in the real world (Khadimally, 2016). This is consistent with the claim that multimodal visual approaches can help to enhance EFL students' understanding of abstract concepts through the illustration of concrete visual materials. But what this research has failed to consider is the special difficulties which existence for EFL learner in dealing with reception of visual and verbal information at the same time. Because this is the foundation of Audio-visual comprehension. This can be clearly seen in the context of foreign language learning when images are used, such as in the 'Selfie Project', to prompt spoken or written language production (Victoria, 2021). These findings align with the present study's focus on how visual aids can enhance verbal production in EFL contexts, suggesting that multimodal tasks (e.g., digital storytelling or image-based reflections) may foster both linguistic and nonverbal expression. However, the limitations of verbal communication in diverse and complex contexts highlight the necessity of visual aids to improve understanding and engagement (Lennartsson, 2010). This reinforces the idea that EFL learners may rely heavily on visual scaffolding to decode meaning, particularly in linguistically demanding situations. Yet, the study does not explore how cultural differences in visual literacy might affect comprehension—a gap this research could address. Despite the predominance of verbal communication in education, the effective use of visual thinking is crucial for conveying messages and foster-

[1] Zhyrun, I. (2016). Culture through comparison: Creating audio-visual listening materials for a CLIL course. *American Journal of Content and Language Integrated Learning, 9*(2)*,* 345–373. http://dx.doi.org/10.5294/laclil.2016.9.2.5

ing critical thinking skills (Nuzzaci, 2019). This supports the argument that multimodal methodologies could empower EFL students to analyze and articulate ideas more critically. However, the study focuses on general education rather than language learners, leaving room for further investigation into how visual thinking strategies specifically benefit EFL audio-visual processing. This interplay between verbal and nonverbal communication significantly influences the quality of learning experiences, underscoring the need for educators to develop proficiency in both forms (Wahyuni, 2018). These findings align with the present study's focus on multimodal pedagogy but do not examine the role of technology (e.g., videos, interactive graphics) in facilitating this interplay—an area this research could explore.

Moreover, having established the cognitive advantages of multimodal input, the following section examines its role in productive communication skills. Multimodal approaches to communication encompass both verbal and nonverbal elements, emphasizing their interdependence in conveying meaning. Research indicates that nonverbal cues, such as gestures, facial expressions, and prosody, play a crucial role in enhancing the understanding of verbal messages, particularly in contexts such as political discourse and media communication (Madella et al., 2023; Abduraximova, 2024; Harutyunyan, 2023). The evolution of multimodal studies has further expanded to include various media forms, highlighting the significance of visual components in enhancing the impact of verbal texts (Szawerna, 2023). This supports the argument that integrating visual and verbal modalities strengthens communicative effectiveness, which is crucial for EFL students developing audio-visual comprehension. However, this study overlooks the specific challenges EFL learners may face when processing multimodal input, such as cognitive overload or cultural differences in visual interpretation. These findings align with the present study's focus on how multimodal methodologies influence both verbal and nonverbal communication in EFL contexts. The neuroscientific perspective offered by Benetti et al. (2023) reinforces the idea that multimodal learning engages multiple cognitive processes, potentially enhancing retention and comprehension. For instance, studies show that learners who engage in multimodal communication experience fewer communication breakdowns and demonstrate improved language abilities (Bouchey et al., 2021).

In contrast to receptive modalities, productive and interactive modes such as gesture and prosody have received comparatively less attention. The multimodal approach is particularly relevant for students who are already accustomed to technology as part of daily life (Ngongo, 2022). This supports the argument that EFL learners, many of whom are digital natives, may benefit from instructional methods that align with their existing technological literacy, thereby facilitating engagement with multimodal texts. Multimodality refers to all verbal and nonverbal visual semiotic inputs that can be used to interpret dialogical associations in reading material (Herman et al., 2019). These findings align with the present study's focus on how visual and auditory semiotic resources interact to shape comprehension and communication in EFL contexts. As Baldry et al. (2020, p. 157) assert, "we live in a multimodal society," where individuals experience the world multimodality and construct meaning using words, visuals, gestures, sounds, and other resources. This perspective reinforces the idea that language learning should extend beyond traditional verbal instruction to incorporate diverse semiotic modes, which is central to this study's investigation of audio-visual comprehension. This perspective is supported by Bilfaqih and Qomarudin (2017), who argue that texts of various types are inherently multimodal, combining semiotic frameworks to create meaning in both standard and innovative ways. However, this study overlooks the specific challenges EFL learners may face when navigating multimodal texts, such as cognitive overload or cultural disparities in semiotic interpretation - a gap the current research seeks to address. Technological advancements further enhance text multimodality by simplifying the creation and dissemination of multimodal content, allowing learners to engage with diverse semiotic resources as a cohesive communication unit (Bezemer & Kress, 2016). This supports the argument that digital tools can scaffold EFL students' multimodal literacy, but it raises questions about whether current pedagogical practices adequately prepare learners to critically analyse and produce multimodal discourse, a key concern of this study.

In this environment, learners are placed in the role of "semiotic initiators and responders," not just attending to spoken language, but to producing and responding to, a range of texts and images as well as other multimodal resources (Coffin & Donohue, 2014). Frequently this creative mediation consent new readings or uses of the multimodal content that do not necessarily correspond to the original intentions of its design. Multimodal approaches in spoken and visual communication integrate linguistic (words), visual (gestures), and aural resources, demonstrating the interdependence of verbal and nonverbal communication. This is particularly evident in face-to-face communication, where gestures, facial expressions and prosody (intonation) contribute to the meaning and emotion of an utterance (Szawerna, 2023; Kenzhegaliev et al., 2023). Multimodal teaching, in the end, applies textual, visual, and auditory modes to enhance learning and equip students with powerful communication skills (Archvadze, 2023). While these studies highlight the benefits of multimodality, they do not sufficiently address potential challenges, such as cognitive overload in learners when processing multiple modes simultaneously or cultural differences in interpreting nonverbal cues. Additionally, there is limited discussion on how digital versus

in-person multimodal interactions affect learning outcomes differently.

METHOD

Research Design

This study implemented a mixed-methods research design to answer its research questions. Proudfoot (2023) defines mixed methods research as the combination of both qualitative and quantitative threads as part of the same research program (although data analysis/ and or collection may be done separately and findings and inferences may be combined/ integrated through the integration of data. It seeks to establish an in-depth understanding of complex educational phenomena through the collection, analysis and synthesis of a range of data in relation to particular research questions (Creswell, 2024). The method can be particularly advantageous in educational research, as i t could contribute to the consistency of results and lead to a more informed viewpoint on topics, such as teaching practices and students' well-being (Fàbregues et al., 2024; Ercan et al., 2022). An explanatory sequential mixed methods design 'consists of first collecting quantitative data and then collecting qualitative data to help explain or elaborate on the quantitative results' (Creswell, 2017, p.542). The research was conducted according to an explanatory sequential research design, in which the findings obtained with quantitative data are examined in depth with qualitative methods and data. In this context, firstly, quantitative data were collected and analysed by using a repeated-measures ANOVA on the set of three test scores over time. Secondly, qualitative data collection processes and analysis were applied by using thematic analysis of the data collected by a focus group semi-structured-interview to enrich the findings obtained in the first step.

Participants

The study population of the quantitative method consists of randomly selected third-year EFL university students at Anbar University, College of Education for Humanities, and Tikrit University, College of Education for Humanities and College of Education for Women enrolled in the English departments, during the academic year 2023–2024. The total population consists of 373 students. In addition to this population, the researcher randomly selected and invited 214 EFL Iraqi students to participate in the study on the basis of their willingness and agreement. However, some students were excluded due to failure to comply with the study period (including withdrawal during the study), incomplete responses, or non-compliance with test instructions, such as skipping sections or providing inconsistent responses. Only participants who fully adhered to the study protocols were included in the final analysis. The research's quantitative data method was determined by convenience sampling. Convenience sampling was selected to identify the nearest and easily accessible sample that the researcher can obtain. Convenience sampling was used due to logistical constraints (e.g., accessibility during the academic year). To mitigate bias, the sample was stratified by university (Anbar and Tikrit) and department (Humanities/Women), ensuring representation across institutional contexts. The researcher asked 214 students, who had the same educational and language background knowledge, to represent the sample by answering audio-visual comprehension, verbal communication and nonverbal communication tests.

The qualitative study sample consists of certain people and was determined through maximum variation sampling, one of the methods included in purposive sampling. Maximum diversity sampling aims to select a relatively small yet highly representative sample that captures the broadest possible range of perspectives within relevant population groups. This approach seeks to identify commonalities across diverse scenarios, thereby revealing the multifaceted nature of the issue under study (Yıldırım & Şimşek, 2006). In this study, maximum variation was achieved by incorporating participants with differing academic performance levels, age ranges, and degrees of digital literacy. Volunteering is crucial because the participants of the qualitative research will be drawn from the participants in the quantitative method (Creswell, 2017). For this reason, 20 students among

Table 1

The Demographic Data of the Study Participant of the Quantitative Dimension

Variable	Introductory Features	N	%
Sample	College of education for humanities- university of Tikrit	87	40.65%
	College of education for women- university of Tikrit	54	25.23%
	College of education for humanities- Anbar university	73	34.11%
Gender	Female	153	71.49%
	Male	61	28.50%
Age	19-21	178	83.17%
	22-24	36	16.21%

the participants who supported the quantitative part were selected for the qualitative sample with the focus group semi-structured interview. Data were collected from three colleges to form the quantitative sample. In the qualitative study group, at least five students from each college were included to ensure maximum diversity. The gender breakdown of the interviewed students is 13 females and 7 males, with ages ranging from 19 to 22.

Instruments of Data Collection

Data collection was conducted using triangulation of mixed-mode design instruments, as outlined below.

Quantitative Research Instrument

The study employed three quantitative data collection instruments. The first, the Audio-Visual Comprehension Test, measured listening comprehension and inferential understanding through a modified version of Norte Fernández-Pacheco's (2018) test. It included eight fill-in-the-blank items requiring exact words from the vodcast, six true/false questions on explicit facts, and one open-ended follow-up question analyzing speaker tone. The vodcast was selected for its authentic native-speaker dialogue, cultural neutrality, and alignment with B1 CEFR proficiency, with piloting (n=20) confirming appropriate difficulty after minor lexical adjustments. The second, the verbal communication test, evaluated written production via four tasks: five prompt-based dialogue continuations, one role-play scenario, one dialogue-writing task using rejoinders, and one structured topic-development exercise (renamed from «mental visualisation» for clarity). The third, the nonverbal communication test, assessed the identification and interpretation of gestures, facial expressions, and proxemics through 12 matching items (pairing images to nonverbal types), 10 multiple-choice questions (interpreting body language), and three short-answer analyses of emotional tone. All tests were refined for formal tone, replacing colloquialisms, adding exemplar responses, and standardizing scoring, to ensure replicability and alignment with language assessment best practices.

The second and third tests were collected and guided by the researcher. The repeated test design is used to test the effect, as indicated by Wiklund-Hörnqvist et al. (2014). It improves the ability to recall and learning outcomes at a significantly higher level. This method involves administering tests after particular periods and is more efficient than actions such as repeating.

To establish the face validity and content validity of the audio-visual, verbal and nonverbal communication tests, the researcher evaluated them with the help of a panel of 10 experts in the field of applied linguistics and methods of teaching. The experts' modifications and suggestions were taken into consideration in verifying the appropriateness and the final structure of the tests. Finally, the three tests were piloted to ensure the construct validity, language clarity, accuracy, practicality and reliability. According to Oluwatayo (2012), construct validity refers to whether the operational definition of a variable actually reflects the theoretical meanings of a concept. In other words, construct validity shows the degree to which inferences are legitimately made from the operationalisations in one's study to the theoretical constructs on which those operationalisations are based. Therefore, construct validity of all the instruments is established through item analysis such as item discrimination power, item difficulty level, the correlation coefficient between the item score and the total score of the test, the correlation coefficient between the item score and the component to which the item belongs, and internal correlation matrices. The results indicate that the audio-visual test items exhibited an appropriate level of difficulty, ranging between 0.41 and 0.67, while the discrimination power for all items fell within 0.31 to 0.53. For verbal communication, the difficulty level ranged from 0.42 to 0.58, with a discrimination power between 0.35 and 0.55. Similarly, nonverbal communication items demonstrated a difficulty level of 0.33 to 0.46 and a higher discrimination power of 0.43 to 0.63, which aligns with established benchmarks (Tang & Logonnathan, 2016).

Accordingly, the point-biserial correlation coefficient formula is used to calculate the correlation between the total score of the test, the binary score (intermittent) is used for the subject items and the Pearson correlation coefficient is used to find out the correlation between the items' scores and the total score for the 214 participants. Statistical analysis of audio-visual, nonverbal and verbal communication of the test's items reveals that all the correlation coefficient values are more significant than the critical value, which is 0.195 at 212 degrees of freedom and 0.05 significance level.

According to Franzen (2013, p. 15), one of the basic characteristics of a good instrument is reliability, which refers to the consistency or stability of scores values that an instrument elicits. The measurement of test–retest means that if the same respondents complete a test at two different points in time, then the responses should be stable and the set of results should be reproducible. With the use of the Pearson correlation coefficient to estimate reliability, or r-value, of responses, the r-value for the audio-visual comprehension scale is 0.91, that for the verbal communication scale is 0.87 and that for the nonverbal communication scale is 0.89, thus being indicators of good reliability because the values are higher than 0.70. Internal consistency of measure items is another way to test the reliability of this study. Cronbach's alpha is extensively used to determine a test's internal consistency (Franzen, 2013). Most internal consistency testing approaches treat each item as a separate measurement and the test as multiple measurements. The study's three measures (audio-visual comprehension scale, verbal communication scale and nonverbal communication scale) demonstrate high internal consistency, with values of

Table 2

Psychometrics Properties of the Study Instruments

Test Name	Cronbach's Alpha (α)	Test–Retest (r)	Difficulty Index (p)	Discrimination Index (D)
Audio-visual Comprehension	0.88	0.91	0.41-0.67	0.31-0.53
Verbal communication	0.92	0.87	0.42-0.58	0.35-0.55
Nonverbal Communication	0.85	0.89	0.33-0.46	0.43-0.63

0.88, 0.92 and 0.85, respectively (Ravid, 2024). The high Cronbach's alpha (α) and test-retest (r) values (all above 0.85, except for one α at 0.85 and one r at 0.87) indicate strong reliability, while the difficulty (p) and discrimination (D) indices suggest appropriate item variability and effectiveness in distinguishing between high and low performers, confirming robust instrument quality.

Qualitative Research Instrument

To triangulate quantitative findings, the researcher conducted semi-structured focus group interviews with 20 EFL university students divided into four groups (five participants each), each lasting 60–75 minutes, to explore the influence of multimodal visual methodologies on audio-visual comprehension, verbal communication, and nonverbal communication. Prior to implementation, a jury of experts evaluated the interview prompts for validity and relevance, leading to minor refinements, and a pilot test with five students confirmed the clarity and suitability of the questions. Core discussion prompts included questions such as, «How do multimodal materials influence your understanding of complex English-language content?» and «In what ways do these methodologies affect your confidence in verbal communication?», with flexibility for follow-up probes to explore emerging themes. Sessions were audio-recorded (with consent) and transcribed, ensuring rigorous qualitative data collection that complemented the study's quantitative approach while providing nuanced insights into learners' experiences with multimodal tools.

Procedures of Data Collection

The study extended for 60 days and included quantitative and qualitative data collection. For the quantitative data, the study passed through phrases such as 'before' (first test), 'during' (14 days after first test follow-up with a 1-hour lecture or more to provide feedback on the three tests) and 'after' (final tests after 14 days). The study employed various multimodal technological devices, including smartboards, to facilitate learning during the designated period. In the audio-visual comprehension assessment, a vodcast was utilised to present material for answering test questions. The verbal and nonverbal communication evaluations were conducted using visual imagery, which was effectively managed and articulated through PowerPoint presentations. The three important phases of quantitative data are as follows:

- In the 'before phase, the participants were required to complete three assessments (e.g. audio-visual comprehension, verbal and nonverbal communication) designed to evaluate their strengths and weaknesses across the three variables under investigation, utilising a smartboard to present the study questions.

- In the 'during' phase, the students received a lesson lasting one hour or more prior to the assessments. This instruction served as feedback regarding the variables being examined in the tests, allowing students to familiarise themselves with the material. The assessments were administered on consecutive days.

- In the 'after' phase, the stability of the information related to the study variables was assessed. This phase aimed to evaluate retention and both short-term and long-term memory over a specified duration.

The study rigorously controlled key variables to ensure consistency across all phases. Timing was standardized, with fixed intervals between the «before» (baseline), «during» (14-day follow-up with feedback), and «after» (final assessment) phases, and assessments were administered on consecutive days during the intervention period. Materials and technology were kept uniform, with smartboards used for presenting test questions, vodcasts for audio-visual comprehension tasks, and PowerPoint for verbal and nonverbal communication evaluations. Learning conditions were maintained through structured, multimodal instruction, including a minimum one-hour feedback lecture in the «during» phase, ensuring all participants received identical content and technological support. These controls minimized external variability, enhancing the reliability of the quantitative data collected.

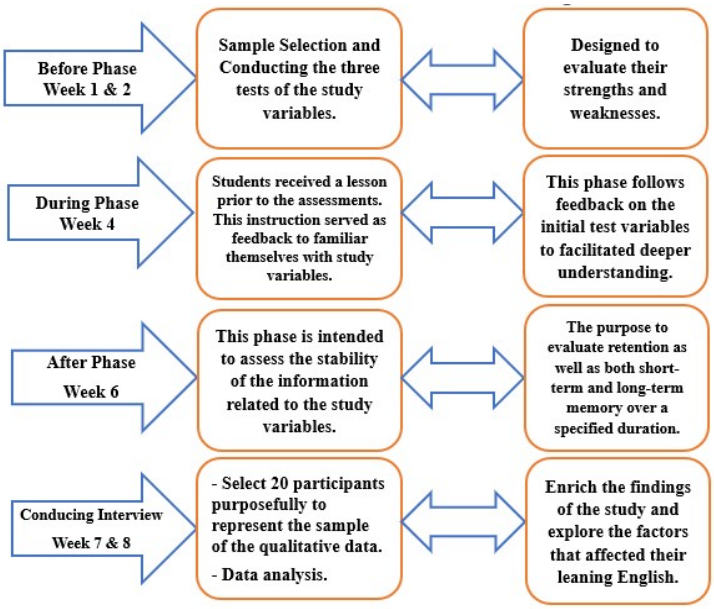
As for qualitative data, the focus group semi-structured interview is used to gather information after the three phases of quantitative data. The interviews were conducted to enrich the findings of the study and explore the factors that affected English learning (e.g. psychological, social, increased student engagement and pedagogical factors). Notes were taken on nonverbal cues and group dynamics during the focus groups, as these can provide additional context to verbal responses.

Data Analysis Procedures

A multimodal transcription of the 'English is Great' vodcasts was conducted on the basis of the audio-visual comprehension test to illustrate the many communicative modes employed during the vodcasts. The analysis was conducted

Figure 1

Procedures of the Study



using the EUDICO Linguistic Annotator (ELAN) by which EUDICO means European Distributed Corpora Project. ELAN is sophisticated annotation software which can be used for annotating audio and video material. Produced at the Max Planck Institute for Psycholinguistics in Nijmegen, Netherlands, ELAN is an essential tool for linguists and other researchers who use video or audio recordings (Cheng, 2024). In this study, ELAN was utilized solely for instructional design purposes. The variety of modes present in the vodcasts — such as participants' gestures, speech, written text, images and music — were categorised into different tiers. This classification allowed for the simultaneous display of precise annotations on the same screen. The teacher notified students that they will view a vodcast on the English language and thereafter respond to questions. To assuage any apprehension, the instructor told them that this was not an examination, but simply a routine audio-visual exercise. Subsequently, a comprehension assessment and a blank sheet of paper were administered to the students. They were allotted 90 seconds to review the test and anticipate the content they would be viewing. Following this reading session, students were directed to turn the test over to deter them from repeatedly referencing both the paper and the screen. While note-taking during the vodcast was not mandatory, students had the choice to engage in it if they desired. Following the elucidation of the directions, the class viewed 'English is Great' once. They were subsequently allotted three minutes to respond to the questions according to their comprehension. After three minutes, the students flipped their tests and rewatched the vodcast, adhering to the same protocol as previously established. Following the second viewing, they were allotted an additional three minutes to complete the test.

The images utilised in the verbal and nonverbal communication assessments were sourced from *Conversation Strategies: Pair and Group Activities for Developing Communicative Competence* by Kehe and Kehe (1994), as well as *Discussion Strategies: Beyond Everyday Conversation* by Kehe et al. (1998). The nonverbal communication referenced were derived from *100+ Body Language Tips* by Rosas (2010). As for the data collected by audio-visual comprehension, verbal and nonverbal communication tests, quantitative data analysis procedures were followed through numerical statistical analysis by using Statistical Package for the Social Sciences version 26. Consequently, repeated-measures ANOVA was conducted on the set of three test scores over time. The Pearson correlation coefficient was employed to assess the relationship between students' performance across each test phase, enabling the calculation of effect sizes for each variable.

In the analysis of qualitative data, semi-structured focus group interviews were subjected to qualitative analysis using Creswell's (2012) thematic analysis steps. These steps are as follows: preparing and organising the data; exploring and coding the data; describing findings and forming themes; representing and reporting findings; interpreting the meaning of the findings; and ensuring the credibility and trustworthiness of the findings (p. 238). Denzin and Lincoln (2018) advised the use of two steps to verify the qualitative findings' objectivity and authenticity. Firstly, an external auditor conducts triangulation and reviews, ensuring the study's accuracy and compatibility by using multiple data gathering sources and instruments. Secondly, the external auditor, who is an experienced applied linguistics specialist, was asked to critically and deliberately review the initial draft of the qualitative analysis of the findings in all its key themes

and subthemes. The external auditor then confirmed that the analytical themes were scientifically sound. These two steps ensured that the researchers' analysis was non-biased and scientifically compelling.

RESULTS

Quantitative Results

Results Related to Audio-Visual Comprehension Test

In order to find out *'the influence multimodal visual methodologies on audio-visual comprehension among EFL university students*.' The results represent sample participants (N = 214) from three colleges of education. The mean score and standard deviation were calculated to assess students' performance over time in the 'before', 'during' and 'after' phases. The mean scores of these tests in the three phases are 46.22, 56.58 and 67.59, respectively, as shown in Table 3.

To examine the variations in students' performance on the audio-visual comprehension test over time, repeated-measures ANOVA was performed to assess the statistical significance of the changes among the three tests over time. Mauchly's test of sphericity indicated that the assumption of sphericity was not violated (P = 0.213). The partial eta squared (partial η2) was calculated to assess the variance in comprehension scores attributable to the three tests over time, as illustrated in Table 4.

According to Table 4, repeated-measures one-way ANOVA indicated a significant difference in the number of the three tests according to the 'before', 'during' and 'after' phases, F(2, 426) = 244.656, (p < .001). The obtained partial η2 was 0.535. The percentage of variance in the dependent variable is substantial and positively influenced, indicating that students achieve improved audio-visual comprehension

Table 3

Descriptive Statistics Table According to Students' Audio-Visual Comprehension Test Performance Over Time

AVC Phases	Mean	Std. Deviation
Before	46.22	10.65
During	56.58	9.475
After	67.59	9.90

Figure 2

Phases of Student Performance in Audio-Visual Comprehension Tests Over Time

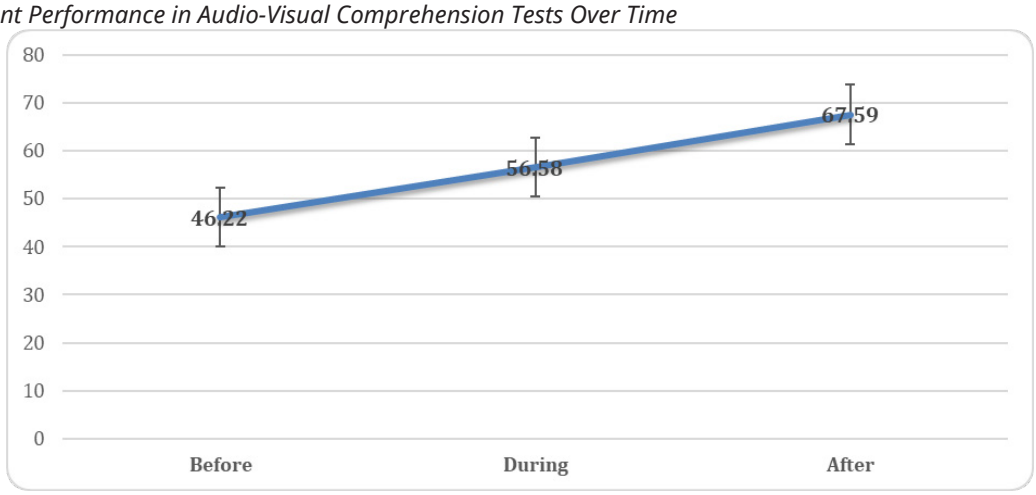


Table 4

Within-Subjects Effects Tests (Sphericity Assumed)

Source		Type III Sum of Squares	df	Mean Square	F	Sig.	Partial Eta Squared
Audio-visual	Sphericity Assumed	48875.629	2	24437.815	244.656	.000	.535
Error Audio-visual	Sphericity Assumed	42551.704	426	99.887			

results when multimodal visual images are utilised. In an educational context, a «significant effect» typically refers to a meaningful or important impact of an intervention, program, teaching method, or policy on student learning outcomes, engagement, or other educational measures. The analysis involved conducting three post hoc t-tests. This phase of the analysis aimed to assess the differences among the three pairs of conditions: before–during, before–after and during–after. Post hoc comparisons between conditions were conducted using three paired sample t-tests. A paired sample analysis revealed a significant difference between the before–during conditions, t (213) = −11.018, $p < .001$. A second paired sample t-test revealed a significant difference between the before–after measurements, t (213) = −20.899, $p < .001$. The third paired sample t-test revealed a significant difference between the during and after conditions, t (213) = −11.796, $p < .001$. All pairs exhibited a statistically significant difference at the $p < 0.05$ level. The reported effect size of $d = 1.072$ indicates a large practical difference between the compared conditions (before–during, before–after, and during–after). For example, if the mean score in the before condition was 50 with a standard deviation of 10, the during condition would average around 60.72, showing a substantial increase of ~10.72 points. Similarly, the after condition would be even higher if the effect is cumulative. This means that not only were the differences statistically significant ($p < .001$), but they also represented meaningful real-world changes, as an effect size above 0.8 is generally considered large. Thus, the findings demonstrate strong, practically important effects across all comparisons. These results indicate a clear progression in students' performance over time, with mean scores increasing significantly from «before» phase to the «during» and «after» phases. While no significant difference is found between the during and after phases. The lack of significant difference between the during and after phase may suggest that while students continued to benefit from these multimodalities, the rate of improvement plateaued after initial exposure.

Results Related to Verbal Communication Test

Regarding *the influence of multimodal visual methodologies on EFL students' verbal communication among the three rounds of test.* Initially, the average score and standard deviation reveal the progression of students' performance over time using the terms 'before,' 'during' and 'after'. Table 5 shows that the respective mean scores before, during and after the three tests were 36.97, 53.95 and 56.14.

Repeated-measures ANOVA was performed on the three test results to assess the statistical significance of the differences observed over time. Mauchly's test of sphericity also revealed a violation of the sphericity assumption, with a P-value of 0.07. Huynh-Feldt epsilon[b] of 0.963 was used to correct an increase in the type I error rate. As shown in Table 6, partial η2 was used to figure out how much of the difference in verbal communication test scores can be explained by the three tests over time.

The repeated-measures one-way ANOVA F-value (1.926, 410.292) = 271.990; $p < .001$) demonstrated a notable disparity in the total number of tests related to the terms 'before',

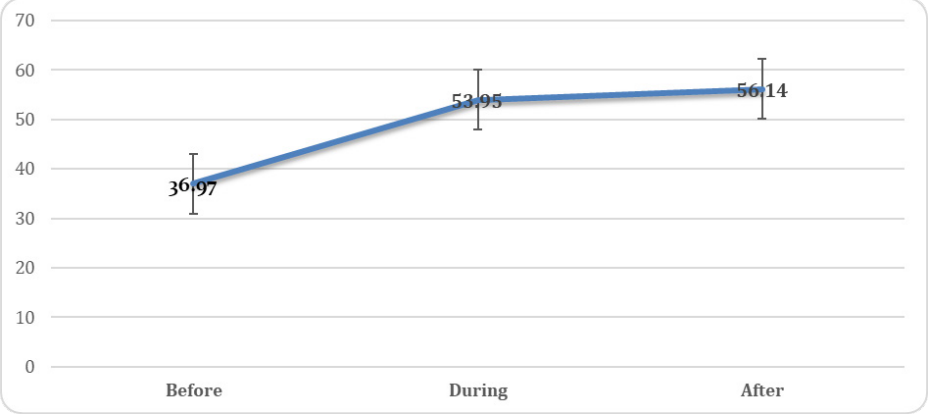
Table 5

Descriptive Statistics Table According to Students' Verbal Communication Test Performance Over Time

VC Phases	Mean	Std. Deviation
Before	36.97	10.39
During	53.95	10.16
After	56.14	12.28

Figure 3

Phases of Student Performance in Verbal Communication Tests Over Time



'during' and 'after'. The partial η2 value was 0.561. The percentage of variation in the dependent variable is substantial and positive, suggesting that the multimodal visual image enhanced students' verbal communication. Analysis indicated that each condition exhibited significant variability. The study utilised three post-hoc t-tests and analysed the differences among before–during, before–after and during–after conditions. Three paired sample t-tests were utilised for post hoc condition comparisons. A preliminary paired sample analysis indicated a significant difference between the before-during conditions, t (213) = −16.949, $p < .001$. A second paired sample t-test indicated a significant difference between the before-after conditions, t (213) = −20.982, $p < .001$. The third paired sample t-test indicated a significant difference between the during-after conditions, t (213) = −2.354, $p < .001$. All pairs exhibited significant differences ($p < 0.05$). A significant effect size $d = 0.707$ supported these results. Cohen's *d* is a measure of effect size that indicates how many standard deviations apart the means of two paired conditions are. A *d* of 0.707 falls between a medium and large effect, meaning the differences between conditions are practically meaningful, not just statistically significant. The results show that turn-taking dominates, suggesting that effective communication heavily depends on participants' ability to manage conversational turns, potentially enhancing engagement and understanding.

Results Related to Nonverbal Communication Test

In order to find out *'How do multimodal visual methodologies influence EFL students' nonverbal communication and what nonverbal cues that are strengthened by utilising multimodal approaches?*', was addressed. The mean scores and standard deviations reveal the progression of students' performance over time using the terms 'before', 'during' and 'after'. Figure 4 shows that the respective mean scores before, during and after the three tests were 37.75, 54.37 and 62.24, as illustrated in Table 7.

Students' performance in the nonverbal communication test varied over time. Repeated-measures ANOVA was conducted on the three tests to assess the statistical significance of

Table 6

Tests of Within-Subjects Effects (Huynh-Feldt)

Source		Type III Sum of Squares	df	Mean Square	F	Sig.	Partial Eta Squared
Verbal communication	Huynh-Feldt	66437.171	1.926	34490.333	271.990	.000	.561
Error Verbal communication	Huynh-Feldt	52028.162	410.292	126.808			

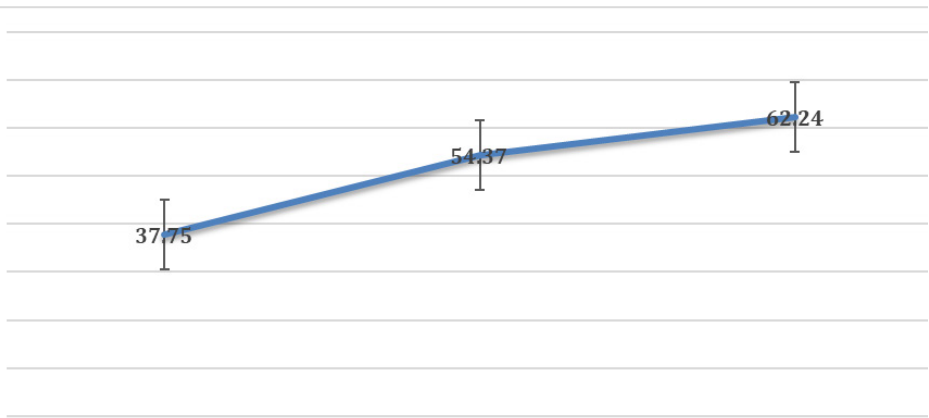
Table 7

Descriptive Statistics Table According to Students' Nonverbal Communication Test Performance Over Time

NVC Phases	Mean	Std. Deviation
Before	37.74	12.19027
During	54.37	10.15926
After	62.24	11.12358

Figure 4

Phases of Student Performance in Nonverbal Communication Tests Over Time



the differences observed over time. Mauchly's test of sphericity revealed a violation of the sphericity assumption, with a P-value of 0.07. Huynh-Feldt epsilon[b] of 0.965 was used to correct an increase in the type I error rate. Quintana and Maxwell (1994) recommended using ε if ε is greater than 0.75. The partial η2 was employed to determine the extent to which the variance in nonverbal communication test scores can be attributed to the three tests conducted over time, as illustrated in Table 8.

Repeated-measures one-way ANOVA F (1.931, 411.202) = 270.521; p <.001) showed a significant difference in the number of three tests based on the phrases 'before', 'during' and 'after'. The partial η2 was 0.559. The percentage of variation in the dependent variable is significantly high and positive, suggesting that the multimodal visual image enhanced students' nonverbal communication. The analysis comprised three post hoc t-tests. This phase of the investigation analysed the differences among the three condition pairings: before–during, before–after, and during–after. Post hoc comparisons between conditions were conducted using three paired sample t-tests. A preliminary paired sample analysis revealed a significant difference between the before and during conditions, t (213) = −15.523, p < .001. A second paired sample t-test revealed a significant difference between the before and after intervention, t (213) = −20.901, p < .001. The third paired sample t-test revealed a significant difference between the during and after periods, t (213) = −8.085, p < .001. All pairs exhibited a statistically significant difference at the p < 0.05 level. The quoted results indicate that all pairwise comparisons (before vs. during, before vs. after, and during vs. after) showed statistically significant differences with very large effect sizes (Cohen's d = 1.00). The results indicate that there are significant differences in the performance of EFL university students across various types of nonverbal communication. Specifically: Facial Expressions are the most effective type of nonverbal communication, significantly outperforming all other types. Kinesics are more effective than kinesthetics, eye contact, proxemics, and artifacts but less effective than facial expressions. Kinesthetics, eye contact, and proxemics show similar effectiveness and are significantly better than artifacts. These findings indicate that facial expressions play a crucial role in nonverbal communication, followed by other types in the context of the performance test.

Qualitative Results

The qualitative results from the interview aligned with the previously mentioned quantitative ones, suggesting that the students experienced a high level of understanding. Thus, the integration of the multimodal visual images and the repeated testing should enrich students' audio-visual comprehension, verbal communication and nonverbal communication meaningfully. To explain further, the analysis of the interview extracts according to the four factors (psychological, social, increased engagement and pedagogical factors) revealed that the students expressed their memory encoding, as indicated in Table 9.

Students reported that the incorporation of audio-visual tools significantly influences listening comprehension and is crucial in contemporary language learning, based on four factors. Also, multimedia materials create engaging learning experiences that promote understanding by stimulating multiple senses, especially when included in instructional activities, and have a major beneficial impact on the growth of language proficiency in general and listening comprehension in particular. Students qualitatively emphasized that multimedia materials significantly improve learning by engaging multiple senses and creating immersive experiences, which aligns with the quantitative data showing a clear progression in comprehension scores from the before to the during and after phases. Together, these results demonstrate that multimedia tools play a crucial role in boosting comprehension, with their strongest impact occurring upon initial integration, followed by maintained, though not further increasing, advantages.

The students' responses to the interview questions show that visual representations greatly boost self-esteem and confidence when expressing their ideas. They enhance understanding and retention, encourage social connections in virtual spaces, and make narratives more captivating and relevant. Visuals that evoke emotions enhance the bond with the audience, resulting in greater engagement. Moreover, visuals enhance verbal communication and offer valuable feedback, thus improving verbal skills, as shown by the beneficial effects of educational resources that include visuals. Quantitatively, the prevalence of turn-taking suggests that managing conversational flow is key to effective dialogue, further supporting engagement and comprehension. Together, these findings indicate that combining visual elements with structured interaction techniques creates a

Table 8

Tests of Within-Subjects Effects (Huynh-Feldt)

Source		Type III Sum of Squares	df	Mean Square	F	Sig.	Partial Eta Squared
Nonverbal communication	Huynh-Feldt	66931.523	1.931	34670.063	270.521	.000	.559
Error Nonverbal communication	Huynh-Feldt	52699.810	411.202	128.160			

Table 9

Interview Extracts Indicating Students' Reference to "Audio-Visual Comprehension"

No.	Factors	Extracts
Audio-visual comprehension	Psychological factors	S01 "Ummm, I prefer visual images that make me more enjoyable/interesting/fun/exciting."
		S04 "I like Linking spoken words with visual cues supports vocabulary acquisition and is especially beneficial for language learners."
	Social factors	S05 "I feel comfortable when technologies employed such audio-visual that provide opportunities for realistic language learning experiences."
		S11 "I believe that audiovisual tools serve as catalysts for cross-cultural exchange and international cooperation."
	Increase students engagement	S18 "I feel more energetic when I watch the video."
		S08 "I enjoy how multimedia resources like audio-visual stimulate my senses and provide interesting learning opportunities that advance my comprehension."
		S03 "Compared to the audio track, I think I'm paying closer attention and being more focused."
	Pedagogical factors	S13 "In my opinion, using audiovisuals instead of only traditional audio helps the audience better understand the situation at hand."
		S02 "Visual materials contribute to an inclusive environment by accommodating different learning style."

Table 10

Interview Extracts Indicating Students' Reference to "Verbal communication"

No.	Factors	Extracts
Verbal communication	*Psychological factors*	*S19 «Excellent, I favor visual representations that enhance my self-esteem and empower me to confidently share my thoughts and opinions.»*
		S16 "I feel more comfortable talking about something when I use visual imagery, which improves my comprehension and memory."
	Social factors	*S11 "I found that Visual images fosters verbal social interactions within online communities."*
		S03 "Amazing, Visual elements facilitate my verbal storytelling and cultures norms by making the narrative more compelling and relatable."
	Increase students engagement	*S20 "I like the way of Visual images that drive my willingness to engage in conversation."*
		S06 "Wonderful, I feel that emotional stimuli, such as compelling images or videos, can create a strong connection with the audience, increasing the engagement through likes, shares, and comments."
	Pedagogical factors	*S14 "I love educational textbooks' that employ visuals specially the book real speaking and listening improved my verbal skills."*

more dynamic and effective learning environment, where emotional resonance, cognitive clarity, and collaborative communication synergistically improve outcomes. This alignment underscores the value of multimodal approaches in education, leveraging both visual and conversational strategies to maximize participation and learning.

The qualitative results from the student interviews emphasized that the interplay of psychological, social, engagement, and pedagogical factors in nonverbal communication is vital for effective teaching and learning, enabling educators to refine their methods and enhance student performance. These findings align with the quantitative results, which revealed significant differences in EFL university students' performance across various types of nonverbal communication, with facial expressions emerging as the most effective, significantly outperforming other types. Kinesics ranked next in effectiveness, surpassing kinesthetics, eye contact, proxemics, and artifacts but remaining less impactful than facial expressions, while kinesthetics, eye contact, and proxemics showed similar effectiveness and were all significantly better than artifacts. Together, these results highlight the crucial role of facial expressions in nonverbal communication, followed by other types, further supporting the importance of integrating these components into teaching strategies to optimize student outcomes.

Table 11

Interview Extracts Indicating Students' Reference to "Nonverbal communication".

No.	Factors	Extracts
Nonverbal communication	Psychological factors	S06 "Visual images provided me with the opportunity to interpret nonverbal signals, such as the emotional tone of a lesson or the seriousness of a topic, which can impact my comprehension and retention of information."
		S12 "I enjoyed utilising visual images for learning nonverbal communication, which necessitates self-awareness and emotional intelligence, enabling me to comprehend signs such as gestures and facial expressions."
	Social factors	S14 "In my opinion, images help us comprehend how cultural norms influence nonverbal communication. Because Different cultures may perceive gestures, eye contact, and personal space in different ways."
		S10 "I enjoyed how positive nonverbal behaviours, such as maintaining eye contact or utilising open body language, may foster trust and better connections, which are necessary for effective teaching."
	Increase students engagement	S17 "I prefer to convey myself visually rather than verbally. For example, using thumbs-up or emoji reactions in online lectures enables me to swiftly express and grasp levels or emotional states."
		S20 "I like the way that visuals help me understand body language movements—such as nodding or fidgeting—and help me adjust my teaching strategies in real time to better meet my needs."
	Pedagogical factors	S09 "It exposes us to the effective use of visual resources, improves instructional clarity, and simplifies the delivery of complicated ideas."
		S07 "It was an excellent opportunity for me to learn nonverbal communication through the use of visual representations, such as gesturing to indicate transitions or utilising facial expressions to convey expectations."

DISCUSSION

The results of this study demonstrate that multimodal visual methodologies enhance EFL learners' audio-visual comprehension, verbal fluency, and nonverbal communication. Students exposed to subtitled videos, contextual images, and gesture-based instruction consistently outperformed their peers in comprehension tasks. They also demonstrated a richer use of facial expressions and body language, confirming the integrative effect of multimodal input. These findings are consistent with Fay et al. (2013), who showed that gestures often precede or substitute for verbal expression and function as primary tools for conveying meaning. The positive outcomes extended across different learning preferences and were most evident when visual and verbal channels were activated simultaneously. Participants reported increased motivation, engagement, and communicative confidence, especially in tasks involving repeated exposure and varied media formats. While the study confirmed promising effects in verbal, written, and embodied modalities, some areas, such as vodcast-based movement tasks, would benefit from more rigorous statistical examination to clarify the role of non-linguistic cues in shaping performance.

These findings align with prior research on the benefits of multimodal instruction in language learning. Bairstow and Lavaur (2012) and Lin (2016) highlighted the contribution of L2 subtitles to listening comprehension, while Mohsen (2016) and Pardo-Ballester (2016) emphasized the role of audio-visual input in activating top-down strategies. Batty (2015), Lesnov (2017), and Hsieh (2020) further noted the importance of authentic multimodal contexts for receptive processing. Although Lin (2016) reported variable outcomes for L1 and L2 subtitles, her study lacked longitudinal scope. The present research addresses this gap by showing that sustained exposure to multimodal cues fosters improvement not only in comprehension but also in expressive and interactive skills over time.

Theoretical Integration

The observed learning gains can be understood through several theoretical frameworks. Cognitive Load Theory explains how distributing input across visual and auditory channels reduces working memory overload. Working Memory Theory supports this view by showing that dual coding enhances retention and recall. Sociocultural theory adds another layer by framing visuals and gestures as scaffolding devices that

support internalization of complex linguistic forms. The influence of learner-specific variables such as proficiency, anxiety, learning style, and cultural background also emerged as relevant. For example, subtitled videos were particularly effective for low-proficiency learners, and students from high-context cultures relied more on nonverbal cues. The effectiveness of instruction was moderated by media design quality, where clear and interactive resources supported engagement while poorly designed materials impeded learning. These findings reinforce Mayer's Cognitive Theory of Multimedia Learning (2005, 2009), which emphasizes dual-channel input, limited capacity, and active learning in second language acquisition.

Pedagogical Implications

The study suggests that multimodal instruction should be adapted to learner level and context. Subtitled videos in L1 may support beginners, while authentic L2 input can challenge advanced learners. Explicit instruction in nonverbal communication should be integrated into lesson planning to develop expressive fluency. Tasks such as vodcasts, storytelling, and gesture-based activities offer opportunities for multimodal practice. Repeated testing also contributed to vocabulary acquisition, consistent with Metsämuuronen (2013) and Vojdanoska et al. (2009). The emotional and social benefits of multimodal tasks were particularly visible in group activities and gamified formats. These observations align with Zhang and Zou (2022), Salamanti et al. (2023), Kress and Selander (2021), and Mayer (2021), who emphasized the motivational and self-efficacy effects of learner-centered multimodal design.

Limitations

Although the study revealed consistent patterns, several limitations should be acknowledged. The participant pool was limited to three Iraqi universities, which constrains generalizability. Individual cognitive factors such as spatial ability and working memory were not systematically measured. While facial expressions and body movements emerged as dominant nonverbal factors, artifacts such as props and visual symbols were rated as less effective, despite research by Bambaeeroo and Shokrpour (2017), which demonstrated that teacher appearance and use of physical space can significantly influence learner attention and academic engagement. This discrepancy may indicate that such static cues operate contextually and require more nuanced measurement. Furthermore, the finding that turn-taking, rather than visuals, was the strongest predictor of communicative success introduces a thematic ambiguity. It remains unclear whether visual support facilitates turn-taking or whether the two operate independently. Doumont (2002) noted that visuals convey intuitive meaning, but their interaction with social communication structures may require more nuanced investigation.

CONCLUSION

This study provides empirical evidence that multimodal visual methodologies significantly enhance audio-visual comprehension, as well as verbal and nonverbal communication in EFL contexts. The integration of quantitative and qualitative data demonstrates that multimodal input supports language learning through a combination of cognitive, emotional, and social mechanisms. Statistically significant gains across multiple testing phases suggest that multimodal tools strengthen the connection between linguistic and sensory processing, contributing to improved retention and communicative performance. Students' interview responses further confirm the affective value of multimodal instruction, emphasizing reduced anxiety, increased motivation, and greater engagement.

The study contributes to the field by offering a triangulated analysis that moves beyond additive models of learning. Multimodal input was shown to not only supplement traditional instruction but to transform the way learners internalize and apply language. These results align with dual-coding and sociocultural theories, illustrating how visual and nonverbal cues facilitate both comprehension and interactional competence. Importantly, the study highlights the relevance of turn-taking, gestural expression, and contextual cues in bridging receptive and productive skills.

Future studies should focus on isolating the specific mechanisms behind multimodal learning, including cognitive load distribution, emotional engagement, and interactive structure. Experimental designs with randomized control groups would help clarify causal relationships, while cross-cultural comparisons could shed light on the influence of sociolinguistic context. The role of multimodal instruction in virtual and AI-enhanced learning environments also warrants further attention. Adaptive platforms that adjust visual input to learner profiles may offer promising avenues for increasing the efficiency and personalization of EFL instruction.

AI DISCLOSURE STATEMENT

The authors confirm that AI-assisted tools, including deepseek and ChatGPT, were utilized to enhance the manuscript's language clarity and coherence. These tools were employed primarily for linguistic refinement and to assist in addressing reviewer feedback - particularly in aligning critical theory with the study's scope. However, all conceptual arguments, analytical frameworks, and interpretative insights were independently developed by the authors and rigorously reviewed to uphold scholarly originality and academic integrity.

DECLARATION OF COMPETING INTEREST

None declared.

AUTHORS' CONTRIBUTION

Ibrahim Hassan Ali: conceptualization; data curation; formal analysis; funding acquisition; inquiry; methodology; administration of the project; resources; writing - original draft; review and editing.

Istabraq Tariq Jawaad Alazzawi: supervision; investigation; re-sources; validation; review and editing (literature review and discussion sections).

REFERENCES

Abdikarimova, M., Tashieva, N., & Abdullaeva, Z. (2021). Developing students' verbal communication skills and speech etiquette in English language teaching. *Open Journal of Modern Linguistics*, *11*(1), 83-89. http://dx.doi.org/10.4236/ojml.2021.111007

AbdulGhafoor, C. & Challob, A.I. (2021). The influence of using online discussion on developing EFL students' grammar knowledge.ٰ *Korean Journal of English Language and Linguistics, 21*, 837-855. https://doi.org/10.15738/kjell.21..202109.837

Abdullah, A. H., Soh, H. M., Mokhtar, M., Hamzah, M. H., Ashari, Z. M., Ali, D. F., & Abd Rahman, S. N. S. (2020). Does the use of smart board increase students' higher order thinking skills (HOTS)? *IEEE Access, 9*, 1833-1854.ٰ http://dx.doi.org/10.1109/ACCESS.2020.3042832

Abduraximova, M. (2024). Multimodal discourse analysis in English: Integrating verbal and non-verbal modes. *New Uzbekistan Journal of Academic Research*, *1*(7), 62-65. https://doi.org/10.5281/zenodo.11242460

Adhitya, N., & Valiansyah, V. (2024). The impacts of native speaker teachers' nonverbal communication in EFL classrooms. *The Journal of English Language Teaching, Literature, and Applied Linguistics*, *5*(2), 95–108. https://doi.org/10.37742/jela.v5i2.110

AL Nofali, H. A., & Gasim, A. (2024). Adopting verbal and non-verbal communication strategies in oral discussion by Omani university students. *International Journal for Multidisciplinary Research*. https://doi.org/10.36948/ijfmr.2024.v06i05.28823

Arbab, I. A. E. (2020). *Effect of using audio-visual materials on students' language achievements (A case study of secondary schools' students at Eastern Gazeera)* [Unpublished doctoral dissertation]. Sudan University of Science & Technology.

Archvadze, E. (2023). Exploring multimodality-social-semiotic analysis of communication in TEFL. *Language and Culture*. http://dx.doi.org/10.52340/lac.2023.08.05

Bairstow, D., & Lavaur, J. M. (2012). Audiovisual information processing by monolinguals and bilinguals: Effects of intralingual and interlingual subtitle. *Audiovisual Translation and Media Accessibility at the Crossroads*, *36,* 273–293. http://dx.doi.org/10.1163/9789401207812_016

Baldry, A., Thibault, P. J., Coccetta, F., Kantz, D., & Taibi, D. (2020). Multimodal ecological literacy: Animal and human interactions in the animal rescue genre. *Multiliteracy advances and multimodal challenges in ELT environments* (pp.155-218). Forum.

Bambaeeroo, F., & Shokrpour, N. (2017). The impact of the teachers' non-verbal communication on success in teaching. *Journal of Advances in Medical Education & Professionalism*, *5*(2), 51-59.

Batty, A. O. (2015). A comparison of video-and audio-mediated listening tests with many facet Rasch modelling and differential distractor functioning. *Language Testing, 32*(1), 3–20. http://dx.doi.org/10.1177/0265532214531254

Becker, S. R., & Sturm, J. L. (2017). Effects of audiovisual media on L2 listening comprehension: A preliminary study in French. *Calico Journal*, *34*(2), 147-177. http://dx.doi.org/10.1558/cj.26754

Benetti, S., Ferrari, A., & Pavani, F. (2023). Multimodal processing in face-to-face interactions: A bridging link between psycholinguistics and sensory neuroscience. *Frontiers in Human Neuroscience*, *17*, 1108354. http://dx.doi.org/10.3389/fnhum.2023.1108354

Bezemer, J. & Kress, G. (2016). *Multimodality, learning and communication: A social semiotic frame*. Routledge. https://doi.org/10.4324/9781315687537

Bilfaqih, Y., & Qomarudin, M. N. (2017). *Multimodal analysis* (vol. 1). Dee Publish. https://doi.org/10.23887/jet.v7i1.54336

Bouchey, B., Castek, J., & Thygeson, J. (2021). Multimodal learning. In J. Ryoo, & K. Winkelmann (Eds.), *Innovative Learning Environments in STEM Higher Education*. SpringerBriefs in Statistics. Springer. https://doi.org/10.1007/978-3-030-58948-6_3

Bromberek-Dyzman, K., Jankowiak, K., & Chełminiak, P. (2021). Modality matters: Testing bilingual irony comprehension in the textual, auditory, and audio-visual modality. *Journal of Pragmatics*, *180*, 219-231. http://dx.doi.org/10.1016/j.pragma.2021.05.007

Burgoon, J. K. (2003). Nonverbal communication skills. *Handbook of Communication and Social Interaction Skills.* Lawrence Erlbaum Associates, Inc.

Campbell, R. (2007). The processing of audio-visual speech: empirical and neural bases. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *363*(1493), 1001–1010. https://doi.org/10.1098/rstb.2007.2155

Chen, G., & Fu, X. (2003). Effects of multimodal information on learning performance and judgment of learning. *Journal of Educational Computing Research*, *29*(3), 349-362. http://dx.doi.org/10.2190/J54F-B24D-THN7-H9PH

Cheng, J. (2024). Application of ELAN in human-machine mutual translation of English for the news. *Journal of Electrical Systems*, *20*(6s), 2343-2351. http://dx.doi.org/10.52783/jes.3218

Chiriac, R. (2025). The importance of audio-visual materials in teaching ESL vocabulary at primary school level. *LiBRI. Linguistic and Literary Broad Research and Innovation*, *9*(1), 18-25. https://orcid.org/0009-0002-3541-5153

Coffin, C. & Donohue, J. (2014). *A language as social semiotic-based approach to teaching and learning in higher education* (Language Learning Monograph Series). Wiley-Blackwell. https://doi.org/10.1111/lang.2014.64.issue-s

Creswell, J. W. (2012). *Educational research: Planning, conducting, and evaluating quantitative and qualitative research* (4th ed.). Pearson Education.

Creswell, J. W. (2017). *Educational research: Planning, conducting, and evaluating quantitative and qualitative research*. Edam.

Creswell, J. W. (2024). My 35 years in mixed methods research. *Journal of Mixed Methods Research*, 18(3), 203-215. http://dx.doi.org/10.1177/15586898241253892

Denzin, N. K., & Lincoln, Y. S. (Eds.). (2018). *The sage handbook of qualitative research* (5th ed.). SAGE publications, Inc.

Doumont, J. L. (2002). Verbal versus visual: A word is worth a thousand pictures, too. *Technical Communication*, *49*(2), 219-224.

Ed-Dali, R. (2024). Enhancing EFL learning through multimodal integration: The role of visual and auditory features in Moroccan textbooks. *Journal of World Englishes and Educational Practices*, *6*(3). https://doi.org/10.32996/jweep.2024.6.3.5

Ercan, S. A., Asenbaum, H., Curato, N., & Mendonça, R. F. (2022). *Research methods in deliberative democracy* (p. 529). Oxford University Press. http://dx.doi.org/10.1093/oso/9780192848925.001.0001

Espino, R. M., Castro, F. Z., Jiménez, M. D. C. R., Suárez, J. L. M., & Marrero, A. C. (2006). Modulated frequency systems in students with cochlear implant. *Auditio, 3*(2), 32-36. https://doi.org/10.51445/sja.auditio.vol3.2006.0036

Fàbregues, S., Younas, A., Escalante-Barrios, E. L., Molina-Azorin, J. F., & Vázquez-Miraz, P. (2024). Toward a framework for appraising the quality of integration in mixed methods research. *Journal of Mixed Methods Research*, *18*(3), 270-280. https://doi.org/10.1177/15586898241257555

Farías, M., Obilinovic, K., Orrego, R., & Gregersen, T. (2014). Evaluating types and combinations of multimodal presentations in the retention and transfer of concrete vocabulary in EFL learning. *Revista Signos*, *47*(84), 21-39.

Fay, N., Arbib, M., & Garrod, S. (2013). How to bootstrap a human communication system. *Cognitive Science*, *37*(7), 1356-1367. https://doi.org/10.1111/cogs.12048

Fitria, T. N. (2024). Teaching IELTS speaking skills: How is the students' preparation for taking the test? *Journal of English Education Program*, *5*(2). http://dx.doi.org/10.26418/jeep.v5i2.75380

Franzen, M. D. (2013). *Reliability and validity in neuropsychological assessment.* Springer Science & Business Media. http://dx.doi.org/10.1007/978-1-4757-3224-5

Ghoushchi, S., Yazdani, H., Dowlatabadi, H., & Ahmadian, M. (2021). A multimodal discourse analysis of pictures in ELT textbooks: Modes of communication in focus. *Jordan Journal of Modern Languages and Literatures, 13*(4), 623-644. http://dx.doi.org/10.47012/jjmll.13.4.2

Gilakjani, A. P., Ismail, H. N., & Ahmadi, S. M. (2011). The effect of multimodal learning models on language teaching and learning. *Theory & Practice in Language Studies, 1*(10). http://dx.doi.org/10.4304/tpls.1.10.1321-1327

Hahl, K., Lehtovuori, K., & Pietarila, M. (2025). Video research in language classrooms: Activities and target language use in early language learning. *Scandinavian Journal of Educational Research*, *69*(2), 363-375. https://doi.org/10.1080/00313831.2024.2308874

Halliday, M. A. K., and Hasan, R. (1985). *Language, context, and text: Aspects of language in a social-semiotic perspective*. Oxford University Press.

Haneef, M., Faisal, M. A., Alvi, A. K., & Zulfiqar, M. (2014). The role of non-verbal communication in teaching practice. *Science International*, *26*(1), 513-517.

Harutyunyan, N. (2023). Decoding multimodal texts of media discourse. *Armenian Folia Anglistika*, *19*(1), 83-94. https://doi.org/10.46991/AFA/2023.19.1.083%20

Heimbürger, A. (2013). Context meets culture. In *Modeling and Using Context: 8th International and Interdisciplinary Conference* (pp. 143-156). Springer Berlin Heidelberg.

Herman, Murni, S. M., Sibarani, B., & Saragih, A. (2019). Structures of representational meta-functions of the "Cheng Beng" ceremony in pematangsiantar: A multimodal analysis. *International Journal of Innovation, Creativity and Change*, *8*(4), 34–46. https://doi.org/10.25134/erjee.v10i2.6244

Hsieh, Y. (2020). Effects of video captioning on EFL vocabulary learning and listening comprehension. *Computer Assisted Language Learning, 33*(5–6), 567–589. https://doi.org/10.1080/09588221.2019.1577898

Huang, J. (2006). English abilities for academic listening: How confident are Chinese students? *College Student Journal*, *40*(1), 218.

Jewitt, C., & Kress, G. (Eds.) (2003). *Multimodal literacy*. Peter Lang. http://dx.doi.org/10.3102/0091732X07310586

Kartika, D., Siahaan, S., Herman, H., Rumapea, E. L., & Silalahi, T. F. (2023). Implementation of audio-visual teaching media in improving students' listening comprehension: A case on teaching method. *Journal of English Language and Education*, *8*(2), 86-96. https://doi.org/10.31004/jele.v8i2.428

Kehe, D., & Kehe, P. D. (1994). *Conversation Strategies: Pair and group activities for developing commupnicative competence* (2nd ed.). Pro Lingua Associates.

Kehe, D., Kehe, P. D., & Toos, A. (1998). *Discussion strategies: Beyond everyday conversation*. Pro Lingua Associates.

Kenzhegaliev, S. A., Aitzhan, D. K., & Zhanuzakhova, K. K. (2023). Communicative potential of the means of non-verbal communication. *Bulletin of the Karaganda University. Philology Series*, *112*(4), 117-123. https://doi.org/10.31489/2023Ph4/117-123

Khadimally, S. (2016). Visual communications and learning. *US-China Education Review*, *6*(8), 466-479. http://dx.doi.org/10.17265/2161-623X/2016.08.002

Kress, G. (2009). *Multimodality: A social semiotic approach to contemporary communication*. Routledge.

Kress, G. (2010). *Multimodality: A social semiotic approach to contemporary communication*. Routledge.

Kress, G. R., & van Leeuwen, T. (2001). *Multimodal discourse: The modes and media of contemporary communication*. Arnold.

Kress, G. R., & Van Leeuwen, T. (2006). *Reading images: The grammar of visual design*. Routledge. http://dx.doi.org/10.4324/9781003099857

Kress, G., & Selander, S. (2021). *Multimodal teaching and learning: The rhetorics of the classroom*. Bloomsbury.

Lennartsson, B. (2010, April). Visualization in education-support for the cognitive processes in understanding and learning. In *International Conference on Computer Supported Education* (vol. 2, pp. 375-379). SCITEPRESS.

Lesnov, R. O. (2017). Using videos in ESL listening achievement tests: effects on difficulty. *Eurasian Journal of Applied Linguistics, 3*, 67–91. http://dx.doi.org/10.32601/ejal.461034

Lin, L. F. (2016). The impact of video-based materials on Chinese speaking learners' English text comprehension. *English Language Teaching, 9,* 1–13. http://dx.doi.org/10.5539/elt.v9n10p1

Love, N. (2019). The Semiotics of JR Firth. In T. Sebeok & J. Umiker-Sebeok (Eds.), *The Semiotic Web 1990: Recent Developments in Theory and History*. De Gruyter Mouton.

Madella, P., Wharton, T., & Romero-Trillo, J. (2023). *Non-verbal communication and context: multi-modality in interaction*. Cambridge University Press.

Maguire, L. L. (2005). Literature review: Faculty participation in online distance education: Barriers and motivators. *Online Journal of Distance Learning Administration*, *8*(1).

Mayer, R. E. (2005). *The Cambridge handbook of multimedia learning*. Cambridge University Press. http://dx.doi.org/10.1017/CBO9780511816819

Mayer, R. E. (2009*). Multimedia learning*. Cambridge University Press.

Mayer, R. E. (2014). *The Cambridge handbook of multimedia learning* (2nd ed.). Cambridge University Press.

Mayer, R. E. (2021). *Multimedia learning*. Cambridge University Press.

McDuffie, A. (2021). Verbal communication. *Encyclopedia of Autism Spectrum Disorders*, 5029-5029. http://dx.doi.org/10.1007/978-1-4419-1698-3_25

Metsämuuronen, J. (2013). Effect of repeated testing to the development of Biblical Hebrew language proficiency. *Journal of Educational and Psychological Development, 3*(1). http://dx.doi.org/10.5539/jedp.v3n1p10

Mikhael, J. N., Rafiqa, S., & Siaahan, I. (2022). The effect of teachers verbal and non-verbal communication on students motivation in learning English at Smak Frater Don Bosco Tarakan. *Borneo Journal of English Language Education*, *4*(2). http://dx.doi.org/10.35334/bjele.v4i2.3136

Miller, S. M. (2007). English teacher learning for new times: Digital video composing as multimodal literacy practice. *English Education, 40*(1), 61–83.

Mohsen, M. A. (2016). The use of help options in multimedia listening environments to aid language learning: a review. *British Journal of Educational Technology, 47*, 1232–1242. http://dx.doi.org/10.1111/bjet.12305

Moreno, R., & Mayer, R. (2007). Interactive multimodal learning environments. *Educational Psychology Review*, 19, 309-326.

Namaziandost, E., Nasri, M., & Akbari, S. (2019). The impact of teaching listening comprehension by audio and video aids on the intermediate EFL learners listening proficiencies. *Language, Literature and Culture*, *2*(3), 121-128.

New London Group. (1996). A pedagogy of multiliteracies: Designing social futures. *Harvard Educational Review, 66,* 60–92. http://dx.doi.org/10.17763/haer.66.1.17370n67v22j160u

Ngongo, M. (2021). The investigation of modality and adjunct in spoken text of proposing a girl using Waijewa language based on Halliday's systemic functional linguistic approach. *English Review: Journal of English Education*, *10*(1), 223–234. https://doi.org/0.25134/erjee.v10i1.5382

Norte Fernández-Pacheco, N. (2018). The impact of multimodal ensembles on audio-visual comprehension: Implementing vodcasts in EFL contexts. *Multimodal Communication*, *7*(2). http://dx.doi.org/10.1515/mc-2018-0002

Nuzzaci, A. (2019). A picture is worth a thousand words: Visual thinking between creative thinking and critical thinking in the teaching-learning processes. *Img Journal*, (1), 234-253. https://doi.org/10.6092/issn.2724-2463/11071

Olelewe, C. J., Dong, C., Abdullahi, M., & Nwangwu, C. E. (2023). Effects of using a video-clip instructional strategy on students' performance in a computer networking course. *Technology, Pedagogy and Education, 32*(3), 351-365. http://dx.doi.org/10.1080/1475939X.2023.2201931

Oluwatayo, J. A. (2012). Validity and reliability issues in educational research. *Journal of Educational and Social Research, 2*. https://doi.org/10.5901/jesr.2012.v2n2.391

Pardo-Ballester, C. (2016). Using video in web-based listening tests. *Journal of New Approaches in Educational Research, 5,* 91–98. http://dx.doi.org/10.7821/naer.2016.7.170

Pintado, B. R., & Fajardo, T. (2021). Learning idioms through the multimodal approach. *Learning*, *12*(24). http://dx.doi.org/10.7176/JEP/12-24-03

Pratolo, B. W. (2019). Integrating body language into classroom interaction: The key to achieving effective English language teaching. *Humanities & Social Sciences Reviews*, *7*(3), 121–129. https://doi.org/10.18510/hssr.2019.7319

Proudfoot, K. (2023). Inductive/deductive hybrid thematic analysis in mixed methods research. *Journal of Mixed Methods Research*, *17*(3), 308-326. http://dx.doi.org/10.1177/15586898221126816

Purwaningtyas, T. (2020). Didactic symbol of visual images in EFL textbook: Multi-modal critical discourse analysis. *Pedagogy: Journal of English Language Teaching*, *8*(1), 51-63. http://dx.doi.org/10.32332/pedagogy.v8i1.1959

Putri, Y. S., Setyaningsih, E., & Putra, K. A. (2024). Embracing multimodality to teach EFL ыtudents: english course teachers' perspectives. *International Journal of Educational Research and Social Sciences*, *5*(4), 731–735. https://doi.org/10.51601/ijersc.v5i4.847

Quintana, S. M. & Maxwell, S. E. (1994). A monte Carlo comparison of seven an adjustment procedure in repeated measures designs with small sample sizes. *Journal of Educational and Behavioural Statistics, 19*, 57–71. http://dx.doi.org/10.3102/10769986019001057

Ramirez, D., and Alonso, I. (2007). Using digital stories to improve listening comprehension with Spanish young learners of English. *Language Learning & Technology, 11(1)*, 87–101. http://goo.gl/U5z8S0 (2016-05-19).

Ravid, R. (2024). *Practical statistics for educators*. Rowman & Littlefield.

Richards, J. C., & Rodgers, T. S. (2014). *Approaches and methods in language teaching*. Cambridge University Press.

Rojas, J. H. C. (2024). The effect of flipped classroom audiovisual content on students' talking time and speaking skills in an adult EFL class. *Gist: Education and Learning Research Journal*, *28*(28). https://doi.org/10.26817/16925777.1609

Rosas, E. (2010). *100+ body language tips*. Knesix Institute.

Ruswandi, R., & Arief, M. (2024). Inspiring through interaction: The impact of teachers' verbal and non-verbal communication in EFL classes. *Voices of English Language Education Society*, *8*(2). http://dx.doi.org/10.29408/veles.v8i2.26821

Salamanti, E., Park, D., Ali, N., & Brown, S. (2023). The efficacy of collaborative and multimodal learning strategies in enhancing English language proficiency among ESL/EFL Learners: a quantitative analysis. *Research Studies in English Language Teaching and Learning*, *1*(2), 78-89. https://doi.org/10.62583/rseltl.v1i2.11

Shaojie, T., Samad, A. A., & Ismail, L. (2022). Systematic literature review on audio-visual multimodal input in listening comprehension. *Frontiers in Psychology*, *13*. http://dx.doi.org/10.3389/fpsyg.2022.980133

Sueyoshi, A., & Hardison, D. M. (2005). The role of gestures and facial cues in second language listening comprehension. *Language learning*, *55*(4), 661-699. https://psycnet.apa.org/doi/10.1111/j.0023-8333.2005.00320.x

Surguladze, S. A., Calvert, G. A., Brammer, M. J., Campbell, R., Bullmore, E. T., Giampietro, V., et al. (2001). Audio–visual speech perception in schizophrenia: An fMRI study. *Psychiatry Research: Neuroimaging, 106*, 1–14. http://dx.doi.org/10.1016/S0925-4927(00)00081-0

Sutiyatno, S. (2018). The effect of teacher's verbal communication and non-verbal communication on students' English achievement. *Journal of language teaching and research*, *9*(2), 430-437. http://dx.doi.org/10.17507/JLTR.0902.28

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

Szawerna, M. (2023). Theoretical and analytical explorations of multimodality. *Anglica Wratislaviensia*, *61*(1), 9-11. https://dx.doi.org/10.19195/0301-7966.61.1.1

Tang, S. F., & Logonnathan, L. (2016). *Assessment for learning within and beyond the classroom*. Springer Verlag. http://dx.doi.org/10.1007/978-981-10-0908-2

Tussupova, G. K., Temirbekova, A. T., & Kerimbayeva, K. K. (2016). Verbal and non-verbal means of communication in teaching foreign. *New Trends and Issues Proceedings on Humanities and Social Sciences*. 2(5), pp 205-209. http://dx.doi.org/10.18844/PROSOC.V2I5.1123

Van Leeuwen, T. (2020). Multimodality and multimodal research. *The SAGE handbook of visual research methods* (2nd ed.) (pp. 465-483). SAGE.

Vandergrift, L., & Goh, C. C. M. (2012). *Teaching and learning second language listening: Metacognition in action*. Routledge.

Victoria, M. (2021). The verbal and the visual in language learning and teaching: Insights from the 'Selfie Project'. *The Language Learning Journal*, *49*(1), 93-104. http://dx.doi.org/10.1080/09571736.2018.1484797

Vojdanoska M., Cranney, J., & Newell, B. R. (2009). The testing effect: The role of feedback and collaboration in a tertiary classroom setting. *Applied Cognitive Psychology*, *24*(8), 1183–1195. http://dx.doi.org/10.1002/acp.1630

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

Wagner, E. (2010). The effect of the use of video texts on ESL listening test-taker performance. *Language Testing, 27*, 493–513. http://dx.doi.org/10.1177/0265532209355668

Wagner, J., Lingenfelser, F., Bee, N., & André, E. (2011). Social Signal Interpretation (SSI) - A framework for real-time sensing of affective and social signals. *KI-Kuenstliche Intelligenz*, *25*, 251-256. http://dx.doi.org/10.1007/s13218-011-0115-x

Wahyuni, A. (2018). The power of verbal and nonverbal communication in learning. In *1st International Conference on Intellectuals' Global Responsibility* (pp. 80-83). Atlantis Press. http://dx.doi.org/10.2991/icigr-17.2018.19

Wang, C., Yan, J., & Liu, B. (2014). An empirical study on washback effects of the internet-based college English test band 4 in China. *English Language Teaching, 7,* 26–53. http://dx.doi.org/10.5539/elt.v7n6p26

Wiklund-Hörnqvist, C., Jonsson, B., & Nyberg, L. (2014). Strengthening concept learning by repeated testing. *Scandinavian Journal of Psychology*, *55*(1), 10-16. http://dx.doi.org/10.1111/sjop.12093

Yeh, Y. C. (2022). Student satisfaction with audio-visual flipped classroom learning: a mixed-methods study. *International Journal of Environmental Research and Public Health*, *19*(3), 1053. https://doi.org/10.3390/ijerph19031053

Yi, Q., Dong, Z. Y., & Qiao, H. (2024). Enhancing EFL listening and speaking skills: strategies and practice for implementing multimedia and multi-modal approaches. *Journal of Humanities, Arts and Social Science*, *8*(9), 2211–2217. https://doi.org/10.26855/jhass.2024.09.032

Yi, Y. (2014). Possibilities and challenges of multimodal literacies in the learning and teaching of second and world languages. *Language and Linguistics Compass, 8*(4), 158–169. http://dx.doi.org/10.1111/lnc3.12076

Yıldırım, A. and Şimşek, H. (2006). *Qualitative research methods in social sciences*. Seçkin. https://doi.org/10.4236/vp.2023.92003

Zhang, R., & Zou, D. (2022). A state-of-the-art review of the modes and effectiveness of multimedia input for second and foreign language learning. *Computer Assisted Language Learning*, *35*(9), 2790-2816. http://dx.doi.org/10.1080/09588221.2021.1896555

https://doi.org/10.17323/jle.2025.18364

Enhancing EFL Students' Idiomatic Competence: A Comparative Analysis of Lexical, Etymological, and Multimodal Approaches

Sulaiman Alnujaidi

Imam Mohammad Ibn Saud Islamic University (IMSIU), Saudi Arabia

ABSTRACT

Background: Given the importance of idiomatic expressions in second language learning on one hand and their complex lexical, syntactic, and semantic characteristics on the other hand, previous studies have proposed numerous instructional models for teaching idioms. However, despite the various instructional approaches available, there is a research gap comparing their effectiveness in enhancing idiomatic competence, which causes inconsistencies in teaching practices and underscores the need to identify the most effective methods for improving idiomatic competence.

Purpose: To examine and compare the impact of three approaches (Lexical, Etymological, and Multimodal) on enhancing EFL students' acquisition of idiomatic expressions and developing their overall idiomatic competence.

Method: The study used a quasi-experimental pre-test, post-test research design with three intact classes. Three groups of EFL students (n=66) were taught idiomatic expressions using different instructional approaches to compare their effectiveness. The first group (23 students) learned idioms through the Lexical Approach, the second group (22 students) through the Etymological Approach, and the third group (21 students) through the Multimodal Approach. Pre- and post-tests were administered to the three groups, and their scores were analyzed using a mixed model ANOVA to determine significant differences in within-subjects effects (changes in idiomatic competence over time) and between-subjects effects (differences among the instructional approaches: Lexical, Etymological, and Multimodal).

Results: The results demonstrated that there were statistically significant differences between the three approaches (Lexical, Etymological, & Multimodal) in terms of their effects on enhancing L2 learners' idiomatic competence. It was found that Multimodal group outperformed Lexical and Etymological groups significantly. No significant difference was found between Lexical and Etymological groups. This study suggests that the Multimodal Approach is highly effective for idiomatic competence development.

Conclusion: The study concluded that EFL teachers, curriculum developers, textbook designers, and researchers could implement the potentials of the Multimodal Approach to create more effective, interactive, and authentic environments for learning idiomatic expressions. Practical implications and further research recommendations were also suggested.

KEYWORDS

english as a foreign language, idiomatic competence, lexical approach, etymological approach, multimodal approach

Citation: Alnujaidi, S. (2025). Enhancing EFL students' idiomatic competence: A comparative analysis of lexical, etymological, and multimodal approaches. *Journal of Language and Education, 11*(2), 57-74. https://doi.org/10.17323/jle.2025.18364

Correspondence:
Sulaiman Alnujaidi,
ssnojeidi@imamu.edu.sa

Received: November 17, 2023
Accepted: June 10, 2025
Published: June 30, 2025

INTRODUCTION

Idiomatic expressions are a pervasive and indispensable component of native speaker discourse, often used to convey ideas succinctly and vividly across diverse communicative settings such as news media, entertainment, and informal speech (Liontas, 2017; Cooper, 1999). Their idiomaticity reflects deep-rooted cultural, historical, and social knowledge, which adds pragmatic depth to spoken

and written language. For learners of English as a foreign language (EFL), acquiring idiomatic competence is not only key to linguistic proficiency but also essential for engaging meaningfully with authentic language use across cultural boundaries (Crowley et al., 2023; McCarthy & O'Dell, 2017).

Despite their ubiquity, idiomatic expressions present persistent challenges for EFL learners due to their metaphorical, non-literal, and culturally embedded meanings (Celce-Murcia & Larsen-Freeman, 1999; Prodromou, 2003). While native speakers use idioms fluidly and subconsciously, second-language learners tend to avoid them or use them inaccurately - a phenomenon described as the «idiomatic paradox» (Fellbaum, 2009). This gap between frequency of use and learner competence has been widely recognized as a stumbling block in achieving communicative fluency and sociolinguistic awareness in EFL contexts (Liontas, 1999; Irujo, 1986).

Over the past two decades, a growing body of research has explored the role of idiomatic expressions in second language acquisition (Boers et al., 2004; Martinez & Schmitt, 2012; Liontas, 2015). Numerous studies have examined the cognitive difficulty of idioms (Irujo, 1986), their contribution to communicative and cultural competence (Cooper, 1999), and strategies for improving idiom retention and production (Boers, 2001; Vasiljevic, 2015). Instructional models based on lexical chunking (Lewis, 1997), etymological elaboration (Boers et al., 2007), and multimodal input (Freyn & Gross, 2017) have all been proposed, each drawing from distinct theoretical traditions.

However, the field remains fragmented. Most studies assess a single approach in isolation, which makes it difficult to evaluate the relative efficacy of competing methods. In addition, findings across studies are often contradictory or context-specific, which limits their generalizability and practical applicability (Birch et al., 2010; Zhang, 2009). As a result, language teachers are left without clear, evidence-based guidelines for selecting the most effective techniques for idiom instruction. This underscores the need for comparative, empirically grounded research that tests multiple instructional approaches within the same experimental design.

In response to this gap, the present study compares the effectiveness of three pedagogical approaches (the Lexical, Etymological, and Multimodal Approaches) for developing idiomatic competence among EFL learners. Each method targets idiom comprehension and retention through a different instructional mechanism: chunk-based memorization, historical-conceptual elaboration, and multisensory input, respectively. By applying a quasi-experimental design across three learner groups, the study offers a direct, systematic comparison of these instructional approaches.

The study addresses the following research questions:

RQ1: Is there any statistically significant difference between the effects of the Lexical, Etymological, and Multimodal approaches on enhancing idiomatic competence?

RQ2: Which of these instructional approaches is most effective in developing EFL students' idiomatic competence?

Findings from this study are expected to inform EFL curriculum design and contribute to a more nuanced understanding of how idioms can be effectively taught in diverse learning contexts.

LITERATURE REVIEW

Defining Idiomatic Competence

Based on the literature, formulaic sequences encompass all pre-fabricated word combinations stored and retrieved holistically from memory (Wray, 2002). Within this broad category, multiword expressions (MWEs) constitute a core subset of (semi-)fixed, recurrent phrases, including collocations, binomials, speech formulae, lexical bundles, and idioms (Siyanova-Chanturia & Martinez, 2015). Idioms are distinguished from other MWEs by their semantic non-compositionality: their conventionalized, figurative meanings (e.g., spill the beans, pull someone's leg) cannot be deduced from the literal meanings of their individual components (Irujo, 1986; Liontas, 2002; Grant & Bauer, 2004), though they exhibit structural diversity (Wolter, 2019) and exclude phrasal verbs.

Crucially, mastering idioms requires idiomatic competence, which extends Canale and Swain's (1980) sociolinguistic competence. Liontas (1999, 2015) defines idiomatic competence as the ability to understand and use idioms accurately and appropriately across varied sociocultural contexts, akin to native speakers. This competence integrates linguistic knowledge (phonology, morphology, syntax, semantics) with pragmatic knowledge (sociolinguistic/functional, discourse, intercultural awareness) (Liontas, 2015), thereby bridging Canale and Swain's grammatical and sociolinguistic components.

This study operationally adopts Liontas' (2015) framework, positioning idioms as a specialized type of MWE defined by non-literal meaning and sociopragmatic embeddedness, whose effective use necessitates idiomatic competence—an advanced dimension of communicative competence rooted in Canale and Swain's foundational model.

Pedagogical Relevance of Idioms

Learning idiomatic expressions significantly enhances L2 learners' communicative, idiomatic, and cultural competence (Cooper, 1999; Crowley et al., 2023; De Caro, 2009;

Liontas, 1999). Furthermore, the ability to use and understand idioms allows learners to engage more fully in authentic conversations, which can boost their confidence and improve their fluency and proficiency (Irujo, 1986; Liontas, 2017). In addition, idioms often reflect the cultural values and societal norms of the language, which provide learners a deeper understanding of the target culture (Bennett, 1997; Boers et al., 2004; Lundblom & Woods, 2012; McCarthy & O'Dell, 2017). Despite their complex meanings and vivid imagery, which can be challenging to grasp (Geeraerts, 2002; Zhang, 2009), the frequent use of idioms in English underscores their importance (Cooper, 1999). The CEFR's inclusion of idioms in language curricula and proficiency tests further highlights their significance in language learning (Council of Europe, 2018[1]; (Iwashita & Vasquez, 2019[2]; Kaneko, 2020; Read & Nation, 2006[3]).

Acquisition of Idiomatic Competence & Teaching Idiomatic Expressions

The lack of a systematic approach to teaching idiomatic expressions in a language classroom setting (Liontas, 2002) has triggered many researchers to propose several approaches and techniques to introduce idioms to L2 learners, enhance their learning of idiomatic expressions, and develop their overall idiomatic competence. Some of the proposed approaches within the scope of this study are: the Lexical Approach, the Etymological Approach, and the Multimodal Approach.

The Lexical Approach

The Lexical Approach (Lewis, 1997) prioritizes teaching prefabricated lexical «chunks» (e.g., idioms, collocations) as holistic units without grammatical decomposition. This method emphasizes lexical chunking and pedagogical simplicity, and advocates teaching idioms be as unanalyzed wholes to promote fluency (Lewis, 2000; Schmitt, 2000). Empirical studies support its efficacy: Tang (2012) reported improved use of lexical chunks in L2 writing, Li (2014) observed reduced L1 transfer in writing, Tang (2013) documented enhanced listening efficiency, and Brenes (2022) noted better comprehension of contextualized idioms. Criticisms highlight its lack of a structured syllabus, described as «a journey without maps» (Thornbury, 1998, p. 11). Skehan (1998, cited in Thornbury, 1998) further cautioned that over-reliance on memorization risks fossilization and limits generative language use.

The Etymological Approach

The Etymological Approach (Boers et al., 2007) employs idioms' historical origins to facilitate cognitive elaboration. Grounded in Dual Coding Theory, which involves verbal and visual processing (Clark & Paivio, 1991), and Levels of Processing Theory, where deeper analysis aids retention (Craik & Lockhart, 1972), the Etymological Approach evokes mental imagery of literal origins (e.g., *spill the beans* linked to ancient voting practices). Boers (2001) found etymological elaboration significantly improved idiom retention compared to contextual guessing. Liontas (2017) emphasized its role in cultural-historical insight. However, Zhang (2009) observed no advantage for receptive knowledge over rote learning, while Szczepaniak and Lew (2011) noted that etymological notes could distract learners or cause meaning confusion. Bakla et al. (2016) and Zarei and Rahimi (2014) similarly reported minimal retention benefits.

The Multimodal Approach

The Multimodal Approach uses digital tools (e.g., videos, apps) to integrate multisensory engagement (visual, auditory, textual modes) and support learner autonomy (Kress, 2003; Anstey & Bull, 2010). Studies confirm its strengths: Khoshnevisan (2019) observed enhanced motivation and idiom retention through multiple modes, Huang et al. (2022) reported superior long-term phrase acquisition versus unimodal input, and Freyn and Gross (2017) documented significant comprehension gains via digital storytelling. Implementation barriers include inequitable technology access, student adaptation difficulties, and increased teacher workload for resource creation (Romero & Bobkina, 2021). Mixed efficacy results exist: Birch et al. (2010) found no significant performance gains despite learner preference, and Cho and Kim (2021) noted no quality differences in multimodal versus monomodal writing outcomes. Teacher training gaps in digital pedagogy further hinder adoption (Romero & Bobkina, 2021).

Despite the volume of research supporting individual methods, findings across studies are inconsistent and context-dependent. Some emphasize retention (Boers et al., 2007), while others prioritize learner engagement (Khoshnevisan, 2019) or processing efficiency (Tang, 2013). Few studies have adopted a comparative lens to empirically test these methods within a controlled design. Moreover, differences in outcome variables (e.g., comprehension vs. production) and assessment tools limit comparability. These limitations

[1] Council of Europe (2018). *Common European Framework of Reference for Languages: Learning, teaching, assessment. Companion volume with new descriptors.* Press Syndicate of the University of Cambridge.

[2] Iwashita, N., & Vasquez, C. (2019). An examination of discourse competence at different proficiency levels in IELTS Speaking Part 2. *IELTS Research Report Series*, 5, 1–44. https://www.ielts.org/for-researchers/research-reports/online-series-2015-5

[3] Read, J., & Nation, P. (2006). An investigation of the lexical dimension of the IELTS speaking test. *IELTS Research Reports*, 6, 207-231. https://www.ielts.org/for-researchers/research-reports/volume-06-report-7

signal the need for integrative studies that can evaluate instructional efficacy across theoretical paradigms.

Taken together, the reviewed literature highlights the pedagogical value of idioms and outlines three major approaches to idiom instruction. However, the absence of comparative empirical research leaves educators without evidence-based guidance. To fill this gap, the present study directly compares the Lexical, Etymological, and Multimodal approaches to determine their relative effectiveness in fostering idiomatic competence among EFL learners.

METHOD

Research Design

The present study used a quasi-experimental pre-test, post-test research design using three intact classes taught by the researcher. One class, comprising 23 students, was taught idiomatic expressions through the Lexical Approach. The second class, including 22 students, was taught idiomatic expressions through the Etymological Approach. The third class, involving 21 students, was taught idiomatic expressions through the Multimodal Approach. Pre- and post-tests were applied to the three groups and the scores of the three groups were compared to determine whether there were any significant differences.

Participants

The total participants in this study were 66 male, college-level EFL students at a public university in Riyadh, Saudi Arabia. Convenience non-random sampling was utilized; therefore, students were placed in three groups based on their classrooms. All the participants were native Arabic speakers studying English as a foreign language. The participants were homogeneous at the pre-intermediate and intermediate proficiency levels, which corresponds to (Level B1) on the Common European Framework of Reference for Languages (CEFRL). All students have studied English for seven years at school. However, their exposure to English in daily life is only limited to classroom environments and media outlets. At college, students are required to take two English courses (ENG-1 and ENG-2) during their first year of college study. These English courses emphasize the development of students' macro-skills (listening, speaking, reading, and writing) as well as their micro-skills (grammar, vocabulary, and pronunciation) to prepare them for their academic courses in their field of study and for using English in their future professions.

Instruments & Procedures

Pre-Test

A pre-test was administered to the three groups on the first week to measure the participants' background knowledge on idiomatic expressions. After conducting the pre-test, the three groups were exposed to the target idiomatic expressions throughout 10 weeks in three different modes: the Lexical Approach group, the Etymological Approach group, and the Multimodal Approach group.

Post-Test

A post-test was conducted on the last week to check the target idiomatic expressions acquired by the participants in all groups and to find out whether they were any differences between the three groups. The post-test lasted for 45 minutes and included 30 multiple-choice items with three choices per idiom in which students were asked to choose the correct interpretation of the idiom. An example is presented below:

> John accidentally spilled the beans about the new project during the meeting.
> The idiomatic expression 'spilled the beans' means:
> a. to prepare a meal
> b. to reveal a secret
> c. to clean up a mess

Validity & Reliability

The validity and reliability of the pre- and post-tests were ensured through content checks and item testing with a separate group of students. The content checks were performed via a review by three EFL experts. These experts ensured that the test items comprehensively covered a representative sample of idiomatic expressions relevant to the study and also checked for face validity. Their feedback confirmed that the items appeared relevant and appropriate for measuring idiomatic competence. Test-retest reliability was evaluated by assessing the items with a subsample of 20 students from the target population who were not part of the study, taking the test with a two-week interval between the tests. A Pearson correlation test was conducted for reliability and returned a score of .83, indicating high stability of the test scores over time.

Procedure

For the purpose of this study, a quasi-experimental, three-group design was used to compare the impact of the three approaches: (1) Lexical, (2) Etymological, and (3) Multimodal on improving EFL students' idiomatic competence. The present study was carried out over a period of 10 weeks and classes met two times a week (20 sessions). A number of 40 idiomatic expressions (see Appendix 1) were presented to the participants in all three groups over 20 sessions (2 idioms per session). The target idioms were selected from several sources including: *101 American English Idioms* (Collis, 1987), *Essential American Idioms* (Spears, 1999), and *Dictionary of Idioms and Their Origin* (Flavell, 1992).

In the Lexical Approach group, students were taught the target idiomatic expressions as whole chunks and with no analysis at the word level. At this stage, more emphasis was put on idiom meaning and usage rather than the meaning of the constituent words or the origin of the idioms as shown in the following example.

> ***Spill the beans***
> ***Meaning:*** *to reveal a secret*
> **Examples*:***
> 1. Emily was so excited about her promotion that she spilled the beans before the official announcement.
> 2. Mark knew he had to be careful not to spill the beans about the confidential meeting.
> 3. We worked hard to keep it a secret, but he spilled the beans and told everyone.

In the Etymological Approach group, students were introduced to the origins of the target idiomatic expressions and the history of the changing meanings and forms of words which constitute those idiomatic expressions as shown in in the following example.

> ***Spill the beans***
> **Meaning**: *to reveal a secret*
> **Etymology**:
> The phrase likely originates from an ancient Greek voting process involving beans. Voters would cast their ballots by placing one of two colored beans in a vase, with white beans typically indicating yes and black or brown beans indicating no. If someone spilled the beans, the secret results of the election would be prematurely revealed. Thus, «spilling the beans» came to mean disclosing confidential information.

In the Multimodal Approach group, students were presented with the target idiomatic expressions through a web-assisted language learning environment (Figure 1) where text, pictures, audios, and videos interactively provide the figurative meaning of each idiomatic expression along with examples, exercises, and quizzes.

In all groups, students were introduced to the linguistic form and conceptual meaning of the idioms. L1 equivalent of the idioms were sometimes provided when students struggle to get the exact figurative meaning of the idiomatic expressions. To effectively correct students' mistakes with idioms, contextual learning was used by incorporating idioms into meaningful sentences and real-life scenarios, and engaging students in role-playing activities to practice idioms in conversation. Gentle corrections and positive reinforcement were provided when idioms were used correctly, and error analysis was conducted to discuss common mistakes and their underlying reasons, helping students understand and avoid these errors in the future. Exercises and quizzes were provided in the form of multiple-choice, matching, and fill-in-the-blank. However, in the Multimodal Approach group, exercises and quizzes were automated online and feedback was provided promptly and individually for students.

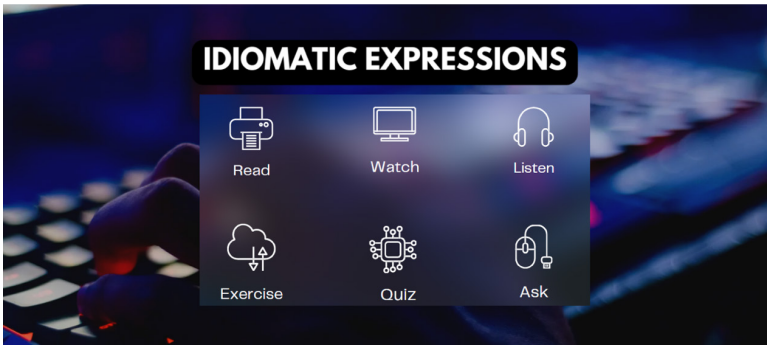
Data Collection and Analysis

Descriptive statistics were calculated to provide an initial summary of the mean and standard deviation of idiomatic competence scores for each instructional approach group at both the pre-test and post-test stages. The mixed model ANOVA, also known as a within-between ANOVA, was utilized to analyze the data and address the research questions due to the study design incorporating both within-subjects factors (repeated measures) and between-subjects factors (independent groups). This type of ANOVA combines elements of both a one-way independent ANOVA and a one-way repeated-measures ANOVA (Tabachnick & Fidell, 2013). It was conducted to assess both the within-subjects effects (changes in *idiomatic competence* over time) and the between-subjects effects (differences among the instructional approaches: *Lexical*, *Etymological*, and *Multimodal*). This statistical method allows for the simultaneous examination of the main effects of time (*pre-test* vs. *post-test*) and instructional group (*Lexical*, *Etymological*, and *Multimodal*), as well as their interaction effect. This comprehensive approach was crucial for determining not only whether there were overall improvements in idiomatic competence but also whether these improvements varied significantly across the different instructional methods.

To identify the specific group differences following the significant findings from the mixed model ANOVA, post hoc comparisons using Tukey's Honestly Significant Difference (HSD) test were conducted. This analysis was essential for pinpointing exactly where the statistically significant differences occurred between the groups. By comparing each

Figure 1

A web-assisted language learning environment for Idiomatic Expressions



pair of groups, the post hoc analysis provided a detailed understanding of the relative effectiveness of each instructional approach. These analyses collectively provided a robust framework for evaluating the instructional interventions, ensuring that the findings were both statistically rigorous and practically meaningful.

RESULTS

Descriptive Statistics of Idiomatic Competence Development

Descriptive statistics (Table 1) were calculated to provide an overview of the mean and standard deviation of idiomatic competence scores for each group at both pre-test and post-test stages. This helps in understanding the initial levels of competence and the improvement after the intervention. The results show that all groups improved their scores from pre-test to post-test, indicating that each instructional approach had a positive effect on idiomatic competence. Notably, the Multimodal group exhibited the highest mean improvement, suggesting that while all instructional methods are beneficial, the Multimodal approach appears to offer the greatest enhancement.

Verification of Statistical Assumptions for Mixed-Model ANOVA

As this study employed the mixed model ANOVA, which includes both the within-subjects and between-subjects effects, it was necessary to meet three assumptions to ensure the equality of variances. The first assumption was normality, which ensures that the data are normally distributed. The second assumption was homogeneity of variance, relevant for between-subjects factors, to ensure that the variances within each group (e.g., the Lexical, Etymological, and Multimodal groups) were equal. The third assumption was sphericity, relevant for within-subjects factors, to ensure that the variances of the differences between all possible pairs of within-subject conditions (e.g., pre-test and post-test) were equal.

To assess the assumption of normality of the pre-test and post-test scores for each group, Shapiro-Wilk test (Table 2) were conducted. The non-significant results ($p > 0.05$) indicate that the data is normally distributed and does not violate this assumption.

To test the assumption of homogeneity of variances, Levene's Test (Table 3) was conducted. The non-significant p-value of the pre-test scores ($p = 0.200$) as well as the non-significant p-value of the post-test scores ($p = 0.127$) indicate that the assumption of homogeneity of variances has been met. These results confirm that the variances within

Table 1

Descriptive Results

	Group	N	M	SD
Pre-test	Lexical	23	15.34	3.35
	Etymological	22	14.04	2.83
	Multimodal	21	13.61	2.74
Post-test	Lexical	23	20.13	2.86
	Etymological	22	21.22	3.23
	Multimodal	21	25.14	1.15

Table 2

Tests of Normality

		Shapiro-Wilk Statistic		
Test	Group	Statistic	df	Sig.
Pre-test	Lexical	0.970	23	0.658
	Etymological	0.975	22	0.781
	Multimodal	0.968	21	0.549
Post-test	Lexical	0.985	23	0.945
	Etymological	0.979	22	0.873
	Multimodal	0.990	21	0.988

each instructional group are equal, satisfying the assumption of homogeneity of variances necessary for conducting the mixed model ANOVA.

To test the assumption of sphericity, Mauchly's Test (Table 4) was calculated. The non-significant p-value for the Time effect (p = 0.059) and Time*Group effect (p = 0.060) indicate that the assumption of sphericity has not been violated for either effect. Therefore, it was concluded that the variances of the differences between the conditions were equal, meeting the assumption of sphericity required for performing the mixed model ANOVA.

Mixed ANOVA Analysis of Within-Subjects and Between-Subjects Effects

The mixed model ANOVA, as shown in Table 5, was calculated to examine both the within-subjects effects (changes in idiomatic competence over time) and the between-subjects effects (differences between groups: *Lexical*, *Etymological*, and *Multimodal*) in a comprehensive manner. The significant main effect of time ($F(1, 64) = 234.35$, $p < .001$, $\eta^2 = .781$) indicates that idiomatic competence scores improved significantly from pre-test to post-test across all groups, reflecting the overall effectiveness of the instructional interventions. The large effect size ($\eta^2 = .781$) suggests that a substantial proportion of the variance in idiomatic competence scores is attributable to the passage of time, reflecting the overall effectiveness of the instructional interventions. The significant interaction effect between time and group ($F(2, 64) = 20.88$, $p < .001$, $\eta^2 = .402$) suggests that the improvement in idiomatic competence varied significantly depending on the instructional approach. The large effect size ($\eta^2 = .402$) suggests that the type of instructional intervention substantially influences how idiomatic competence scores change from pre-test to post-test. The significant intercept ($F(1, 64) = 1517.47$, $p < .001$, $\eta^2 = .959$) reflects the overall level of idiomatic competence across all measurements, with a very large effect size ($\eta^2 = .959$). This high effect size indicates that the baseline level of idiomatic competence is consistently high across the sample. The significant main effect of instructional approach ($F(2, 64) = 9.82$, $p < .001$, $\eta^2 = .235$) indicates that there are significant differences in idiomatic competence among the three instructional approach groups. The medium effect size ($\eta^2 = .235$) indicates that the instructional approach accounts for a notable proportion of the variance in idiomatic competence scores. Overall, the significant main effects and interaction effects highlight the impact of the instructional approaches on enhancing idiomatic competence, showing notable improvements over time and differences between the groups. These findings imply that while all instructional approaches improve idiomatic competence, their effectiveness varies, with the interaction effect highlighting the differing impact over time.

The post hoc comparisons using Tukey's Honestly Significant Difference (HSD) test, shown in Table 6, further clarify the differences between the instructional approaches. The results reveal that the Multimodal group significantly outperformed both the Lexical and Etymological groups ($p < .001$, Cohen's $d = 1.23$ for Lexical vs. Multimodal, Cohen's $d = 1.10$ for Etymological vs. Multimodal). However, there was no statistically significant difference between the Lexical and Etymological groups ($p = .265$, Cohen's $d = 0.30$). These findings imply that the Multimodal approach is significantly more effective than both the Lexical and Etymological approaches in enhancing idiomatic competence, while the Lexical and Etymological approaches are equally effective but less so compared to the Multimodal Approach. This underscores the superior efficacy of the Multimodal Approach in improving idiomatic competence among EFL students.

In summary, the combined results from these tables indicate that all instructional approaches positively impact idiomatic competence, with significant improvements observed over time. The Multimodal Approach is particularly effective, as demonstrated by its higher post-test scores and large effect sizes in the post hoc comparisons. These findings suggest that incorporating multiple modes of learning may provide

Table 3

Homogeneity of Variances (Levene's Test)

Group	Levene Statistic	df1	df2	Sig.
Pre-Test Scores	1.65	2	63	0.200
Post-Test Scores	2.12	2	63	0.127

Table 4

Sphericity (Mauchly's Test)

Effect	Mauchly's W	Approx. Chi-Square	df	Sig.
Time	0.914	5.678	2	0.059
Time * Group	0.876	8.132	2	0.060

a more comprehensive and effective method for teaching idiomatic expressions in a foreign language context.

DISCUSSION

Based on the study findings, there were statistically significant differences between the Lexical, Etymological, and Multimodal approaches in terms of their effect on developing L2 learners' idiomatic competence. It was found that students' idiomatic competence in the Multimodal Approach group was significantly better than their counterparts in the Lexical Approach group and the Etymological Approach group. It was also concluded that there was no statistically significant difference between the Lexical Approach group and the Etymological Approach group; thus implying that these two approaches had no difference in their effect on L2 students' idiomatic competence.

However, the high mean score observed in the Multimodal Approach group requires careful interpretation. Although the Multimodal Approach group exhibited a mean of 25.14, which was significantly different from other groups, this mean score accounts for only 62.5% of the total 40 items, which indicates that the overall performance may not necessarily reflect substantial learning. In other words, even though there was a measurable change in performance, it may not represent a substantial improvement in participants' understanding or retention of idiomatic expressions. Several factors could contribute to this outcome. One potential explanation is the presence of a history effect. In this case, the treatment involved an extended duration, spanning ten weeks, during which participants were exposed to a total of 40 idioms (equating to an average of 4 idioms taught per week). This prolonged exposure to the material could have influenced participants' performance on the post-test. Therefore, while the mean score of the Multimodal Approach group appears to differ significantly from other groups, the context of the study design, including the duration and intensity of the intervention, should be taken into account when interpreting these findings. Further analysis and consideration of other variables, such as individual learning styles and engagement levels, may provide additional insights into the observed outcomes.

The results of this study are consistent with Liontas (2015) who concluded that multimodal learning environments facilitate idiomatic competence and make learning idiomatic expressions more authentic, more natural, and more effective as well as adhere to learners' individual differences. The findings are also analogous with Khoshnevisan (2019) who concluded that multimodality (audio and visual) af-

Table 5

Mixed Model ANOVA Results

Source	*SS*	*df*	*MS*	*F*	*p-value*	η^2
Within-Subjects Effects						
Time	824.48	1	824.48	234.35	.000	.781
Time*Group	146.87	2	73.44	20.88	.000	.402
Error (Time)	227.40	64	3.55			
Between-Subjects Effects						
Intercept	40779.60	1	40779.60	1517.47	.000	.959
Group	823.34	2	411.67	9.82	.000	.235
Error (Group)	1718.30	64	26.85			

Table 6

Tukey's HSD Post Hoc Tests

(I) Group	(J) Group	*Mean Difference (I-J)*	*Std. Error*	*Sig.*	*Cohen's d*
Lexical	Etymological	-1.09	0.49	.265	0.30
	Multimodal	-5.01*	0.49	.000	1.23
Etymological	Lexical	1.09	0.49	.265	0.30
	Multimodal	-3.92*	0.49	.000	1.10
Multimodal	Lexical	5.01*	0.49	.000	1.23
	Etymological	3.92*	0.49	.000	1.10

fords teachers with diverse instructional modes that meet their students' needs, accommodate their diverse learning styles (auditory and visual), and contextualize idiom learning. In addition, the results are in accordance with those obtained by Fadel (2008) who found that learning through a multimodal-based approach that integrates multiple instructional modes results in better performance than a traditional-based approach that incorporates one mode only. The findings of the current study echo those obtained by Zhang (2021) who argued that integrating many modes, which trigger multi-sensory stimuli, can promote students' motivation, increase their auditory and visual input, and enhance their idiom recognition and understanding.

The conclusions drawn from the analysis underscore the critical role of the instructional approach in the effective teaching and learning of idiomatic expressions, which are essential for developing idiomatic competence and overall linguistic proficiency. Despite their importance, idiomatic expressions pose significant challenges in EFL classroom contexts. The inherent complexity of idiomatic expressions, with their figurative meanings diverging from literal interpretations, often creates barriers for learners. These barriers can lead to misunderstandings, confusion, and a tendency to avoid using idioms altogether. To address these challenges, it is crucial to explore and implement alternative instructional methods. Such methods may include contextual learning, visual aids, and interactive activities, which can facilitate a smoother transition from literal to figurative comprehension and improve overall idiomatic expression instruction. By focusing on innovative teaching strategies, educators can enhance students' understanding and usage of idiomatic expressions, which in turn fosters better linguistic competence and confidence in language use.

The findings highlight the remarkable efficacy of the Multimodal Approach in facilitating students' recognition and comprehension of idiomatic expressions. By incorporating diverse sensory modalities and interactive elements into the learning process, this approach offers students a multifaceted and engaging learning experience that fosters deeper understanding and retention of idiomatic structures and meanings. In contrast, the study reveals limited effectiveness in the chunking-based (Lexical) and origin-based (Etymological) approaches in enhancing idiomatic comprehension and production. Despite their potential theoretical underpinnings, these traditional methods appear insufficient in addressing the nuanced complexities of idiomatic language for learners. These findings suggest that incorporating multimedia resources, interactive activities, and real-world contextualization into instruction can effectively alleviate the perceived complexity of idiomatic syntactic and semantic structures.

Overall, the conclusions drawn from the study underscore the pressing need for innovative and adaptive instructional strategies that cater to the diverse needs and learning preferences of students while effectively addressing the inherent challenges of idiomatic expression instruction. Through continued exploration and refinement of instructional approaches, EFL educators, teachers, and researchers can better equip learners with the idiomatic competence necessary for effective communication and language proficiency.

CONCLUSION

This study provides empirical evidence supporting the effectiveness of the Multimodal Approach in enhancing EFL learners' idiomatic competence. By systematically comparing three instructional approaches, the research demonstrates that multimodal instruction leads to significantly better learner outcomes than both lexical chunking and etymological elaboration. This finding adds comparative depth to the existing literature, which has largely evaluated these methods in isolation.

The findings hold practical value for EFL practitioners seeking to address one of the most persistent challenges in language instruction: learners' difficulty with figurative and idiomatic language. Multimodal instruction (through its integration of visual, auditory, and contextual inputs) appears particularly well-suited to support learners with diverse cognitive profiles and preferences. Beyond teaching practice, the study also has implications for syllabus design and educational technology, which suggests that idiom learning can be improved by leveraging multimedia resources and interactive pedagogies.

While the results are promising, they should be interpreted within the context of certain limitations, including the homogeneity of the sample, the lack of delayed post-testing, and potential differences in instructional time across groups. Future research could extend these findings by incorporating a broader demographic base, exploring long-term retention, and integrating learner feedback on instructional experiences. Additional studies might also examine how multimodal instruction interacts with idiom type, proficiency level, and cognitive learning style.

Ultimately, this study contributes to a growing body of work that calls for more innovative, context-responsive approaches to idiom instruction in EFL contexts. By demonstrating the comparative advantages of multimodal learning, it lays the groundwork for a more learner-centered, inclusive, and effective model of figurative language teaching - one that better prepares students for real-world communication and intercultural competence.

ACKNOWLEDGMENTS

The author would like to thank the anonymous reviewers for their invaluable and insightful comments on the earlier version of this paper which have led to considerable improvements to the current version.

FUNDING STATEMENT

The author declares that this article did not receive any funding or financial support.

DECLARATION OF COMPETING INTEREST

None declared.

REFERENCES

Alexander, R. (1987). Problems in understanding and teaching idiomaticity in English. *Anglistik and Eneglichunterricht, 32*(2), 105-122.

Anstey, M., & Bull, G. (2010). Helping teachers to explore multimodal texts. *Curriculum & Leadership Journal, 8*(16), 1-4.

Bennett, M. J. (1997). How not to be a fluent fool: Understanding the cultural dimensions of language. In A. E. Fantini, (Vol. Ed.) & J. C. Richards (Series Ed.). *New ways in teaching culture. New ways in TESOL series II: Innovative classroom techniques* (pp. 16– 21). TESOL.

Bakla, A., Çekiç, A., & Demiröz, H., (2016). Learning English idioms through reading in an LMS etymological notes versus pictorial support. *Erzincan Üniversitesi Eğitim Fakültesi Dergisi, 18*(1), 445-462. https://doi.org/10.17556/jef.88778

Bergstrand, N. (2017). *Etymological elaboration versus written context : A study of the effects of two elaboration techniques on idiom retention in a foreign language*. Halmstad University.

Bircan, P. (2010). *Lexical approach in teaching vocabulary to young language learners*. (Publication No. 28635751) [Master's thesis, Anadolu University]. ProQuest Dissertations and Theses Global.

Birch, D., Sankey, M., & Gardner, M. (2010). The impact of multiple representations of content using multimedia on learning outcomes. *International Journal of Instructional Technology and Distance Learning, 7*(4), *3-19.*

Boers, F. (2001). Remembering figurative idioms by hypothesising about their origin. *Prospect, 16*(3), 35–43.

Boers, F., Demecheleer, M., & Eyckmans, J. (2004). Etymological elaboration as a strategy for learning idioms. In P. Bogaards & B. Laufer (Eds.), *Vocabulary in a second language: Selection, acquisition and testing* (pp. 53-78). John Benjamins Publishing. https://doi.org/10.1075/lllt.10.07boe

Boers, F., Eyckmans, J., & Stengers, H. (2007). Presenting figurative idioms with a touch of etymology: more than mere mnemonics? *Language Teaching Research, 11*(1), *43-62.* https://doi.org/10.1177/1362168806072460

Boers, F., & Webb, S. (2015). Gauging the semantic transparency of idioms: Do natives and learners see eye to eye? In R. Heredia & A. Cieślicka (Eds.), *Bilingual figurative language processing* (pp. 368–392). Cambridge University Press.

Brenes, C. A. N. (2022). Implementing the lexical approach in an integrated English e-course. *Research in Pedagogy*, *12*(1), 60-81. https://doi.org/10.5937/IstrPed2201060N

Brown, H. D. (1994). *Principles of language learning and teaching*. Prentice Hall Regents.

Buckingham, L. (2006). A multilingual didactic approach to idioms using a conceptual framework. Language Design. *Journal of Theoretical and Experimental Linguistics, 8*, 35- 45.

Cain, K., Oakhill, J., & Lemmon, K. (2005). The relation between children's reading comprehension level and their comprehension of idioms. *Journal of Experimental Child Psychology, 90*(1), 65–87. https://doi.org/10.1016/j.jecp.2004.09.003

Canale, M., & Swain, M. (1980). Theoretical bases of communicative approaches to second language teaching and testing. *Applied Linguistics, 1*, 1–47. https://doi.org/10.1093/applin/I.1.1

Celce-Murcia, M. & Larsen-Freeman, D. (1999). *The grammar book: An ESL/EFL teacher's course*. Heinle & Heinle.

Cho, H., & Kim, Y. (2021). Comparing the characteristics of EFL students' multimodal composing and traditional monomodal writing: The case of a reading-to-write task. *Language Teaching Research*, *28*(6). https://doi.org/10.1177/13621688211046740

Clark, J. M., & Paivio, A. (1991). Dual coding theory and education. *Educational Psychology Review, 3*, 149e210. https://doi.org/10.1007/BF01320076

Collis, H. (1987). *101 American idioms*. NTC Publishing Group.

Cooper, T. c. (1999). Processing of idioms by L2 learners of English. *TESOL Quarterly, 33*(2), 233-262. https://doi.org/10.2307/3587719

Craik, F. I. M., & Lockhart, R. S. (1972). Levels of processing: A framework for memory research. *Journal of Verbal Learning and Verbal Behavior*, *11*(6), 671-684. https://doi.org/10.1016/S0022-5371(72)80001-X

Crowley, K., Haugh, S., & Spring, R. (2023). An examination of correlations between multiword expression interpretability and general proficiency test scores. *APU Journal of Language Research, 8*, 46–61. https://doi.org/10.34409/apujlr.8.1_47

De Caro, E. E. R. (2009). The advantages and importance of learning and using idioms in English. *Cuadernos de Lingüística Hispánica, 14*, 121-136.

Dressman, M. (2019). Multimodality and language learning. In Dressman, M., & Sadler, R. W. (Eds.), *The handbook of informal language learning* (pp. 39-56). Wiley-Blackwell. https://doi.org/10.1002/9781119472384.ch3

Fadel, C. (2008). *Multimodal learning through media: What the research says*. Cisco Systems.

Fellbaum, C. (2007). *Idioms and collocations: Corpus-based linguistic and lexicographic studies*. Continuum.

Flavell, L. & Flavell, R. (1992). *Dictionary of idioms and their origins*. Kyle Cathie Limited.

Freyn, A., L., & Gross, S. (2017). An empirical study of Ecuadorian university EFL learners' comprehension of English idioms using a multimodal teaching approach. *Theory and Practice in Language Studies, 7*(1), 984-989. http://dx.doi.org/10.17507/tpls.0711.06

Geeraerts, D. (2002). The interaction of metaphor and metonymy in composite expressions. *Metaphor and Metonymy in Comparison and Contrast, 20*, 435. https://doi.org/10.1515/9783110219197.3.435

Gilakjani, A.P., Ismail, H.N., & Ahmadi, S.M. (2011). The effect of multimodal learning models on language teaching and learning. *Theory and Practice in Language Studies*, *1*(10), 1321-1327. http://doi.org/10.4304/tpls.1.10.1321-1327

Grant, L., & Bauer, L. (2004). Criteria for re-defining idioms: Are we barking up the wrong tree? *Applied Linguistics*, *25*(1), 38–61. https://doi.org/10.1093/applin/25.1.38

Huang, Y., Zhang, Z., Liu, J. Y., & Huang, Y. (2022). English phrase learning with multimodal input. *Frontiers in Psychology, 13*. https://doi.org/10.3389/fpsyg.2022.828022

Irujo, S. (1986). A piece of cake: Learning and teaching idioms. *ELT Journal, 40*(3), 236–242. https://doi.org/10.1093/elt/40.3.236

Jewitt, C., (2008). Multimodality and literacy in school classrooms. *Review of Research in Education,* 32(1), pp. 241-267. https://doi.org/10.3102/0091732X07310586

John, & Smithback, C. Y. (1991). *Fun with idioms*. Federal Publications.

Kalantzis, M., & Cope, B. (2008). Digital communications, multimodality and diversity: Towards a pedagogy of multiliteracies. *Scientia Paedagogica Experimentalis, 45*(1), 15–50.

Kaneko, M. (2020). Lexical frequency profiling of high-stakes English tests: Text coverage of Cambridge First, EIKEN, GTEC, IELTS, TEAP, TOEFL, and TOEIC. *JACET Journal*, *64*, 79–93. https://doi.org/10.32234/jacetjournal.64.0_79

Khoshnevisan, B. (2019). Spilling the beans on understanding English idioms using Multimodality: An idiom acquisition technique for Iranian language learners. *International Journal of Language, Translation and Intercultural Communication, 8*, 128-143. http://dx.doi.org/10.12681/ijltic.20281

Knowles, L. (2004). The evolution of CALL. *The Journal of Communication & Education: Language Magazine*, *3(12)*, 20-23.

Kress, G. (2003). *Literacy in the new media age*. Routledge.

Lennon, P. (1998). Approaches to the teaching of idiomatic language. *International Review of Applied Linguistics in Language Teaching, 36*(1), 1-11. https://doi.org/10.1515/iral.1998.36.1.11

Lewis, M. (1997). *Implementing the lexical approach: Putting theory into practice*. Language Teaching Publications.

Lewis, M. (2000). *Teaching collocation: Further developments in the lexical approach*. Language Teaching Publications.

Li, Q. (2014). An empirical study on the application of lexical chunk to college English writing. *Journal of Language Teaching and Research, 5*(3), 682–688. https://doi.org/10.4304/jltr.5.3.682-688.

Liontas, J. I. (1999). *Developing a pragmatic methodology of idiomaticity: The comprehension and interpretation of SL vivid phrasal idioms during reading* [Doctoral dissertation, University of Arizona]. UA Repository. http://hdl.handle.net/10150/284736

Liontas, J. I. (2015). Developing idiomatic competence in the ESOL classroom: A pragmatic account. *TESOL Journal, 6*(4), 621-658. https://doi.org/10.1002/tesj.230

Liontas, J. I. (2002). Vivid phrasal idioms and the lexical-image continuum. *Issues in Applied Linguistics, 13*(1), 71–109. https://doi.org/10.5070/L4131005053

Liontas, J. I. (2017). Why teach idioms? A Challenge to the Profession. *Iranian Journal of Language Teaching Research, 5*(3), 5-25.

Littlemore, J., & Low, G. (2006). Metaphoric competence, second language learning, and communicative language ability. *Applied Linguistics, 27*(2)268–294. https://doi.org/10.1093/applin/aml004

Lundblom, E., & Woods, J. (2012). Working in the classroom: Improving idiom comprehension through classwide peer tutoring. *Communication Disorders Quarterly, 33*, 202–219. https://doi.org/10.1177/1525740111404927

Martinez, R., & Schmitt, N. (2012). A Phrasal Expressions List. *Applied Linguistics, 33*(3), 299–320. https://doi.org/10.1093/applin/ams010

McCarthy, M., & O'Dell, F. (2017). *English Idioms in Use*. Cambridge University Press.

Moon, R. (1998). *Fixed expressions and idioms in English: A corpus-based approach*. Oxford University Press.

Nattinger, J. R. & DeCarrico, J. S. (1992). *Lexical phrases and language teaching*. Oxford University Press.

Prodromou, L. (2003). Idiomaticity and the non-native speaker. *English Today,19*(2), 42–48. https://doi.org/10.1017/S0266078403002086

Pintado, B. R., & Fajardo, T. (2021). Learning idioms through the multimodal approach. *Journal of Education and Practice, 12*(24). http://doi.org/10.7176/JEP/12-24-03

Richards, J. C., Platt, J., & Platt, H. (1992). *Longman Dictionary of language teaching and applied linguistics*. Longman.

Richards, J. C. & Rodgers, T. S. (2001) *Approaches and methods in language teaching*. Foreign Language Teaching and Research Press and Cambridge University Press.

Romero, E. D., & Bobkina, J. (2021). Exploring the perceived benefits and drawbacks of using multimodal learning objects in pre-service English teacher inverted instruction. *Education and Information Technologies, 26,* 2961–2980. https://doi.org/10.1007/s10639-020-10386-y

Schmitt, N. (2000) Key concepts in ELT: lexical Chunks. *ELT Journal 54*(4), 400-401. https://doi.org/10.1093/elt/54.4.400

Sinclair, J. (1991). *Corpus, concordance, collocation.* Oxford University Press.

Siyanova-Chanturia, A. (2015). On the 'holistic' nature of formulaic language. *Corpus Linguistics and Linguistic Theory*, 11(2), 285-301. https://doi.org/10.1515/cllt-2014-0016

Siyanova-Chanturia, A., & Martinez, R. (2015). The idiom principle revisited. *Applied Linguistics*, *36*(5), 549–569. https://doi.org/10.1093/applin/amt054

Spears, R. (1999). *Essential American idioms*. NTC Publishing Group.

Spring, R., & Takeda, J. (2024). Teaching phrasal verbs and idiomatic expressions through multimodal flashcards: Confirming the importance of multimedia-based instruction. *STEM Journal, 25*(2), 40–53. https://doi.org/10.16875/stem.2024.25.2.40

Strong, B., & Leeming, P. (2024). Evaluating the application of a gap-fill exercise on the learning of phrasal verbs: Do errors help or hinder learning? *TESOL Quarterly, 58*(2), 726-750. https://doi.org/10.1002/tesq.3248.

Szczepaniak, R., & Lew, R. (2011). The role of imagery in dictionaries of idioms. *Applied Linguistics, 32*(3), 323–347. http://dx.doi.org/10.1093/applin/amr001

Tabachnick, B. G., & Fidell, L. S. (2013). *Using multivariate statistics* (6th ed.). Allyn and Bacon.

Taylor, L. K., Bernhard, J. K., Garg, S., & Cummins, J. (2008). Affirming plural belonging: Building on students' family-based cultural and linguistic capital through multiliteracies pedagogy. *Journal of Early Childhood Literacy, 8*(3), 269–294. https://doi.org/10.1177/1468798408096481

Tang, J. (2012). An empirical study on the effectiveness of the lexical approach to improving writing in SLA. *Journal of Language Teaching and Research, 3*(3), 578–583. https://doi.org/10.4304/jltr.3.3.578-583.

Tang, J. (2013). Input of chunks and its effects on l2 learners' listening competency. *Theory and Practice Language Studies, 3*(7), 1264-1269. https://doi.org/10.4304/TPLS.3.7.1264-1269

Thornbury, S. (1998). The lexical approach: A journey without maps? *Modern English Teacher*, *7*(4), 7–13.

Vasiljevic, Z. (2015). Effects of etymology and pictorial support on the retention and recall of L2 idioms. *Electronic Journal of Foreign Language Teaching, 12*(1), 35-55. https://e-flt.nus.edu.sg/v12n12015/vasiljevic.pdf

Wolter, B. (2019). Key issues in teaching multiword items. In S. Webb. (Ed.), *The Routledge handbook of vocabulary studies* (pp. 493-510). Routledge. https://doi.org/10.4324/9780429291586-31

Wray, A. (2002). *Formulaic language and the lexicon.* Cambridge University Press.

Wray, A., & Perkins, M. R. (2000). The functions of formulaic language: An integrated model. *Language and Communication*, *20*(1), 1–28. https://doi.org/10.1016/S0271-5309(99)00015-4

Zarei, A. A., & Rahimi, N. (2014). Etymology, contextual pragmatic clues, and lexical knowledge in L2 idioms learning. *Iranian Journal of Applied Language Studies, 6*(2), 179-203. http://doi:10.22111/ijals.2014.2193

Zhang, L. (2009). *The effect of etymological elaboration on L2 idiom acquisition and retention in an online environment* [Master's thesis, Iowa University]. Digital Repository. https://doi.org/10.31274/etd-180810-1881

Zhang, Q. (2021). A study on the application of multimodal metaphor in English idiom teaching. In *2nd International Conference on Education Studies: Experience and Innovation (ICESEI 2021)* (pp. 111-115). Atlantis Press. https://doi.org/10.2991/assehr.k.211217.017

APPENDIX 1

A List of 40 Metaphorical Expressions Used in the Study

	English	Figurative Meaning
1	*"a dead-end street"* e.g. *Their relationship is a dead-end street.*	experiencing very severe difficulties and looks likely to end very soon
2	*"a guinea pig"* e.g. *He volunteered to act as a guinea pig in the experiment.*	a subject of research, experimentation, or testing
3	*"add fuel to the fire"* e.g. *John only added fuel to the fire when he accused the other team of cheating.*	to do or say something that makes a miserable situation even worse
4	*"an open book"* e.g. *Sarah's feelings were written all over her face; she was truly an open book to everyone around her.*	to have nothing to hide
5	*"beat a dead horse"* e.g. *He keeps trying to explain it to him but I think he's beating a dead horse.*	to waste effort on something when there is no chance of succeeding
6	*"beat around the bush"* e.g. *Don't beat around the bush and tell me frankly what you think of my proposition.*	avoid talking about the main topic; not speaking directly or precisely; avoid the important point
7	*"behind the scenes"* e.g. *A lot of hard work has been going on behind the scenes.*	to do something secretly rather than publicly
8	*"bend over backwards"* e.g. *He bent over backwards trying to please his potential clients so that they would give him the contract.*	to do all in one's power (usually to achieve something or accommodate somebody); to make every effort to do something, especially to help someone
9	*"between the hammer and the anvil"* e.g. *She was between the hammer and the anvil when her parents asked her to choose between going to college or getting married.*	Facing two equally unpleasant, dangerous, or risky alternatives, where the avoidance of one ensures encountering the harm of the other.
10	*"bite off more than one can chew"* e.g. *By accepting two part-time jobs, he is clearly biting off more than he can chew.*	to try to do more than one is able to do; to attempt to do something which is hardly achievable
11	*"build castles in the air"* e.g. *I told him he should stop building castles in the air and train for a sensible profession.*	having extravagant hopes and plans that will never be carried out and entertaining daydreams that will never come to pass
12	*"burn the candle at both ends"* e.g. *Joseph's been burning the candle at both ends for weeks, working two jobs during the week and a third on weekends.*	To work very hard and for long hours, especially till late at night or the early hours of the morning
13	*"By the skin of one's teeth"* e.g. *We managed to complete the project on time by the skin of our teeth.*	to barely succeed or survive; a situation from which one has barely managed to escape or achieve something
14	*"by the sweat of one's brow"* e.g. *He earned his money by the sweat of his brow.*	by one's own hard work

	English	Figurative Meaning
15	*"chip off the old block"* e.g. *Stephen is a chip off the old block. He's a good football player, just like his father.*	someone who is similar to one's parents in behavior, character, or personality
16	*"cry over spilled milk"* e.g. *It's no use crying over spilled milk; it was a bad investment, the money has been lost, and there's nothing we can do.*	to cry about past events that cannot be undone; to feel sorry about something that has already happened
17	*"fan the flames"* e.g. *My sarcastic comment only fanned the flames during the argument I had with my wife.*	to make a situation worse
18	*"get caught red-handed"* e.g. *Tom was stealing the car when the police drove by and caught him red-handed.*	to get caught in the middle of doing something illegal or forbidden
19	*"get under my skin"* e.g. *The new manager is getting under my skin.*	to be irritating; to bother a person; to annoy someone
20	*"hit the nail on the head"* e.g. *Stephen hit the nail on the head when he said that what the company was lacking in was clear vision and focus.*	to say, do, or get something that is exactly right
21	*"jump through hoops"* e.g. *We had to jump through hoops to get my Dad admitted to hospital.*	to have to do a lot of things that seem difficult or unnecessary in order to achieve something
22	*"keep the wolf from the door"* e.g. I don't earn much but it's enough to keep the wolf from the door.	to have just enough money for basic things like food and somewhere to live
23	*"kill two birds with one stone"* e.g. *I killed two birds with one stone and picked the kids up on the way to the supermarket.*	to succeed in achieving two things in a single action
24	*"miss the train"* *e.g. He missed the train on that huge project.*	miss an opportunity
25	*"on cloud nine"* e.g. *Jim has been on cloud nine since his team won the game.*	overjoyed and extremely excited & happy
26	*"paint the town red"* e.g. *We are getting all dressed up next week and we are going to paint the town red*.	to celebrate and have a wild time; to go out and have a lot to drink
27	*"pull his leg"* e.g. *I think he was just pulling your leg when he said you've failed in the exam.*	to make fun of, fool, or tease someone
28	*"put all one's eggs in one basket"* e.g. *I'm applying for several jobs because I don't really want to put all my eggs in one basket.*	to depend for your success on a single person or plan of action
29	*"roll with the punches"* e.g. *Strong industries were able to roll with the punches during the recession.*	to be able to deal with a series of difficult situations
30	*"spill the beans"* e.g. *The employees spilled the beans about the mega project.*	to reveal a secret

	English	Figurative Meaning
31	*"spread oneself too thin"* e.g. *Working two jobs and trying to maintain a social life, Mary often feels like she's spreading herself too thin.*	to try to do too many things at the same time, so that you cannot give enough time or attention to any of them
32	*"swim with the tide"* *e.g. When I was a teenager, I had some radical opinions but I had to swim with the tide then.*	to go along or agree with the prevailing or popularly held opinion or perspective
33	*"take a shot in the dark"* e.g. *I don't have the map, but I'll take a shot in the dark and try to find the restaurant on my own.*	trying something without having all the necessary facts or details; taking a risk without a clear understanding of the potential outcomes
34	*"take the bull by the horns"* e.g. *You should take the bull by the horns and tell him to leave.*	hold things tightly
35	*"the apple of his eye"* e.g. *Tom's youngest daughter is the apple of his eye.*	someone whom you cherish above all others
36	*"the elephant in the room"* e.g. The company is focusing on the wrong issue and ignoring the elephant in the room.	An obvious truth or fact that is being intentionally ignored or left unaddressed
37	*"there's more than one way to skin a cat"* e.g. *I appreciate that you want to help me lose weight, but there's more than one way to skin a cat.*	there are many ways to do something, there are many ways to achieve a goal
38	*"turn a blind eye"* *e.g. The principal decided to turn a blind eye to the student's misconduct this time with a hope that they won't do it again.*	To deliberately overlook; to intentionally ignore something; to bend the rules; to make an exception
39	*"wet behind the ears"* e.g. *The new salesman's amateur techniques proved to everybody at the meeting that he was wet behind the ears.*	immature or poor skill; to be inexperienced; to be new at something or somewhere and so lack the necessary experience
40	*"with an iron hand"* e.g. *For ten years, the President ruled with an iron hand.*	to rule, manage, or control a group of people very firmly, having complete power over everything they do

APPENDIX 2

A Sample of the Pre- and Post-tests

1. *Their relationship is a dead-end street.*
The idiomatic expression "a dead-end street" means:
    a) a place with no exit
    b) a relationship with no future
    c) a safe place to rest

2. *He volunteered to act as a guinea pig in the experiment.*
The idiomatic expression "a guinea pig" means:
    a) a small furry animal
    b) a curious person
    c) a test subject

3. *John only added fuel to the fire when he accused the other team of cheating.*
The idiomatic expression "add fuel to the fire" means:
    a) to make a bad situation worse
    b) to help someone win
    c) to encourage teamwork

4. *Sarah's feelings were written all over her face; she was truly an open book to everyone around her.*
The idiomatic expression "an open book" means:
    a) a very emotional person
    b) a person who hides nothing
    c) a good writer

5. *He keeps trying to explain it to him but I think he's beating a dead horse.*
The idiomatic expression "beating a dead horse" means:
    a) to insist on something pointless
    b) to punish someone too harshly
    c) to give up quickly

6. *Don't beat around the bush and tell me frankly what you think of my proposition.*
The idiomatic expression "beat around the bush" means:
    a) to walk in circles
    b) to avoid getting to the point
    c) to speak too quickly

7. *A lot of hard work has been going on behind the scenes.*
The idiomatic expression "behind the scenes" means:
    a) at a theater
    b) in a private or hidden way
    c) after the event ends

8. *He bent over backwards trying to please his potential clients so that they would give him the contract.*
The idiomatic expression "bent over backwards" means:
    a) to try very hard to help someone
    b) to change one's opinion
    c) to stretch before working

9. *She was between the hammer and the anvil when her parents asked her to choose between going to college or getting married.*
The idiomatic expression "between the hammer and the anvil" means:
a) to face a difficult choice
b) to feel crushed by pressure
c) to be in a crowded place

10. *By accepting two part-time jobs, Tom is clearly biting off more than he can chew.*
The idiomatic expression "biting off more than he can chew" means:
a) to overeat during meals
b) to accept more responsibility than one can handle
c) to speak with food in one's mouth

11. *I told him he should stop building castles in the air and train for a sensible profession.*
The idiomatic expression "building castles in the air" means:
a) to construct tall buildings
b) to have unrealistic dreams or plans
c) to become a famous architect

12. *Joseph's been burning the candle at both ends for weeks, working two jobs during the week and a third on weekends.*
The idiomatic expression "burning the candle at both ends" means:
a) to stay home all day
b) to spend money carelessly
c) to overwork or exhaust oneself

13. *We managed to complete the project on time by the skin of our teeth.*
The idiomatic expression "by the skin of our teeth" means:
a) with great effort and just barely
b) with no effort at all
c) in a painful way

14. *Jane earned his money by the sweat of her brow.*
The idiomatic expression "by the sweat of her brow" means:
a) through hard physical or mental work
b) by stealing or cheating
c) in a cool and relaxed manner

15. *John accidentally spilled the beans about the new project during the meeting.*
The idiomatic expression 'spilled the beans' means:
a) to prepare a meal
b) to reveal a secret
c) to clean up a mess

https://doi.org/10.17323/jle.2025.19916

Intelligent Approaches to Computer Testing of Perception and Production Skills of Russian EFL Speakers

Marina Kolesnichenko [1], **Vitalii Kapitan** [2]

[1] Far Eastern Federal University, Vladivostok, Russia

[2] National University of Singapore, Singapore

ABSTRACT

Background: This study addresses a gap in applied phonetics by developing an evidence-based, neural network-driven computer phonetic test for Russian EFL learners. Integrating interdisciplinary methods, the system targets Russian-specific pronunciation deviations and delivers adaptive feedback, thereby advancing both perception and production skills while aligning technological innovation with pedagogical effectiveness.

Purpose: The purpose of the study is twofold: (1) to design and develop a computer-based system employing deep learning neural networks for objectively assessing Russian EFL students' perception and production skills, and (2) to evaluate the effectiveness and reliability of this system through repeated testing, statistical analysis of learner performance and user feedback.

Methods: A pre-test identified frequent segmental deviations, informing a targeted item pool. The software was developed in Microsoft Visual Studio 2022 (C#) using the Microsoft Speech Recognition Engine. The perception module used randomized audio stimuli (WAV files), while the speech recognition one recorded response via built-in microphones for automated accuracy evaluation. Twenty-five Russian EFL students (B1–B2 CEFR, aged 19–22) completed three test iterations at one-week intervals. Post-test questionnaire assessed usability and perceived learning gains. Data were analysed using descriptive statistics and correlation analysis.

Results: We designed a computer-based system employing deep learning neural networks and assessed its efficiency in Russian EFL learners. The study found a 14.5% overall improvement in participant performance, with results showing a clear linear increase supported by a high R^2 value. Students performed better in perception tasks than in production practice. Pearson correlation analysis indicated consistent performance between consecutive attempts, supporting robust test-retest reliability. Both modules showed high internal consistency ($\alpha = 0.90$ for perception, $\alpha = 0.88$ for production). Participants rated the tool as useful and interesting, although they suggested improving the speech recognition function due to minor technical flaws.

Conclusion: The module focused on testing perception skills can serve as an effective and engaging learning tool. While the pronunciation control component shows potential, its performance can be further enhanced through additional testing with high-sensitivity microphones to refine speech recognition accuracy. Overall, continued exploration of CAPT systems presents a promising direction for future research and innovation.

KEYWORDS

interdisciplinary approach, perception and production skills, speech recognition, neural networks, computer testing in phonetics, CAPT, Russian EFL speakers

Citation: Kolesnichenko, M., & Kapitan, V. (2025). Intelligent approaches to computer testing of perception and production skills of Russian EFL speakers. *Journal of Language and Education, 11*(2), 75-93. https://doi.org/10.17323/jle.2025.19916

Correspondence:
Vitalii Kapitan,
kapitanvy@gmail.com

Received: March 14, 2024
Accepted: June 10, 2025
Published: June 30, 2025

INTRODUCTION

Computational linguistics emerged in the mid-twentieth century as a response to the growing challenges in information technology, with a primary focus on enabling computers to process and understand natural language (Luz, 2022). This specialized field has facilitated technological advancements, including the development of speech recognition systems, linguistic corpora, and machine translation tools. By leveraging interdisciplinary collaboration between linguistics, computer science, and engineering, computational linguistics has laid the

foundation for innovative educational systems and improved data interaction across scientific and educational domains (Omid, 2022; Urip, 2022). These developments have set the stage for further exploration of language processing technologies and their applications in language learning.

Artificial neural networks have played a pivotal role in advancing computational linguistics by modelling complex relationships between input and output data, mirroring the functioning of biological neural networks in the human brain. These networks consist of interconnected layers-input, hidden, and output-that enable the system to perform tasks such as classification, clustering, and forecasting (Ivanko et al, 2019). The integration of neural networks into language technologies has resulted in the creation of automatic speech recognition (ASR) and speech-to-text (STT) systems, which are now widely used in both research and practical applications (Belenko & Balakshin, 2017; Mehrish et al., 2023; Dovchin, 2024). As a result, they have become integral to the ongoing evolution of computational linguistics and its role in educational innovation.

At the same time the study of speech production and perception has significantly contributed to the field of applied phonetics, providing valuable insights into the mechanisms underlying spoken language (Crystal, 1970; Flege & Davidian, 2008; Vishnevskaya, 2014; Munro & Derwing, 2020). Research in this area has provoked the development of pronunciation training tools and methodologies, with a particular emphasis on the variability and complexity of natural speech. These findings have been instrumental in shaping computer-assisted pronunciation training (CAPT) systems, which aim to enhance learners' pronunciation skills through targeted, technology-driven instruction (Fouz-González, 2020; Alsuhaibani et al., 2024). The intersection of applied phonetics and computational tools has thus expanded the possibilities for effective language learning interventions.

CAPT systems are reported to aid in improving foreign language pronunciation (Wang & Munro, 2004; González & Ferreiro, 2024; Rogerson-Revell, 2021). Many of them leverage the principle of High-Variability Phonetic Training (HVPT), exposing learners to a wide range of speech samples to improve perceptual accuracy. Several studies propose innovative developments in this domain (Barriuso & Hayes-Harb, 2018; Thomson & Derwing, 2015; O'Brien et al, 2018). Notable examples include the "English Accent Coach"[1], which uses gamification and multiple native speakers to increase learner motivation and effectiveness. Some researchers offer computer-based systems for pronunciation training in other languages save English. Blok (2019) developed a methodology for evaluating consonant pronunciation errors in German speech among Russian-speaking students. Similarly, Pashkovskaya (2010) created a flexible program that includes rhythmic-rhyming tasks aimed at improving the phonetics and intonation of Russian for students of various nationalities, with recommendations for pronunciation training in 17 languages. Other CAPT software integrate automatic speech recognition (ASR) and artificial intelligence to increase the potential for individualized learning outcomes (Rogerson-Revell, 2021).

Despite these technological advancements, the use of CAPT tools has often been criticized for placing greater emphasis on technological aspects, such as ASR and visual feedback, at the expense of sound pedagogical principles. This imbalance has created a gap between innovative technological tools and their educational value, with critics pointing to an over-reliance on repetitive drilling or mechanical exercises. Scholars (Pennington and Rogerson-Revel, 2019; Zou et al., 2024) have highlighted the need for a more thoughtful alignment between technology and teaching methodologies.

Additionally, many studies are noted for lacking robust designs, particularly the absence of control groups or delayed post-tests, which compromises the reliability and generalizability of the findings (Bliss et al., 2018; Agarwal, 2019). Moreover, the predominant focus on university-level learners limits the applicability of insights to other learning contexts, as noted by Mahdi and Al Khateeb (2019), and Rogerson-Revell (2021). Further criticism centers around the tendency of CAPT tools to adopt a one-size-fits-all approach. These systems often provide generalized feedback and learning materials, without sufficient customization to address individual learner needs. Derwing and Munro (2015) and Levis (2018) have consistently emphasized the necessity of more sophisticated and personalized feedback systems to enhance the effectiveness of pronunciation training.

Building on these insights, the present study aims to bridge the gap between technology and pedagogical effectiveness in computer-assisted pronunciation training for Russian EFL learners. By leveraging deep learning neural networks and incorporating principles from teaching methodologies, applied phonetics and psycholinguistics we seek to develop and empirically validate a computer-based system that addresses and evaluates Russian-specific pronunciation deviations.

Consequently, the aims of the present study were twofold: (1) to develop a computer-based system utilizing deep-learning neural networks to monitor Russian EFL students' perception and production skills; (2) to quantitatively evaluate the system's effectiveness through pre- and post-test comparisons, error rate analysis, and user feedback, ensuring accuracy, reliability, and practical value for EFL instruction.

[1] Thomson, R. I. (2017). English accent coach [Computer program]. Version 2.3. https://www.englishaccentcoach.com/

LITERATURE REVIEW

Linguistic, Psycholinguistic, and Cognitive Aspects in Modelling English Language Sound System for a Test

Phonetic and phonological competence is a critical component of communicative competence for EFL learners in non-English-speaking environments. This competence comprises internalized knowledge of the target language's sound system, perceptual and articulatory skills, and the ability to apply them effectively in communication (Goncharova, 2006). At each stage of knowledge formation, it is important to monitor acquired skills and analyze any potential deviations. Computer testing can be a useful tool in this process. However, designing such tests requires an understanding of skill formation processes, including cognitive, linguistic, psycholinguistic, and pedagogical aspects (Pennington & Rogerson-Revell, 2019; Flege & Bohn, 2021).

The motor theory of speech perception posits that phoneme acquisition relies on articulatory and acoustic cues rather than isolated auditory input (Stratton, 2025). Research supports this, demonstrating how resonant frequencies in the vocal tract, acoustic changes, and speech timing, influence speech quality (Leonov & Sorokin, 2007; Lam et al., 2012). Clear speech contributes more to identification accuracy revealing that information in the signal from the productions was crucial in facilitating word identification (Redmon et al., 2020; Sereno et al., 2025). Additionally, kinesthetic feedback plays a critical role in articulation control, allowing speakers to self-monitor and adjust their speech. For instance, for Russian EFL learners, contrasting native and English phonemes (e.g., apical-alveolar [s], [z] vs. dorsal-dental [c], [ɜ]) enhances awareness of phonetic differences. Effective foreign language acquisition thus relies not on imitation alone but on systematic analysis and comparison of speech signals (Pashkovskaya, 2010).

This systematic comparative practice should aid EFL learners in overcoming the side-effects of phonetic transfer. The latter is defined as the way where native-language sound system interferes with target-language perception and production (Mooney, 2019). To mitigate transfer effects, learners must develop a second phonological system through structured practice, reinforcing new auditory and articulatory patterns (Shevchenko, 2017). It is necessary to develop the movements for correct articulation of sounds (Stratton, 2025), using both auditory and motor analyzers through exercises that can restructure the speech functional system and develop new perceptual and articulatory images in the brain of a foreign language learner.

Phonological systems are hierarchically structured, with phonemes serving as mental prototypes for allophonic variations (Kulikov, 2005). Syllables, as the smallest functional units, provide perceptual cues for phonemic contrasts[2] (Bondarko, 1969). Cognitive linguistics extends this schema, demonstrating that speakers categorize sounds mentally, even without motor execution (Nesset, 2008). Modern cognitive phonology has shifted its focus from cataloging phonemic inventories to investigating the dynamic processes underlying phonological system formation and change (Ohala, 2013). The phonological system's schematic structure, rooted in structural and generative phonology, allows for the categorization and mental representation of phonological units such as phonemes, allophones, and syllables. These schematic relationships reinforce the cognitive processes involved in language learning and are enhanced through targeted exercises, gradually expanding the learner's phonetic and phonological repertoire, and enabling them to internalize L2 sound patterns algorithmically, a principle that also underpins computational speech recognition. The abstraction and algorithmization of these schematic relationships have facilitated the development of logical models for computer-based speech recognition systems, including those employing deep learning neural networks.

Deep Learning Neural Networks and Their Role in Speech Recognition

Linguistic analysis was founded on information theory techniques in the early days of computer technology when digital computers had just recently been introduced. These techniques fit in nicely with the philosophical and psychological ideas of the day. The application of deep neural networks has undergone a dramatic change in the last few decades (Mcshane & Nirenburg, 2021; Backus et al, 2023; Dovchin, 2024). With the use of this technology, performance on tasks involving natural language processing as well as numerous other speech, vision, and cognitive issues have improved dramatically. Tasks that were previously thought to be unsolvable can now be explored and solved due to advances in neural network technologies.

The concept of artificial intelligence and the era of formal language theory have arrived. This resulted in novel approaches for language processing and analysis while cognitive science was developing (Church & Liberman, 2021; Tikhonova & Raitskaya, 2023; Joshi et al, 2025). The application of deep learning neural networks, which have numerous layers of neurons and are based on principles of how the human brain functions, is one of the major ideas of modern times for the advancement of machine learning and artificial intelligence. Unlike previous methods that usually required significant manual feature engineering, these net-

[2] Potapova, R. K. (1986). Syllabic phonetics of Germanic languages. Vysshaya shkola.

works learn hierarchical representations directly from data. For speech processing tasks, convolutional neural networks (CNNs) (Fukushima, 1980, p. 396; LeCun, 1998, p. 2278) and recurrent neural networks (RNNs) (Mikolov et al., 2010, p. 1045; Graves et al., 2013; Su & Kuo, 2022) have proven effective for speech processing tasks Li et al, 2024; Rudregowda et al, 2024).

CNNs, originally developed for image recognition, excel at extracting spatial features from spectrograms, visual representations of audio frequencies over time. The CNN architecture uses convolutional layers, which apply filters to all inputs to detect patterns regardless of their position. These networks process spectrograms through a series of operations: convolutions to extract features, activation functions such as ReLU to introduce nonlinearity, and pooling to reduce dimensionality while preserving important information (Fig. A.1, see Appendix A for complete proofs). This approach allows CNNs to identify critical acoustic features such as formant transitions and phonetic boundaries that distinguish speech sounds.

RNNs address the sequential nature of speech by incorporating memory mechanisms that retain information across time steps. Unlike traditional feedforward networks, RNNs incorporate feedback connections that allow previous outputs to influence ongoing processing, creating an internal state that functions as a dynamic memory. This architecture is particularly suited to modeling temporal dependencies in speech, where the interpretation of current sounds depends on prior context (Fig. A.2, see Appendix A for complete proofs). Advanced variants such as long short-term memory (LSTM) networks overcome the limitations of basic RNNs by selectively retaining information in extended sequences, making them particularly valuable for recognizing related speech patterns.

The CNN and RNN combination create hybrid systems that use the strengths of both architectures. In these systems, CNNS first processes a spectrogram to extract reliable acoustic features, which are then served in RNN, which simulate temporary speech dynamics. This approach was surprisingly effective, since it solves both the problems of extraction and the consistent nature of speech recognition in a single structure. Modern speech recognition systems often realize this hybrid approach along with attention mechanisms that help the network focus on the corresponding parts of the input sequence, further increasing the accuracy of recognition in different linguistic contexts and performance conditions (Soundarya et al., 2023; Mehrish et al., 2023).

Deep neural networks have advanced automatic speech recognition, evolving with improvements in hardware and algorithms. Modern systems use convolutional and recurrent networks, making deep learning vital in computational linguistics to improve recognition accuracy.

Through our literature review, we found that researchers use various methods and tools to address participant needs in learning environments. However, further investigation is needed to identify suitable CAPT tools for classroom integration. A key challenge is that many CAPT systems are designed for specific research goals and may not apply well to diverse learning contexts. This is understandable as scientists have different research goals and focus on different aspects (Nickolai et al, 2024). However, any effort in developing and implementing CAPT tools in modern education is valuable and relevant. Having examined linguistic, psycholinguistic, and cognitive aspects of phonetic skills formation and intelligent approaches to solving speech processing tasks, we set a goal to create a tool using neural networks that could effectively control the progress of Russian EFL students in acquiring auditory and pronunciation skills in English[3,4]. We, then, tested its effectiveness in Russian EFL learning environment.

METHOD

The following sections provide a detailed overview of the participants, context and approaches used in the study.

Participants

The study involved twenty-five first-year students enrolled in the Bachelor's Degree programs in Applied Linguistics or English Philology, majoring in English as a foreign language (EFL). The participants comprised eighteen female and seven male students, aged between 18 and 19 years (B1–B2 CEFR). All participants had recently completed an Introductory Phonetics Course designed specifically for Russian EFL learners. This sample was selected to represent typical learners at the initial stage of formal phonetic training within the Russian higher education context.

[3] Korshunova, Y. S., Kapitan, V. Y. & Kolesnichenko, M. A. (2020a) Komp'uterniy test dly kontroly sluhoproiznositel'nix navikov [Computer test for monitoring students' auditory pronunciation skills]. (Certificate of State Registration of Computer Program No. RU 2020612357). The Federal Institute of Industrial Property, Rospatent. https://www1.fips.ru/fips_servl/fips_servlet?DB=EVM&DocNumber=2020612357

[4] Korshunova, Y. S., Kapitan, V. Y. & Kolesnichenko, M. A. (2020b). Razrabotka sistemi komp'uternogo testirovania dly kontroly sluhoproiznositel'nix navikov u studentov [Development of a computer testing system for monitoring students' hearing and speaking skills]. *Materials of the regional scientific-practical conference of students, graduate students and young scientists in natural sciences* (pp. 82-83). Far Eastern Federal University.

Context

Russian EFL students follow an intensive (36 hours) one-month Introductory Phonetics Course (IPC) that presents a part of a broader Practical Phonetics curriculum aimed at developing their phonetic competence. The course curriculum includes a variety of activities intended to activate and deepen students' phonetic knowledge. These activities encompass understanding interference phenomena between Russian and English phonological systems, comparative analysis of the two languages' sound structures, auditory discrimination of English speech sounds versus native Russian sounds, error identification and correction, self-recording of speech samples, etc. The primary objective of the IPC is to facilitate the formation of a secondary phonological system in English, enabling students to recognize and overcome common pronunciation difficulties. Special emphasis is placed on training auditory and motor analyzers to develop both perceptual and productive pronunciation skills essential for mastering English phonetics. To additionally monitor the development of these skills by the end of the IPC a computer phonetic test was created.

Task

To address our goals, as outlined in the Introduction, we structured the research process into four sequential tasks presented below.

Pre-Test Preparation and Administration

The initial phase involved the preparation and administration of a pre-test to identify typical phonetic challenges faced by Russian EFL learners. This phase was based on a comprehensive analysis of the typology of Russian and English phonetic systems (Arakin, 2008), phonetic transfer effects and differentiation between typical and fossilized articulation deviations in English - Russian language pairs (Vishnevskaya, 2014). The pre-test aimed to reveal and verify among the participants the most prevalent difficulties in English phonetic acquisition. For instance, vowel and consonant substitution, challenges with vowel length, devoicing of final consonants, and syllable division. A detailed account of the specific test items and their corresponding phonetic phenomena is provided in Appendix B.

The next task of the study was to process the results of the pre-test and develop a computer-based system for testing auditory and production skills, particularly focusing on the sound and syllabic structure of English vocabulary.

Development of a Computer-Based Auditory Skills Testing Module

To create this module, we used the Microsoft Visual Studio 2022 development environment and the C# programming language (Schildt, 2010). This enabled us to create a Windows application with a user-friendly interface, using the latest version of the .Net Framework 4.8 for optimal performance on Windows OS. The Windows Forms technology, which is part of the .NET Framework, provides a set of managed libraries that simplify the development and implementation of applications, ultimately ensuring a high level of usability for end-users, such as teachers and students[5]. The application was designed to read audio files and task texts from a secure directory and to allow for the dynamic expansion of tasks without requiring source code modification or recompilation. To ensure the authenticity and accuracy of the auditory stimuli, recordings were produced by a native American English speaker (a 32-year-old male English teacher) using a Sony ICD-TX650 digital voice recorder in a traditional laboratory setting. All audio files were classified and securely stored within the computer system.

The first module focuses on assessing phonemic hearing. The test tasks are designed to evaluate the ability to perceive the meaningful units of the English language and to perform phonemic actions, such as phonemic differentiation, determination of the distinctive function of a phoneme, segmentation of words into phonetic components (syllables), sound analysis of words, and selection of sounds in a specific order.

Development of a Computer-Based Speech Recognition Module

The third task of the study was to develop and implement a function for English speech recognition to control Russian EFL subjects' pronunciation. To solve that issue, we utilized intelligent neural network technologies, specifically speech recognition tools based on the Microsoft Speech Recognition Engine. This technology allows for real-time conversion of audio streams into text using the Speech Recognition Engine object, which enables the application to recognize words spoken into the microphone[6].

In total, two modules comprise 60 questions, divided into 6 task blocks. Thus, the main directory contains 6 subdirectories with different types of tasks, and each of them contains audio (except for tasks for the second module) and text documents in the *.ssv format.

[5] Windows Forms overview. (2023). https://learn.microsoft.com/en-us/dotnet/desktop/winforms/windows-forms-overview?view=netframeworkdesktop-4.8.

[6] Get Started with Speech Recognition (Microsoft.Speech) - Microsoft Speech Platform SDK 11 Documentation. https://documentation.help/Microsoft-Speech-Platform-SDK-11/4ca93e5c-65c9-433a-95c7-4343d8db269c.htm.

Data Collection, Analysis and Post-Evaluation

The final task was to analyze and evaluate the results of testing our tool to ensure its accuracy and effectiveness. The respondents had three attempts (with a week's interval between them) to work with the test. Suggesting that the subjects should have three attempts we targeted the following: (1) checking whether there might be progress in students' performance (overall performance improvement across attempts); (2) finding out drawbacks in the test semantic structure and evaluating its technical characteristics, (3) evaluating test reliability and validity.

Following the final test administration, a 10-item feedback questionnaire was developed, incorporating both structured (tabular) and open-ended response formats (see Fig. C.1, Fig. C.2 with sample questions in Appendix C for complete proofs). This instrument was designed to provide qualitative feedback and user experience on such aspects as: time allocation, accuracy of responses (correct vs incorrect answers), usability of the interface, clarity of task instructions, perceived difficulty of the test, usefulness, and overall technical performance of the testing system.

Procedure

Participant Briefing

The participants were given a clear explanation of the goals of the study. Prior to their involvement, they provided verbal consent, signifying their willingness to participate as subjects in a testing procedure.

Testing Environment, Protocol and Administration

Testing took place in a computer lab and was integrated as in-class activity regarding the respondents' busy academic schedules. Each subject had their own individual workstation equipped with an Acer computer and headphones. This setup allowed for independent completion of tasks. The students were given three attempts to pass the test, with a one-week interval between each attempt. This repeated-measures design allowed for the assessment of progress and the reliability of the testing instrument over time. Following the final testing, they were asked to answer a questionnaire (see Fig. C.1, Fig. C.2 with sample questions in Appendix C for complete proofs). The participants completed 60 tasks in the phonetic test: 45 tasks in the first module, and 15 tasks in the second one. The results of the completed test were saved as a pdf file where student's first name, last name, and group number were included in the file name. Each student was to enter their information and begin the first test section. Figure D.1 displays the start screen (see Fig. D.1 Appendix D for complete proofs).

Security and Anti-Cheating Measures

To ensure protection against cheating, all test files are stored in encrypted and hidden archives and directories. The files are sorted by directories with the task number, and the audio files are converted into WAV format. A text document in the *.csv format is created for each directory with audio files. This document serves as an additional protection against cheating, as the CSV format is not recognized by default in Windows OS. It contains a formulated task, the number of audio recordings, and answer options. The order of displaying answer options is randomized in the program. The first part of the computer test begins after the program starts and the necessary data is filled in. It displays the tasks of the first block of questions, where subjects are asked to choose one correct answer from three options.

Test Interface and Task Flow

Here, we provide some tasks for illustration: *Click on the "Play" button and choose the word in which you will hear a short sound.* (see Fig. D.2 Appendix D for complete proofs).

The words are presented in audio form and are not visible to the respondents. Three versions of the audio recording are given, for example, with words such as: *do, two, good*. Each answer option has a "Play" button. After listening to each option, the listener must choose the correct answer, click the "Choice" button, and move on to the next question ("Next"). Test takers also have the option to skip the listening part by clicking on the "Skip listening part" button and proceed to the second part of the test, which checks pronunciation.

Here is one more example: *Click on the "Play" button to hear the word. Listen carefully, determine which sound you hear [w] or [v].* Audio is offered, and there are two possible answers. It is necessary to determine which sound is pronounced [w] or [v], (see Fig. D.3 Appendix D for complete proofs).

After completing the auditory section, the program automatically calculated the number of correct responses and prompted participants to proceed to the pronunciation assessment. In this section, participants were shown a word on the screen, given two seconds to prepare, and then instructed to pronounce the word clearly upon receiving a visual cue. The speech recognition module then displayed the recognized text, allowing for immediate feedback on pronunciation accuracy.

The task is formulated as follows: *Clicking on the "Start" button you will see the WORD for reading. You will have 2 seconds to read that word. After that, you will see the command «Speak». Please, pronounce the WORD distinctly. For the next word, please click on the «Next» button*.

This block contains 15 words to check pronunciation (see Fig. D.4 Appendix D for complete proofs). After 15 tasks are completed, there appears a window on the screen with the calculated number of correct answers.

Data Storage

Following completion of all test sections, participants were asked to complete a questionnaire developed by the authors. All responses and results were securely saved in PDF format. This systematic approach facilitated efficient data collection and subsequent analysis.

RESULTS

Overall Performance Improvement Across Attempts

We developed and tested a deep-learning-based software to monitor Russian EFL students' perception and production skills in an Introductory Phonetics course. Feedback was collected through a post-test questionnaire, then analyzed and visualized using MS Excel and Python. As shown in Fig. 1 and 2, students' performance improved by 14.5% between the first and third attempts, demonstrating the system's effectiveness.

To analyze the trend, the scores of students presented above were averaged over multiple attempts. R^2 analysis involves calculating the coefficient of determination, a statistical measure that assesses how well the regression line fits the data by quantifying the proportion of variance in the dependent variable explained by the independent variable. By performing the R^2 analysis on the average scores over multiple attempts, we assess how well the increase in scores can be explained by the number of attempts or the implementation of the proposed test. The results show a clear linear increase, which was further supported by a high confidence R^2 value (R^2 value measures the trendline: the closer R^2 is to 1, the better the trendline fits the data.). This upward trend suggests that the implementation of the proposed test has significantly improved the subjects' performance.

A positive correlation between attempts as well as clear linear trend and high R^2 value confirm that our solution is an effective pedagogical tool for enhancing Russian EFL students' phonological competence through systematic practice.

Next, we considered the two parts of the test separately.

Perception Skills Module: Performance Trends

When analyzing the first part of the test, which assessed auditory perception skills, students demonstrated consistently positive results across all three attempts. The data exhibited a clear upward linear trend, as depicted in Fig. 3 and 4.

Pronunciation Skills Module: Performance Trends

The second part of the test, focused on pronunciation and speech recognition, yielded heterogeneous results. Unlike the perception section, participants' performance did not follow a linear trend. Instead, polynomial approximation was necessary to model the data accurately (see Fig. 5 and 6).

Figure 1
Comparison of Students' Total Performance over Three Attempts

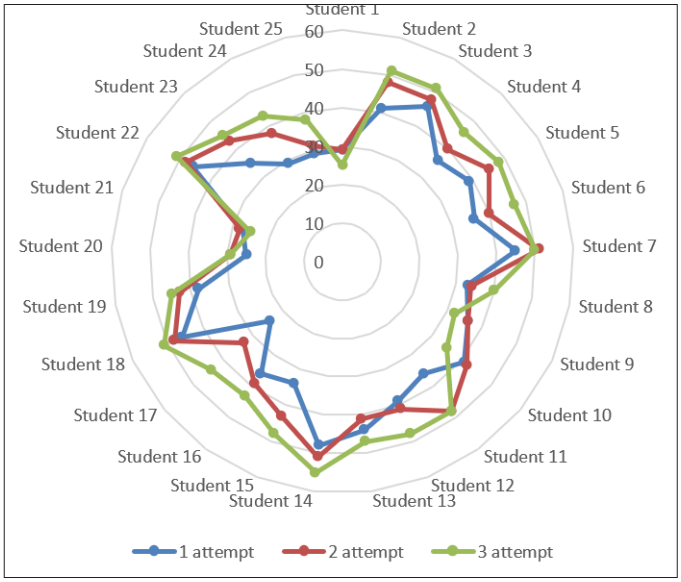


Figure 2
Trendline of the Average Score over Three Attempts

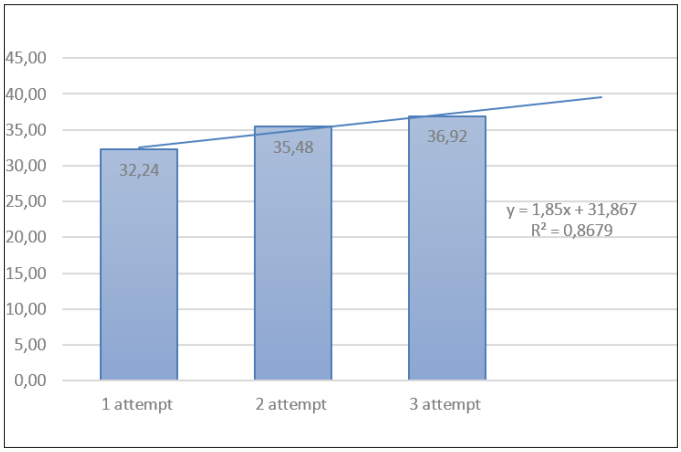


Figure 3

Comparison of Students' Performance in the Perception Part of the Test over Three Attempts

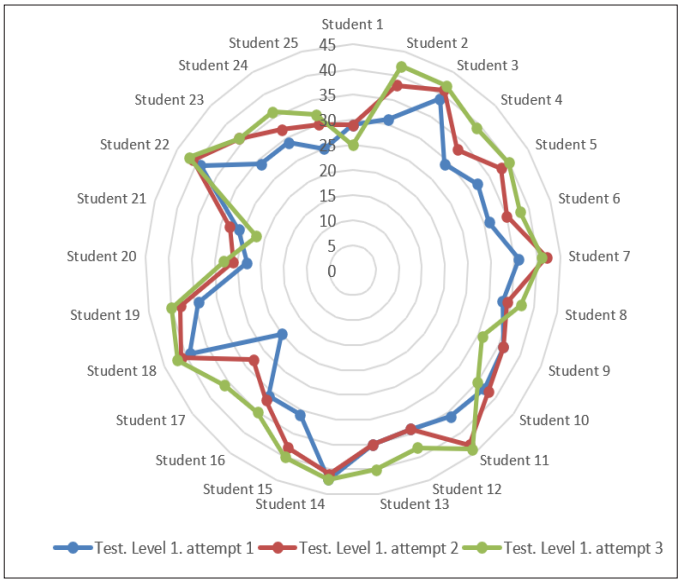


Figure 4

Trendline of the Average Score over Three Attempts after Passing the Perception Part of the Test

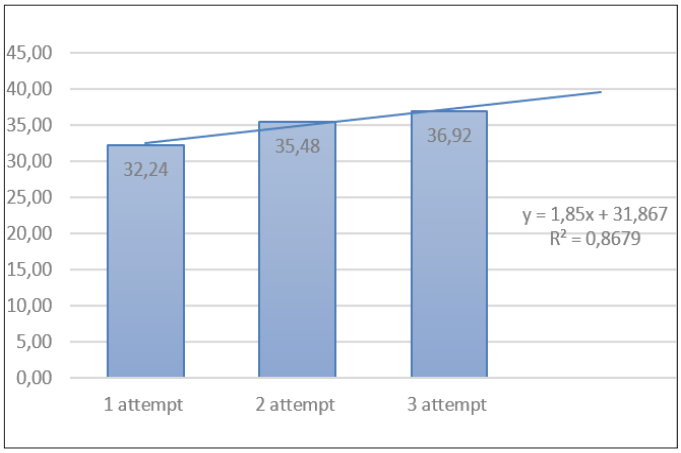


Figure 5

Comparison of Students' Performance in the Speech Recognition Part of the Test over Three Attempts

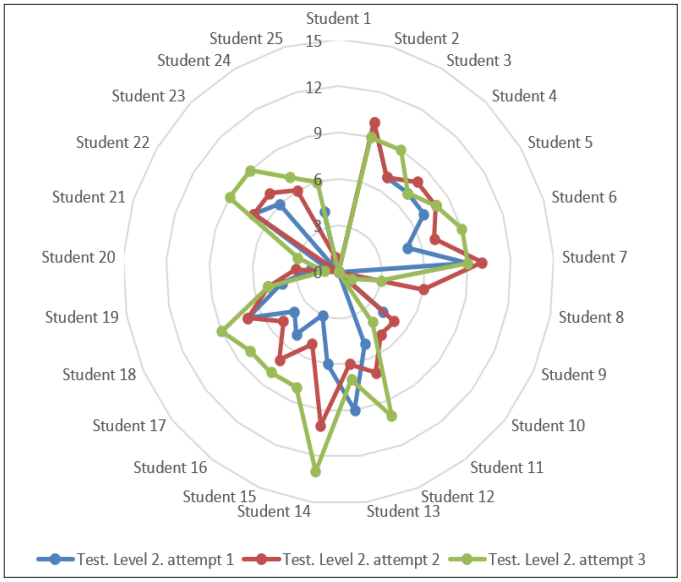
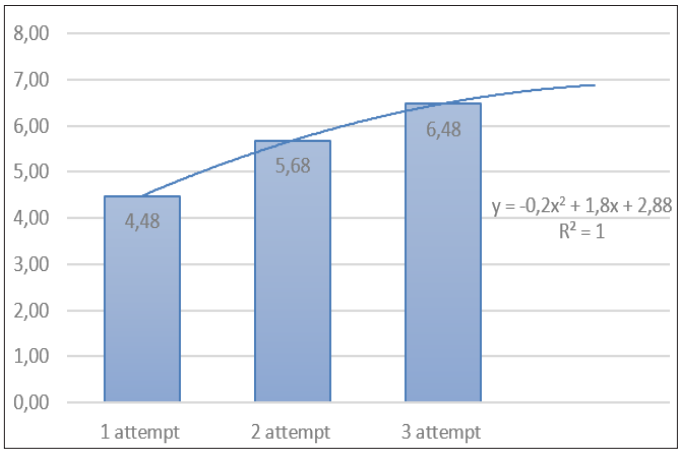


Figure 6

Trendline of the Average Score over Three Attempts after Passing the Speech Recognition Part of the Test



The analysis shows a tendency towards a smooth increase, which indicates a sufficient level of completion of the tasks of the second part of the test, as well as the gradual mastery of the material by students.

Test Completion Time Analysis

Analysis of the time required to complete the test across attempts revealed a non-linear pattern. While initial observations suggested a reduction in completion time, polynomial trend analysis (Fig. 7 and 8) indicated a gradual increase in time spent on subsequent attempts. Questionnaire responses clarified that increased time was often due to students' desire to double-check answers or further engage with challenging tasks.

Qualitative Feedback and User Experience

Post-test questionnaire responses provided valuable insights into user experience. In particular, the subjects' com-

Figure 7

Comparison of Time Spent on the Test across Three Attempts

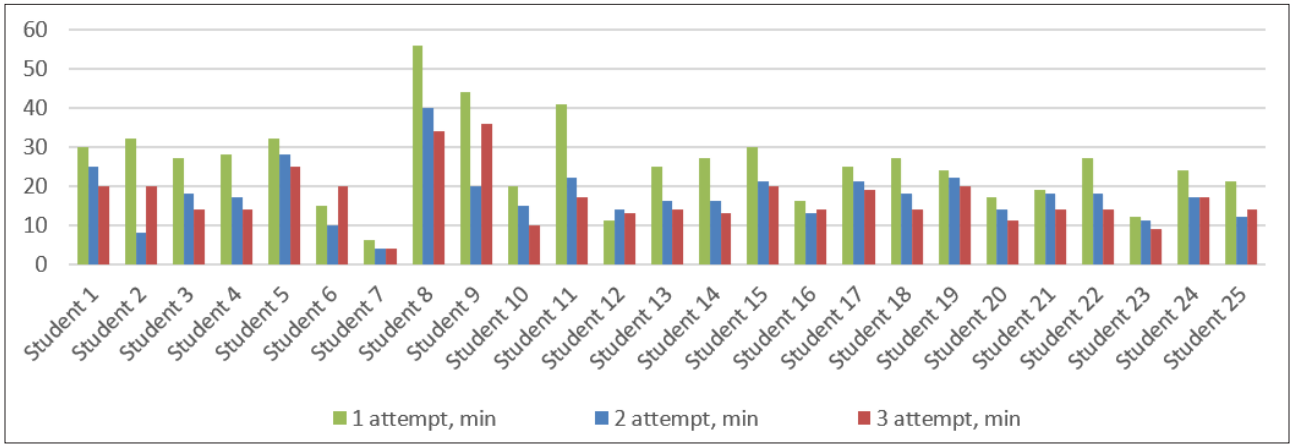
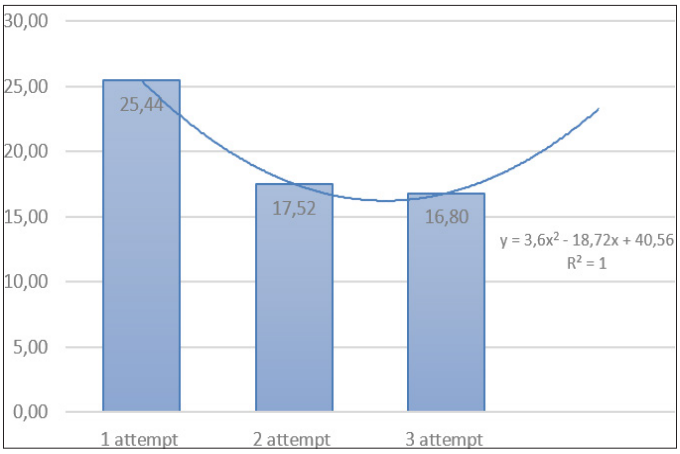


Figure 8

Trendline of Time Variation between Three Attempts



ments made it possible to understand why the test time increased during subsequent attempts. In some forms there were such answers as "*interesting, I went through several times because I had doubts about certain questions, decided to double-check myself*", "*useful, encouraged me to use a dictionary*". From a technical point of view, the first module of the Test has already been implemented in a pre-release form, which is why students found working with it more intuitive and useful, for example, we received the following comments: "*helps you assess your knowledge, test yourself*", "*required me to think, but it's interesting*".

The second set of tasks with speech recognition technology proved to be technically challenging for the subjects. They noted that the accuracy of pronunciation is affected by the sensitivity of the microphones. Some pointed out that the words appeared on the screen inconsistently, for example, some subjects wrote "*not all words were recognized accurately* ", "*it was unclear whether I should repeat the word again, but a new word was already appearing for reading*".

Other helpful feedback was the following: *"I would appreciate the option to go back to a previous question, I have to start over", "the system does not give comments, only a report on the number of correct answers".*

It should be noted that this is the first version of the application, designed to demonstrate its core functionality. Despite some technical issues in the second module of the test, the participants stated the absolute usefulness of this method of control, interest and increased motivation to achieve personal results in pronunciation.

Evaluation of Reliability and Validity of the Developed Test

We also assessed reliability and validity, metrics that are critical in social science research, to ensure that the measurement instruments developed capture the constructs the test is intended to measure with sufficient accuracy (Drost, 2011).

Test-Retest Reliability

Test-Retest Reliability is a method used to assess the consistency or stability of a measurement over time. It involves administering the same test to the same group of individuals several times. One way to calculate this is to calculate the Pearson correlation between scores from different trials (Attempt 1 vs. Attempt 2, Attempt 2 vs. Attempt 3 and Attempt 3 vs. Attempt 1 for Part 1 and Part 2 of the test). A higher Pearson correlation indicates greater test-retest reliability, meaning that the test produces consistent results when administered to the same individuals over time. In this analysis, correlation coefficients between 0.700 and 0.890 are considered to reflect strong reliability, while values between 0.500 and 0.690 indicate moderate reliability. The analysis of Pearson correlation coefficients revealed strong test-retest reliability across multiple assessment attempts. In Part 1 strong positive correlations were observed between Attempt 1 and Attempt 2 (0.860) and between Attempt 2 and Attempt 3 (0.811). Students who performed well in the first attempt also tended to perform well in the second attempt, and after that third one, suggesting reliability in their performance across these trials. A moderate correlation was found between Attempt 3 and Attempt 1 (0.597). Students' performance on the third attempt is not as strongly related to their performance on the first attempt. This drop in correlation suggests that some results changed significantly in students' performance between the first and third attempts, reflecting less consistency over time. Similarly, Part 2 demonstrated strong positive correlations between Attempt 1 and Attempt 2 (0.727) and between Attempt 2 and Attempt 3 (0.767), with a moderate to strong correlation between Attempt 3 and Attempt 1 (0.679). These findings indicate consistent student performance between consecutive attempts, with some expected variability over longer intervals. Overall, the assessment demonstrates robust test-retest reliability.

Internal Consistency (Cronbach's Alpha)

Cronbach's Alpha is a measure of internal consistency or reliability of a set survey questions that are supposed to measure the same construct. A higher Cronbach's Alpha indicates better reliability of the items. The 0.90 value of the Cronbach's Alpha for Part 1 falls in the "excellent" internal consistency range, while the 0.88 value for Part 2 is at the upper end of the "good" range. These high values suggest that the items within each part are measuring the same underlying construct consistently, indicating strong reliability of the assessment instrument. In social science research, these alpha values exceed generally accepted thresholds of sufficiency, representing very good indicators of reliability that satisfy methodological standards in this field.

Analysis of Construct and Criterion-Related Validity

Test validity is the extent to which a test accurately measures what it is supposed to measure. In the context of educational assessments in the social sciences, validity ensures that the test accurately reflects students' understanding of the subject matter. This analysis focuses solely on assessing the validity of a test based on student scores across three attempts.

Construct validity assesses whether the test measures the theoretical construct it is intended to measure. Let us examine the correlations between the attempts given above to prove construct validity. The strong positive correlations between Attempt 1 vs. Attempt 2 and between Attempt 2 vs. Attempt 3 suggest consistency in what the test measures across these attempts. This consistency supports the construct validity of the test and indicates that the test consistently measures the intended construct over time.

Criterion-related validity examines how well one measure predicts an outcome based on another measure, i.e. earlier attempts can predict performance on later attempts. Specifically, the strong positive correlation between Attempt 1 and Attempt 2 scores (0.860 and 0.727) indicates that students' performance on the initial attempt is a good predictor of their performance on the subsequent attempt. Similarly, the strong correlation between Attempt 2 and Attempt 3 scores (0.811 and 0.767) reinforces this predictive relationship between consecutive assessments. However, the weaker correlation between Attempt 1 and Attempt 3 scores (0.597 and 0.679) suggests that the test's ability to predict performance diminishes over non-consecutive attempts. This pattern highlights that while the test effectively estimates immediate future performance (supporting criterion-related validity in the short term), its predictive power over longer intervals is less pronounced.

DISCUSSION

The present study contributes to the expanding field of computer-assisted pronunciation training (CAPT), particularly as it pertains to the persistent challenges encountered by Russian EFL learners. Within the broader context of second language (L2) phonological acquisition and educational technology, our findings reinforce the growing consensus that targeted, technology-mediated interventions can meaningfully enhance learners' perceptual and productive pronunciation skills (González & Ferreiro, 2024; Alsuhaibani et al., 2024). By integrating established EFL phonetic instruction principles (Wang et al., 2025; Wei, 2025) into a CAPT system tailored for Russian university students, our research bridges a notable gap between theoretical frameworks and their practical application in language learning environments.

Our investigation employed a mixed-methods design, combining a pre-test for contrastive phonetic analysis, a computer-based assessment tool with perception and production modules, three iterative test administrations at weekly intervals, and a post-test questionnaire to capture user

experience. This methodological approach enabled both cross-sectional and longitudinal analysis of pronunciation development among 25 first-year Russian EFL students, providing a sound (Nickolai et al, 2024) foundation for interpreting our results.

The principal findings of this study show that overall participants' performance across attempts improved by 14.5% with the results showing a clear linear increase, which was further supported by a high confidence R2 value. Overall, students performed well on the tasks given, demonstrating confident results. Both the perception and production modules offered valuable insights into learner performance. Although some technical improvements were suggested for the production module, the system as a whole was rated as useful and efficient. The analysis of Pearson correlation coefficients indicates consistent student performance between consecutive attempts, with some expected variability over longer intervals. In total, the assessment demonstrates robust test-retest reliability. Both perception and production modules exhibited statistically significant internal consistency ($\alpha = 0.90$ and $\alpha = 0.88$ correspondently) confirming its reliability as a monitoring tool. Analysis of construct and criterion-related validity highlights that while the test effectively estimates immediate future performance (supporting criterion-related validity in the short term), its predictive power over longer intervals is less noticeable.

Interpreting these results, the marked improvement in perceptual discrimination supports the theoretical position that explicit training in articulation can facilitate perceptual gains (Pashkovskaya, 2010; Flege et al., 2021; Stratton, 2025). This finding aligns with the view that perception and production are mutually reinforcing processes in L2 phonological development, especially when training is tailored to learners' specific L1-L2 transfer zones. Our pre-test analysis and subsequent learner-oriented task design (Marefat et al., 2025) based on a thorough analysis of transfer zones and typical phonetic deviations were instrumental in targeting these areas, thereby enhancing the relevance and impact of the intervention.

Comparatively, the limited accuracy of the speech recognition system is consistent with recent literature highlighting the technical and contextual barriers to automated pronunciation evaluation (Souza & Gottardi, 2022; Shadiev, 2023; Nickolai et al, 2024). While some studies have used CAPT solutions based on more complex web-based AI modules (e.g., Dovchin, 2024), the performance of our system's production module was influenced by several contextual factors, such as variability in microphone quality in typical educational environments, limitations in acoustic modeling for non-native accents, and the absence of contextualized speech input. These challenges reflect the inherent complexity of reliably assessing L2 speech production and highlight the need for further development - potentially through the integration of higher-quality audio equipment or more advanced recognition algorithms.

The outcomes of this study were generally anticipated, given the theoretical and empirical foundations underpinning our intervention design. However, the magnitude of improvement in perceptual discrimination exceeded initial expectations, suggesting that even short-term, focused training can yield measurable gains. Conversely, the persistent challenges in automated production assessment highlight the need for continued innovation in CAPT technologies and methodologies.

Several limitations must be acknowledged. The study's sample was limited to first-year Russian university students, which may constrain the generalizability of findings to other learner populations or educational settings. Additionally, the test's focus on Russian university EFL learners, whose instructional goal is often native-like pronunciation, may limit its applicability in contexts with differing learner objectives (Hino, 2021). The production module's technical limitations, omission of capturing suprasegmental features and spontaneous speech, also limit the scope of our conclusions. Furthermore, the absence of a control group and the relatively short intervention period may limit the strength of causal inferences.

This research advances our understanding of CAPT's potential and limitations in Russian EFL contexts. The demonstrated efficacy of the auditory perception module provides a promising avenue for future development, while the challenges encountered in automated production assessment point to the need for further technological and pedagogical refinement. Expanding the system to address suprasegmental features and to accommodate a broader range of learner profiles represents a logical next step in optimizing technology-mediated pronunciation instruction.

CONCLUSION

This investigation has yielded some insights into the efficacy of computer-assisted phonetic assessment for Russian EFL learners. The study's principal findings demonstrate that technology-mediated training produces measurable improvements in L2 phonological perception, corroborating established psycholinguistic models and pronunciation teaching principles of EFL speech learning. The auditory discrimination module emerged as particularly effective, suggesting that structured perceptual training forms a foundation for phonological competence development.

The contribution of this research is threefold. First, it provides empirical validation for computer-based approaches

to monitoring phonetic challenges specific to Russian-English interlanguage. Second, it establishes a methodological framework for developing targeted pronunciation training protocols. Third, it identifies key technical limitations in current automated speech recognition applications for pedagogical contexts, regarding accented speech evaluation.

From a pedagogical perspective, these findings underscore the value of integrating diagnostic assessment tools within pronunciation curricula. The demonstrated effectiveness of iterative perceptual training suggests promising applications for autonomous learning environments. However, the technical constraints observed in production evaluation highlight the need for more sophisticated acoustic modeling approaches in CAPT systems.

Future investigations should prioritize: (1) longitudinal studies tracking the retention of training effects, (2) expansion of assessment parameters to encompass suprasegmental features, and (3) development of adaptive algorithms capable of processing non-native phonological variation. Such advancements would substantially enhance the validity and pedagogical utility of computer-assisted pronunciation training systems.

ACKNOWLEDGMENTS

The authors would like to express their gratitude to Y. S. Korshunova and A. O. Korol for their assistance in developing the computer test and preparing the materials for this article.

FUNDING

V. Kapitan conducted the study with financial support from the "MAPLE" project, grant number 22-5715-P0001, under the National University of Singapore Faculty of Science, Ministry of Education, Tier 1 grant “Data for Science and Science for Data collaborative scheme”

DECLARATION OF COMPETING INTEREST

None declared.

AUTHORS’ CONTRIBUTION

Marina Kolesnichenko: conceptualization; data curation; formal analysis; methodology; project administration; supervision; visualization; writing – original draft; writing – review & editing.

Vitalii Kapitan: software; data curation; investigation; methodology; visualization; resources; funding acquisition; writing – original draft; writing – review & editing.

REFERENCES

Agarwal, C., & Chakraborty, P. (2019). A review of tools and techniques for computer aided pronunciation training (CAPT) in English. *Education and Information Technologies, 24*(6), 3731-3743. https://doi.org/10.1007/s10639-019-09955-7

Alsuhaibani, Y., Mahdi, H. S., Al Khateeb, A., Al Fadda, H. A., & Alkadi, H. (2024). Web-based pronunciation training and learning consonant clusters among EFL learners. *Acta Psychologica*, *249*, 104459. https://doi.org/10.1016/j.actpsy.2024.104459

Arakin, V. D. (2008). Comparative typology of English and Russian languages (4rd ed.). Fizmatlit.

Backus, A., Cohen, M., Cohn, N., Faber, M., Krahmer, E., Laparle, S., Maier, E., Van Miltenburg, E., Roelofsen, F., Sciubba, E., Scholman, M., Shterionov, D., Sie, M., Tomas, F., Vanmassenhove, E., Venhuizen, N., & De Vos, C. (2023). Minds: Big questions for linguistics in the age of AI. *Linguistics in the Netherlands*, *40*, 301–308. https://doi.org/10.1075/avt.00094.bac

Barriuso, T. A., & Hayes-Harb, R. (2018). High variability phonetic training as a bridge from research to practice. *The CATESOL Journal*, *30*(1), 177-194. https://doi.org/10.5070/B5.35970

Belenko, M. V., & Balakshin, P. V. (2017). Comparative analysis of speech recognition systems with open code. *International Research Journal*, 4(58), 13-18. https://doi.org/10.23670/IRJ.2017.58.141

Bliss, H., Abel, J. & Gick, B. (2018). Computer-assisted visual articulation feedback in L2 pronunciation instruction: A review. *Journal of Second Language Pronunciation, 4*, 129-153. https://doi.org/10.1075/jslp.00006.bli

Blok, E. (2019). The planning and customization of introductory L2 phonetic courses on the basis of a numeric scale for assessing non-native speaker mistakes. *Rhema*, (4), 34-52. https://doi.org/10.31862/2500-2953-2019-4-34-52.

Bondarko, L. V. (1969). *Slogovaya struktura rechi i differentsial’nye priznaki fonem (eksperimental’no-foneticheskoe issledovanie na materiale russkogo yazyka)* [The syllabic structure of speech and the distinctive features of phonemes (experimental-phonetic research in the Russian language)] [Unpublished doctoral dissertation]. Leningrad State University after A. A. Zhdanov. https://search.rsl.ru/ru/record/01010234177

Church, K., & Liberman, M. (2021). The future of computational linguistics: On beyond alchemy. *Frontiers in Artificial Intelligence*, *4,* 625341. https://doi.org/10.3389/frai.2021.625341

Crystal, D. (1970). Prosodic systems and language acquisition. *Prosodic Feature Analysis* (pp. 77-90). Didier Montreal and Paris.

Derwing, T. M., Munro, M. J. (2015) *Pronunciation fundamentals: Evidence-Based perspectives for L2 teaching and research*. John Benjamins.

Dovchin, S. (2024). Artificial Intelligence in Applied Linguistics: A double-edged sword. *Australian Review of Applied Linguistics*, *47*(3), 410–417. https://doi.org/10.1075/aral.24145.dov.

Drost, E. A. (2011). Validity and reliability in social science research. *Education Research and Perspectives*, *38*(1), 105-123. https://search.informit.org/doi/10.3316/.

Flege, J. E., & Bohn, O.-S. (2021). The Revised Speech Learning Model (SLM-r). In R. Wayland (Ed.), *Second Language Speech Learning* (1st ed., pp. 3–83). Cambridge University Press. https://doi.org/10.1017/9781108886901.002.

Flege, J. E., & Davidian, R. D. (2008). Transfer and developmental processes in adult foreign language speech production. *Applied Psycholinguistics*, *5*(4), 323-347. https://doi.org/10.1017/S014271640000521X.

Fouz-González, J. (2020). Using apps for pronunciation training: An empirical evaluation of the English File Pronunciation App. *Language Learning & Technology*, *24*(1), 62–85. https://doi.org/10125/44709

Fukushima, K. (1980). Neocognitron: A self-organizing neural network model for a mechanism of pattern recognition unaffected by shift in position. *Biological Cybernetics*, *36*(4), 193-202. https://doi.org/10.1007/BF00344251

Goncharova, N. L. (2006). *Formirovaniye inoyazychnoy fonetiko-fonologicheskoy kompetentsii u studentov-lingvistov: na materiale angliyskogo yazyka* [Forming foreign language phonetic-phonological competence of linguistic students based on the material of the English language] [Unpublished doctoral dissertation]. North Caucasus State Technical University. https://search.rsl.ru/ru/record/01003042566

González, M. D. L. Á. G., & Ferreiro, A. L. (2024). Web-assisted instruction for teaching and learning EFL phonetics to Spanish learners: Effectiveness, perceptions and challenges. *Computers and Education Open*, *7*, 100214. https://doi.org/10.1016/j.caeo.2024.100214

Graves, A., Mohamed, A. R., & Hinton, G. (2013, May). Speech recognition with deep recurrent neural networks. In *2013 IEEE international conference on acoustics, speech and signal processing* (pp. 6645-6649). IEEE. https://doi.org/10.1109/ICASSP.2013.6638947

Hino, N. (2021). Language education from a post-native-speakerist perspective: The case of English as an international language. *Russian Journal of Linguistics*, *25*(2), 528-545. https://doi.org/10.22363/2687-0088-2021-25-2-528-545

Ivanko, A. F., Ivanko, M. A., Sizova, Y. A. (2019). Neural networks: General technological characteristics. *Scientific Review. Technical Sciences*, (2), 17-23.

Joshi, A., Dabre, R., Kanojia, D., Li, Z., Zhan, H., Haffari, G., & Dippold, D. (2025). Natural Language Processing for dialects of a language: A survey. *ACM Computing Surveys*, *57*(6), 1–37. https://doi.org/10.1145/3712060.

Kulikov, V. G. (2005). Phonological contexts and frames: Toward the unified methodology of cognitive linguistics. *Issues of Cognitive Linguistics*, (2), 28-40.

Lam, J., Tjaden, K., & Wilding, G. (2012). Acoustics of Clear Speech: Effect of Instruction. *Journal of Speech, Language, and Hearing Research*, *55*(6), 1807–1821. https://doi.org/10.1044/1092-4388(2012/11-0154).

LeCun, Y., Bottou, L., Bengio, Y., & Haffner, P. (1998). Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, *86*(11), 2278-2324. https://doi.org/10.1109/5.726791

Leonov, A. S., & Sorokin, V. N. (2007). K analizu rezonansnykh chastot rechevogo trakta [To the analysis of the resonant frequencies of the speech tract]. *Informacionnye Process, 4*(7), 386 - 400.

Levis, J. M. (2018). *Intelligibility, oral communication, and the teaching of pronunciation*. Cambridge University Press.

Li, Z., Basit, A., Daraz, A., & Jan, A. (2024). Deep causal speech enhancement and recognition using efficient long-short term memory Recurrent Neural Network. *PLOS ONE*, *19*(1), e0291240. https://doi.org/10.1371/journal.pone.0291240

Luz, S. (2022). Computational linguistics and natural language processing. *The Routledge handbook of translation and methodology* (pp. 373-391). Routledge. https://doi.org/10.4324/9781315158945-27

Mahdi, H. S., & Al Khateeb, A. A. (2019). The effectiveness of computer-assisted pronunciation training: A meta-analysis. *Review of Education*, 7(3), 733-753. https://doi.org/10.1002/rev3.3165

McShane, M., & Nirenburg, S. (2021). *Linguistics for the Age of AI*. Mit Press. https://doi.org/10.7551/mitpress/13618.003.0003

Marefat, F., Hassanzadeh, M., Noureddini, S., & Ranjbar, M. (2025). Reporting practices in applied linguistics quantitative research articles across a decade: A methodological synthesis. *System*, *131*, 103627. https://doi.org/10.1016/j.system.2025.103627

Mehrish, A., Majumder, N., Bharadwaj, R., Mihalcea, R., & Poria, S. (2023). A review of deep learning techniques for speech processing. *Information Fusion*, *99*, 101869. https://doi.org/10.1016/j.inffus.2023.101869

Mikolov, T., Karafiát, M., Burget, L., Cernocký, J., & Khudanpur, S. (2010). Recurrent neural network-based language model. *Interspeech*, *2*(3), 1045-1048. https://doi.org/10.21437/Interspeech.2010-343

Mooney, D. (2019). Phonetic transfer in language contact: Evidence for equivalence classification in the mid-vowels of Occitan–French bilinguals. *Journal of the International Phonetic Association*, *49*(1), 53-85. https://doi.org/10.1017/S0025100317000366

Munro, M. J., & Derwing, T. M. (2020). Foreign accent, comprehensibility and intelligibility, redux. *Journal of Second Language Pronunciation*, *6*(3), 283-309. https://doi.org/10.1075/jslp.20038.mun

Nesset, T. (2008). Ronald W. Langacker, Cognitive Grammar: A basic introduction. Oxford: Oxford University Press, 2008. Pp. x+562. *Journal of Linguistics*, *45*(2), 477–480. https://doi.org/10.1017/S0022226709005799.

Nickolai, D., Schaefer, E., & Figueroa, P. (2024). Aggregating the evidence of automatic speech recognition research claims in CALL. *System*, *121*, 103250. https://doi.org/10.1016/j.system.2024.103250

O'Brien, M. G., Derwing, T. M., Cucchiarini, C., Hardison, D. M., Mixdorff, H., Thomson, R. I., Strik, H., Levis, J. M., Munro, M. J., Foote J. A. & Levis, G. M. (2018). Directions for the future of technology in pronunciation research and teaching. *Journal of Second Language Pronunciation*, *4*(2), 182-207. https://doi.org/10.1075/jslp.17001.obr

Ohala, J. J. (2010). The relation between phonetics and phonology. In *The Handbook of Phonetic Sciences* (pp. 653–677). Wiley Blackwell. https://doi.org/10.1002/9781444317251.ch17

Omid, M. (2022). Review of research on the use of Information and Communication Technologies (ICTs) in ELT-related academic writing classrooms. *Journal of Language and Education*, *8*(2), 165-178. https://doi.org/10.17323/jle.2022.13395

Pashkovskaya, S. S. (2010). *Differenciruyushaya model obucheniya russkomu proiznosheniyu* [Differentiating model of teaching Russian pronunciation] [Unpublished doctor dissertation]. The Pushkin State Russian Language Institute. https://search.rsl.ru/ru/record/01004949907

Pennington, M. C., & Rogerson-Revell, P. (2019). *English pronunciation teaching and research: Contemporary perspectives*. Palgrave Macmillan. https://doi.org/10.1057/978-1-137-47677-7

Redmon, C., Leung, K., Wang, Y., McMurray, B., Jongman, A., & Sereno, J. A. (2020). Cross-linguistic perception of clearly spoken English tense and lax vowels based on auditory, visual, and auditory-visual information. *Journal of Phonetics*, *81*, 100980. https://doi.org/10.1016/j.wocn.2020.100980

Rogerson-Revell, P. M. (2021). Computer-assisted pronunciation training (CAPT): Current issues and future directions. *RELC Journal*, 52(1), 189-205. https://doi.org/10.1177/0033688220977406

Rudregowda, S., Patilkulkarni, S., Ravi, V., H.L., G., & Krichen, M. (2024). Audiovisual speech recognition based on a deep convolutional neural network. *Data Science and Management*, *7*(1), 25–34. https://doi.org/10.1016/j.dsm.2023.10.002

Schildt, H. (2010). *C# 4.0: The complete reference*. McGraw-Hill.

Sereno, J. A., Jongman, A., Wang, Y., Tupper, P., Behne, D. M., Gu, J., & Ruan, H. (2025). Expectation of speech style improves audio-visual perception of English vowels. *Speech Communication*, *171*, 103243. https://doi.org/10.1016/j.specom.2025.103243

Shadiev, R., & Liu, J. (2023). Review of research on applications of speech recognition technology to assist language learning. *ReCALL*, *35*(1), 74–88. https://doi.org/10.1017/S095834402200012X

Shevchenko, T. I. (2017). Cognitive phonology: Theoretical and applied aspects. *Vestnik of Moscow State Linguistic University. Humanities*, *5*(776), 106-115.

Soundarya, M., Karthikeyan, P. R., & Thangarasu, G. (2023). Automatic speech recognition trained with convolutional neural network and predicted with recurrent neural network. In *2023 9th International Conference on Electrical Energy Systems,* (pp. 41-45). IEEE. https://doi.org/10.1109/ICEES57979.2023.10110224

Souza, H. K. D., & Gottardi, W. (2022). How well can ASR technology understand foreign-accented speech? *Trabalhos Em Linguística Aplicada*, *61*(3), 764–781. https://doi.org/10.1590/010318138668782v61n32022

Stratton, J. M. (2025). The effects of production training on speech perception in L2 learners of German. *Journal of Phonetics*, *108*, 101370. https://doi.org/10.1016/j.wocn.2024.101370

Su, Y., & Kuo, C. C. J. (2022). Recurrent neural networks and their memory behavior: A survey. *APSIPA Transactions on Signal and Information Processing*, *11*(1), e26 (1-38). http://dx.doi.org/10.1561/116.00000123

Thomson, R. I., & Derwing, T. M. (2015). The effectiveness of L2 pronunciation instruction: A narrative review. *Applied Linguistics*, *36*(3), 326-344. https://doi.org/10.1093/applin/amu076

Tikhonova, E., & Raitskaya, L. (2023). ChatGPT: Where is a silver lining? Exploring the realm of GPT and large language models. *Journal of Language and Education*, *9*(3), 5-11. https://doi.org/10.17323/jle.2023.18119

Urip, S., Reli, H., Faruq, U. M., & Mujiyono, W. (2022). Determinants of Technology Acceptance Model (TAM) towards ICT use for English language learning. *Journal of Language and Education*, *8*(2), 17-30. https://doi.org/10.17323/jle.2022.12467

Vishnevskaya, E. M. (2014*). Metodika korrekcii fossilizacii foneticheskih navykov bakalavrov pedagogicheskogo obrazovaniya (na materiale anglijskogo yazyka kak vtorogo inostrannogo)* [Methodology for correcting the fossilization of phonetic skills of bachelors of pedagogical education (based on the material of English as a second foreign language)] [Unpublished doctor dissertation]. Sholokhov Moscow State University for Humanities. https://search.rsl.ru/ru/record/01007483765

Wang, J., Ahmad, N. K. B., Jamil, H. B., & Darmi, R. (2025). Resonating voices: Unpacking EFL teachers' beliefs regarding pronunciation instruction in Chinese tertiary context. *Journal of Curriculum and Teaching*, *14*(1), 30. https://doi.org/10.5430/jct.v14n1p30

Wang, X., & Munro, M. J. (2004). Computer-based training for learning English vowel contrasts. *System*, *32*(4), 539-552. https://doi.org/10.1016/j.system.2004.09.011

Wei, Y. (2025). A study of non-native accent correction techniques combining phonetics, machine learning and biomechanics. *Molecular & Cellular Biomechanics*, *22*(1), 725. https://doi.org/10.62617/mcb725

Zou, B., Liviero, S., Ma, Q., Zhang, W., Du, Y., & Xing, P. (2024). Exploring EFL learners' perceived promise and limitations of using an artificial intelligence speech evaluation system for speaking practice. *System*, *126*, 103497. https://doi.org/10.1016/j.system.2024.103497

APPENDIX A

SCHEMATIC DIAGRAMS: CONVOLUTIONAL AND RECURRENT NEURAL NETWORK

Figure A.1

Schematic Diagram of a Convolutional Neural Network

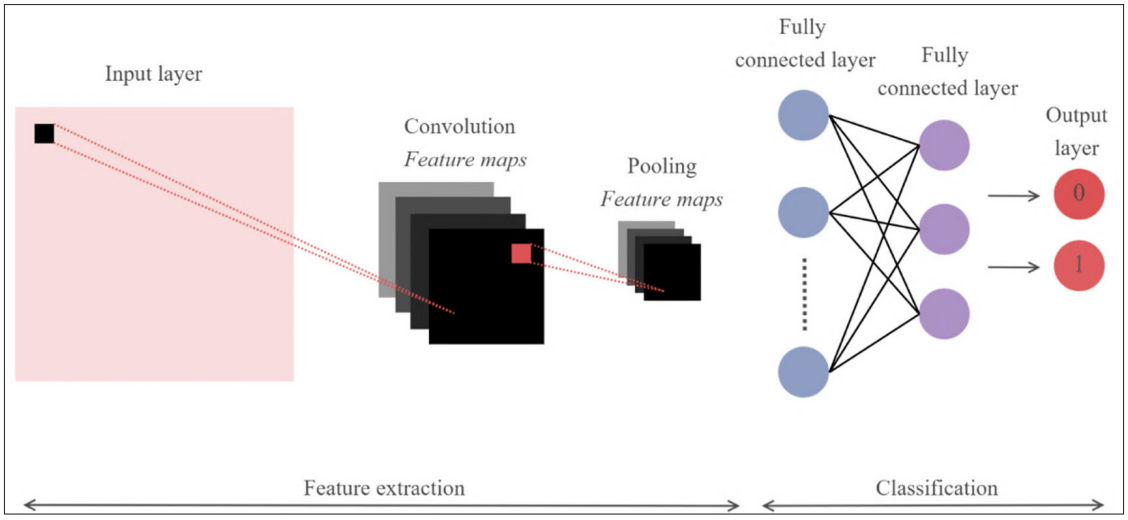
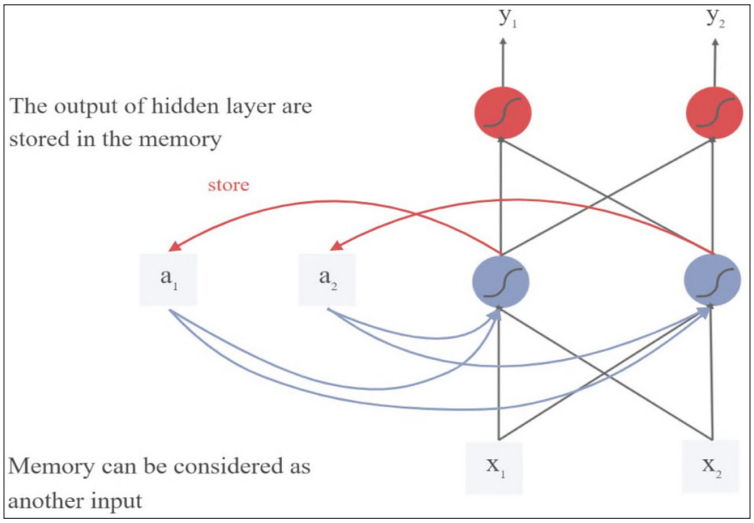


Figure A.2

Schematic Diagram of a Recurrent Neural Network



APPENDIX B

A DETAILED ACCOUNT OF THE SPECIFIC PRE-TEST ITEMS AND THEIR CORRESPONDING PHONETIC PHENOMENA

1. Incorrect articulation of vowels, such as replacing the English sound [ɜ] with Russian [jo] or [э], and English [ɑ:] with Russian [а].

2. Difficulty in distinguishing between vowel length and positional vowel length.

3. Violations in the implementation of English diphthongs, eg., monophthongization of [oʊ] with replacement by [о].

4. Difficulties in implementing the opposition of voiceless/voiced final consonants in English words.

5. Non-discrimination of phonemes [w] and [v], replacing English [s] with Russian [с], and English interdentals [θ] and [ð] with Russian [с], [ф], [т], and [д], [з], respectively.

6. Differences in syllable division, such as the addition of a subsequent consonant to a short vowel in a stressed syllable, which can cause difficulties in determining word stress and result in incorrect pronunciation of vowels in syllables.

APPENDIX C

QUESTIONNAIRE SAMPLES

Figure C.1

Sample Questions

1. Indicate the correct answers in percentages for each of the attempts:

First attempt %		Second attempt %		Third attempt %	
First part of the test	Second part of the test	First part of the test	Second part of the test	First part of the test	Second part of the test

2. Indicate the time to complete the test for each of the attempts:

First attempt	Second attempt	Third attempt

3. Did you use a dictionary to check pronunciation? (Underline the correct option):
   1. Yes
   2. No

4. Are the tasks clearly formulated? (Underline the correct option):
   1.Yes
   2. No

Figure C.2

Sample Questions

5. What recommendations could you suggest for improving this test? _____
6. Have you ever encountered situations where functions of the test were inconvenient? Which? _____
7. Do you think that completing such tasks helps you improve your listening and pronunciation skills in English? _____
8. Was it interesting to complete the test tasks in the first and second modules? What exactly sparked your interest? _____

APPENDIX D

GRAPHICAL USER INTERFACE SAMPLES

Figure D.1
Graphical User Interface (GUI) of Testing Software

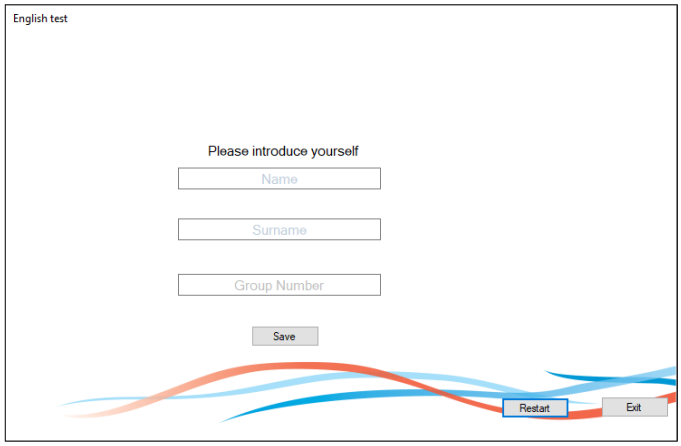


Figure D.2
GUI of the Computer Test for Vowel Recognition

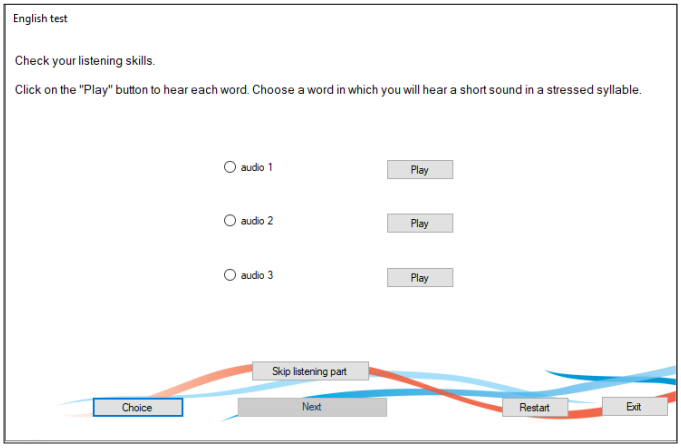


Figure D.3
GUI of the Computer Test for Consonant Recognition

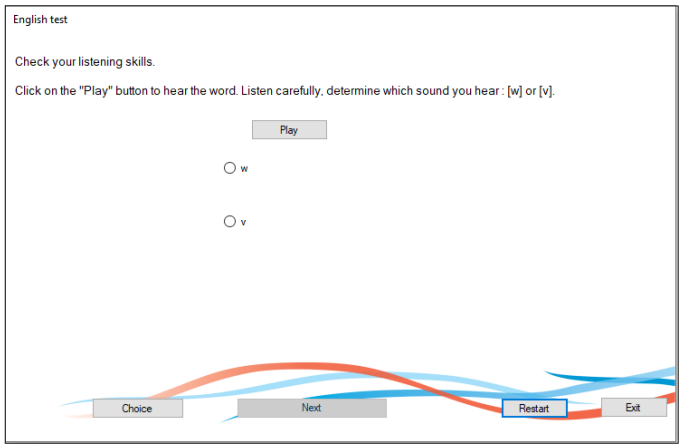
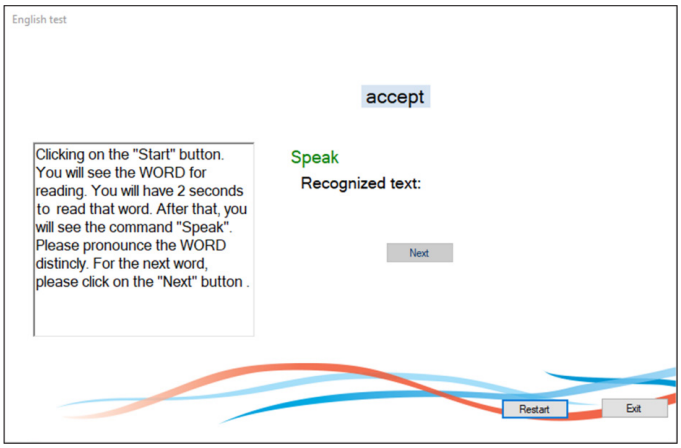


Figure D.4
GUI of the Computer Test with the Speech Recognition Part



https://doi.org/10.17323/jle.2025.21727

GPTBot Development for Translation Purposes: Flowchart, Practical Case and Future Prospects

Helena Ortiz-Garduño, Daniel Torres-Salinas

University of Granada, Granada, Spain

ABSTRACT

Background: This paper explores the development and evaluation of a GPTBot tailored for institutional translation tasks. It addresses a gap in applied research on how generative AI can be adapted for domain-specific translation workflows, particularly in academic institutions.

Purpose: To design and implement UGRBot, a chatbot based on ChatGPT-4 that supports the translation of institutional texts at the University of Granada (UGR) while also outlining a structured and replicable methodology for creating specialised chatbots to enhance translation processes.

Method: The methodology includes: (1) chatbot development using a knowledge base of 57 bilingual institutional documents; (2) evaluation of output quality using BLEU scores, comparing UGRBot with DeepL and Google Translate; and (3) a focused assessment on the translation of 100 institutional terms.

Results: A reference corpus in English of 14,521 words was compiled from UGR administrative and regulatory documents, with human translations serving as the benchmark. BLEU scores were computed using the Natural Language Toolkit library in Python, employing 4-gram analysis for full-text evaluation and bigram analysis for terminology translation.

Conclusion: Results show that UGRBot outperformed both baseline systems in the translation of specialised institutional terminology, achieving the highest BLEU score in this area. However, limitations include lower performance across full-length texts. In conclusion, this research documents the development of a domain-specific GPTBot and its implementation in an institutional context, offering a transferable framework for integrating generative AI into specialised translation workflows.

KEYWORDS

generative AI, ChatGPT-4, translation process, chatbots, GPTBot, institutional translation

INTRODUCTION

The growing adoption of generative artificial intelligence (GAI) models such as ChatGPT has prompted new applications in professional translation workflows. Recent research has focused on three main trends: (1) the use of GAI for general-purpose translation and post-editing (Sahari et al., 2023; Gao et al., 2024); (2) prompt engineering and domain adaptation techniques to enhance translation quality (Zhao et al., 2023); and (3) recognised challenges in aligning AI outputs with context-specific conventions (Siu, 2023a; Ghassemiazghandi, 2024). These trends collectively underscore the growing potential of GAI in translation workflows and many authors suggest that we are on the verge of a new era in translation, with the industry undergoing a transformative technological revolution (Sánchez-Gijón, 2022, Vela-Valido, 2021). However, the role of existing AI-based translation in high-stakes institutional translation remains underexplored.

Despite ongoing efforts to apply generative AI to translation, little attention has been paid to its adaptation for domain-specific tasks, such as institutional document translation, which requires terminological precision, contextual relevance, and adherence to internal style guides. Moreover, commercial tools such as DeepL and Google Translate present

Citation: Ortiz-Garduño, H., & Torres-Salinas, D. (2025). GPTBot development for translation purposes: Flowchart, practical case and future prospects. *Journal of Language and Education, 11*(2), 94-110. https://doi.org/10.17323/jle.2025.21727

Correspondence:
Helena Ortiz-Garduño,
helenaortiz@ugr.es

Received: June 06, 2024
Accepted: June 10, 2025
Published: June 30, 2025

limitations in handling stylistic constraints and ensuring terminological precision in institutional settings. Given these challenges, the central research question guiding this study is how can a customised GPTBot be developed and implemented to enhance the accuracy of institutional translation processes, particularly in the context of translating specialised documents at the University of Granada (UGR).

To address this question, this study aims to develop and evaluate UGRBot, a customised GPTBot for Spanish-English institutional translation tasks at the University of Granada. Specifically, it seeks to: (1) design and configure a UGRBot using ChatGPT-4 and a bilingual institutional knowledge base; (2) assess the system's functionality in translation, terminology extraction, revision, and stylistic assistance; and (3) benchmark its performance against commercial translation engines using BLEU score analysis and functional testing.

While the study is situated within the University of Granada, its findings offer broader relevance for institutions seeking to integrate domain-specific generative AI tools into their translation workflows. The development framework presented here may serve as a replicable model for similar administrative and regulatory contexts.

LITERATURE REVIEW

GAI for General-Purpose Translation

The emergence of generative artificial intelligence (GAI) models has marked a turning point in the field of artificial intelligence by presenting itself as a form of intelligence capable of using natural language processing (NLP) and deep learning to understand human-produced text and generate similar text (Curry et al., 2024). Developed by OpenAI, the latest model available to date[1] is based on the GPT-4 architecture, a large multimodal model capable of processing images and text[2], as well as producing textual output (OpenAI, 2023[3]). In general translation tasks, ChatGPT-4 demonstrates strong capabilities across various subtasks such as text generation, classification, summarisation, sentiment analysis, and machine translation (Hassani & Silva, 2023; Lilli, 2023; Zappavigna, 2023). Several studies confirm that ChatGPT translations rival commercial systems like Google Translate and DeepL (Jiao et al., 2023; Mohsen, 2024). According to Mohsen (2024), ChatGPT-4's superior performance is driven by its large training dataset and advanced algorithms, enabling it to handle diverse genres effectively while integrating updates that reduce biases and errors. Furthermore, Ghassemiazghandi (2024) highlights that translations generated by ChatGPT-4 surpass those produced by the computer-assisted translation tool MateCat and nearly mirror the quality of human translations.

The functionalities of GPT models as translation tools are not limited to translation activity as such but encompass tasks such as contextual clarification and cultural explanation of expressions, explanation of technical terminology and simplification of complex texts, error detection, grammar checking and quality assessment, and stylistic editing and recommendations (Siu, 2023a; Siu, 2023b). In this way, translation professionals can exploit the linguistic fluency and grammatical knowledge embedded in the large language models, without losing full control over the final translation (Siu, 2023a). It is also worth noting that ChatGPT excels in the mechanical phases of translation tasks, but its usefulness decreases in tasks that require judgement, such as fine-tuning and double-checking (Sahari et al., 2023).

Challenges and GPT Customisation for Domain-Specific Translation

While ChatGPT performs well in fluency-oriented tasks (Gao et al., 2024), its ability to maintain institutional terminology and context-specific conventions remains limited (Siu, 2023a; Ghassemiazghandi, 2024). This stems from the fact that these models have been trained on large volumes of general multilingual data, which results in outputs that tend to be overly generic and lack contextual specialisation, as the generated text is inevitably influenced by various prior knowledge rather than being based solely on a specific entry (Gao et al., 2023). For this reason, the importance of configuring the GAI model based on the specific needs for the task at hand is essential to maximise its abilities and obtain relevant results (Zhao et al., 2023). In the specific case of the application of ChatGPT for translation tasks, the use of prompts that focus on the specific translation task and take into account the context can significantly improve its performance (Gao et al., 2023). Thus, a potential approach could lean towards the implementation of chatbots specialised in specific tasks, such as translation, to ensure more accurate and tailored results (Jiao et al., 2023). In this way, the fact that users have the possibility to customise the chatbot for specific use cases, favours more accurate and fluent translations that meet individual needs (Siu, 2023a). The possibility of including a proprietary knowledge base that allows adapting to the user's needs is essential to develop chatbots that are focused on specific functionalities.

[1] ChatGPT-4 was launched on 14 March 2023 and is the latest model currently available to the public, available through the ChatGPT Plus paid subscription plan offered by OpenAI.

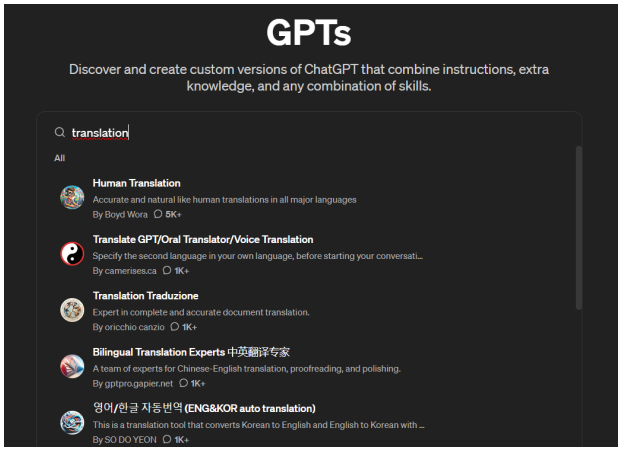
[2] UNESCO. (2023). *ChatGPT e inteligencia artificial en la educación superior: Guía de inicio rápido* [ChatGPT and artificial intelligence in higher education: Quick start guide]. https://unesdoc.unesco.org/ark:/48223/pf0000385146_spa

[3] OpenAI. (2023). *GPT-4 technical report*. http://arxiv.org/abs/2303.08774

The use of prompt engineering is essential to improving ChatGPT's performance for specific translation tasks. Research demonstrates that context-aware prompts can enhance output quality (Zhao et al., 2023; Gao et al., 2023). In addition, integrating proprietary knowledge bases allows models to adapt to user-specific needs, improving accuracy and fluency (Siu, 2023a). In this regard, Yamada (2023) shows how prompt-based strategies enable users to customise ChatGPT for translation workflows, enhancing accuracy even in low-context or technical segments. Similarly, Ngo Cong-Lem et al. (2024) identify persistent limitations in ChatGPT's ability to maintain terminological consistency and global coherence during revision phases, reinforcing the need for task-specific prompt design and structural constraints. Several tools within the ChatGPT-4 'Explore GPTs' store illustrate how user-generated chatbots can be customised for translation, from general tasks to expert translation and proofreading [see Figure 1].

Figure 1

Overview of Translation Chatbots in the GPTs Store



However, while these customised chatbots can be found in the GPTs store, there is currently a lack of academic literature exploring their effectiveness and applications in specialised contexts. Most studies on GPT-based translation tools focus on general-purpose applications, and they often overlook their potential to deal with highly specific domains or contexts, such as institutional translation. Additionally, there is also a lack of complete understanding regarding the nature or quality of the foundation that these models provide, and it remains unclear whether they are fully reliable or trustworthy (Schneider, 2022). Therefore, it is important to define workflows that standardise specific processes, ensuring that users are able to customise these models effectively and consistently for particular tasks and contexts. By focusing on the customisation of the GPTBot specifically for institutional translation tasks at the University of Granada (UGR), this research bridges the gap between the academic research regarding general-purpose AI translation tools and highly specialised applications. By incorporating a proprietary knowledge base and establishing a tailored workflow to configure the model to meet the needs of institutional document translation, this approach demonstrates how customised GPT-based systems can improve translation accuracy and contextual relevance, particularly in institutional settings where consistency and adherence to specific terminology are crucial.

Integrating GAI into Translation Workflows

Integrating GAI into translation workflows requires not only the adoption of advanced technologies, but also a clear framework for ensuring their effective use. Therefore, it is essential to provide guidance on the development of personalised chatbots and to show the importance of following specific guidelines to ensure that quality work is carried out. For this reason, when designing a chatbot specifically for the field of translation, it is essential to identify the stages of the translation process, in order to guarantee the correct development and operation of the GAI model. The translation process encompasses the set of tasks that begin with the receipt of the translation order and culminate with the production of the target text, making use of the necessary tools and strategies to solve the translation problems and carry out the relevant revisions (Hansen, 2013). The purpose of this process is to establish interlinguistic and, as far as possible, intercultural equivalences that allow the meaning of a source text to be transferred to a target text, taking into account the specifications of the translation assignment given by the client (Parra-Galiano, 2006). In general terms, the three main phases of the translation process can be classified into pre-drafting, drafting and post-drafting (Dimitrova, 2010; Mossop, 2000).

The initial phase of pre-drafting lays the groundwork for the translation work. It includes planning, orientation and detailed reading of the source text. During this initial stage, the translator carries out a pre-translational analysis, considering extratextual (intention, function, sender, receiver, etc.) and intratextual (subject matter, content, presuppositions, lexis, syntax, etc.) factors (Nord, 1991). This analysis allows the translator to orient themself as to how to approach the text, taking into account the author's intention and the expectations of the target text's audience.

The drafting phase consists of the actual translation of the source text into the language of the target text. This stage involves the transfer of the meaning of the received message into the target language, either on the basis of an equivalence relationship between lexical items or, in the case of a different text function, according to the function of the target text (Nord, 1991).

The post-drafting phase constitutes the revision phase of the translated text. This stage is essential to guarantee the quality of the target text and consists of determining whether the final product complies with the specifications of the translation order through a series of criteria, with the aim of making the relevant corrections or improvements, before

considering the target text as final and ready for delivery to the client (Parra-Galiano, 2006).

The translation process, while generally based on these main phases, is adapted according to the specific needs and operational context of the translator or organisation. The individual translator's practice differs significantly from the operation of a translation company in terms of procedure, available resources and capacity to handle large projects. For example, in the case of organisations belonging to the European Union, the workflow is highly standardised to handle the huge number of documents requiring translation in their multiple language combinations, as well as relying on CAT tools from the pre-processing phase (document type, domain, source and target language(s), deadline, etc.), pre-translation through a series of translation memories, to the use of terminology databases and translation management systems to maintain document consistency and quality[4]. Similarly, the Translation Centre for the Bodies of the European Union[5] deals with client requests through a standardised workflow that starts with receiving requests, preparing reference material, assigning the work to in-house or freelance translators depending on the specialisation and needs of the assignment, and carrying out a thorough technical review and quality control by in-house translators.

In this context, maintaining consistency, accuracy, and efficiency is a significant challenge, especially when dealing with large volumes of documents and specialised terminology. Traditional translation methods, while effective, often require substantial human intervention. For this reason, an increasing number of corporations and institutions are implementing the use of digital tools to carry out their projects (Rodríguez-de Céspedes, 2020).

Despite advances in generative AI applications for translation, no study has yet explored the implementation of a GPT-based chatbot customised with a bilingual institutional corpus and internal style guide for professional translation workflows. This study responds to that gap by developing and evaluating UGRBot, a domain-specific GPTBot tailored to the translation needs of the University of Granada. In doing so, it contributes not only to the theoretical understanding of prompt engineering and model adaptation but also offers a replicable model for the integration of GAI into institutional translation practices.

Evaluation Metrics in Translation

Human evaluation remains the traditional benchmark for assessing translation quality, as it allows evaluators to consider contextual adequacy, terminological precision, and stylistic coherence. However, it is also time-consuming, costly, and can lead to inconsistencies, particularly when multiple reviewers are involved or when evaluation criteria are insufficiently standardised (Läubli et al., 2020). These limitations are especially problematic in experimental studies that require comparability across translation systems.

To address these challenges, automatic metrics such as the BLEU score (Bilingual Evaluation Understudy) have become widely adopted in machine translation research. BLEU calculates n-gram overlap between candidate and reference translations, offering a fast and replicable method for benchmarking translation outputs (Papineni et al., 2002). Although BLEU does not fully capture semantic or pragmatic adequacy, its simplicity and standardisation make it a useful complement to human evaluation, particularly when comparing systems or tracking improvements (Callison-Burch et al., 2006).

In this study, BLEU is employed alongside human review to assess the translation performance of UGRBot. This dual approach ensures methodological rigour while addressing both quantitative benchmarking and qualitative validation in institutional translation workflows.

METHOD

Research Design

This study follows a design-based research approach aimed at developing, configuring, and evaluating UGRBot, a GPT-based translation assistant tailored to the institutional needs of the University of Granada. The core objective is to assess whether a customised GPTBot (integrating a bilingual institutional knowledge base) can perform institutional translation tasks more effectively than widely used machine translation engines, such as Google Translate or DeepL. The unit of analysis comprises Spanish–English administrative and regulatory documents produced within the university context, and the evaluation framework integrates BLEU score analysis, as well as qualitative evaluation for functional assessment. The following sections detail the phases for the development of a generative artificial intelligence chatbot using ChatGPT-4, the information regarding the data collection, and the BLEU score analysis used to evaluate the quality of the translations.

GPTBot Development Phases

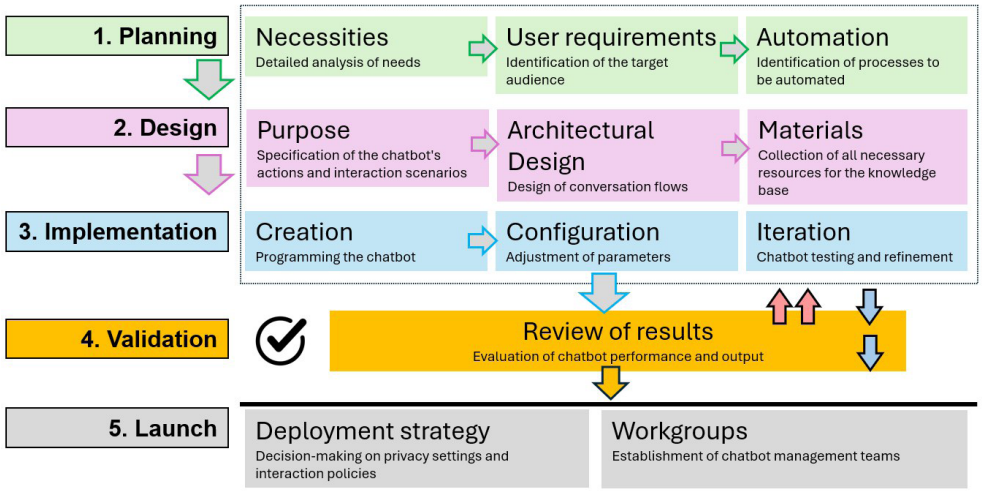
The ChatGPT-4 model includes among its functionalities the possibility to create customised chatbots. To use the latest version of the model, it is necessary to have access to the ChatGPT Plus paid subscription plan. The development

[4] Directorate-General for Translation. (2018). *Translation services in the digital world: A sneak peek into the (near) future: DG TRAD Conference (16-17 October 2017)*. European Union. https://data.europa.eu/doi/10.2861/823102

[5] Translation Centre for the Bodies of the EU. (n.d.). *The Centre's workflow*. https://cdt.europa.eu/en/centres-workflow

Figure 2

Chatbot Development Flowchart



phases, from planning to launch, are shown in Figure 2. The processes to be carried out at each stage are detailed below.

The development of the GPTBot begins with a structured planning phase aimed at establishing a solid foundation for targeted and effective implementation. This phase encompasses three areas: needs analysis, target audience profiling, and task automation identification. The needs analysis involves a comprehensive evaluation to determine the specific requirements of the chatbot, ensuring that the scope and objectives are clearly defined, thus providing a precise framework for its functionality and intended outcomes. Subsequently, it is necessary to profile the target audience to align the chatbot's design with user needs and interaction patterns. This ensures the chatbot design is user-friendly and aligns with the expectations of its intended users, thereby optimising user engagement and satisfaction. Lastly, automatable tasks within existing workflows must be identified. This phase involves spotting repetitive or time-consuming processes that the chatbot can efficiently manage, in order to enhance overall productivity and allow human resources to concentrate on higher-level functions.

The design phase of chatbot development is dedicated to detailing the functional and technical aspects of the GPTBot, divided into three specific subphases: defining the purpose, architectural design, and gathering materials. Defining the purpose involves specifying the core functions of the chatbot, detailing what it is expected to achieve and the interaction scenarios it must handle, which guides the development of relevant features and interactions. The architectural design focuses on the structural design of the chatbot, including the setup of the conversation flow and integration with existing systems. The "Capabilities" parameter enables advanced functionalities such as Web Browsing, to access up-to-date information from the Internet; DALL-E Image Generation, to create creative images from textual descriptions; and Code Interpreter, to enable code interpretation and execution. The appropriate configuration of these parameters depends on the specific objectives of the chatbot and the needs of the target audience.

Essential to the design process is the gathering of resources required to build and support the chatbot. GPTBots designed with the GPT-4 model have the ability to integrate a specific knowledge base to enrich their responses and improve their accuracy on specific topics. This feature allows the chatbot to not only rely on the large dataset it was initially trained with, but also to use updated or specialised information that is relevant to the GPTBot's scope of application. This functionality can be implemented through the "Knowledge" parameter, and feed the GAI model with documents, guides, FAQs or other resources specific to the domain of interest. In this way, the user experience is improved by providing more detailed and contextually appropriate answers. The possibility of integrating a specific knowledge base into the GPTBots allows the use of proprietary data, which enriches the answers and improves their accuracy on specific topics. To date, this knowledge base supports up to 20 documents, with a total limit of 100 gigabytes.

The implementation phase transitions the conceptual design of the GPTBot into a functioning entity, and involves the chatbot's creation, configuration and iteration. The chatbot is developed according to the detailed design specifications outlined in the design phase. The initial setup utilises the "Create" option, which allows direct interaction with ChatGPT to automatically populate designated sections based on predefined conversation parameters. To access the interface for creating a GPTBot, it is necessary to click on the "Explore GPTs > Create a GPT" tab in ChatGPT Plus. This method offers a straightforward approach to configur-

ing the initial prototype, providing a practical foundation for rapid development. However, to meet specific user needs more precisely, it is advisable to proceed beyond this basic setup and engage in more detailed customisation through advanced configuration settings. The configuration process involves the meticulous adjustment of the chatbot's parameters to enhance its interaction capabilities and functional performance. Parameter tuning is essential for refining how the chatbot responds to user inputs, managing data processing, and ensuring that the chatbot behaves in a manner that is both user-friendly and aligned with the intended use cases. The advanced "Configure" interface provides a series of sections to be completed to tailor the chatbot to the specific needs to be met.

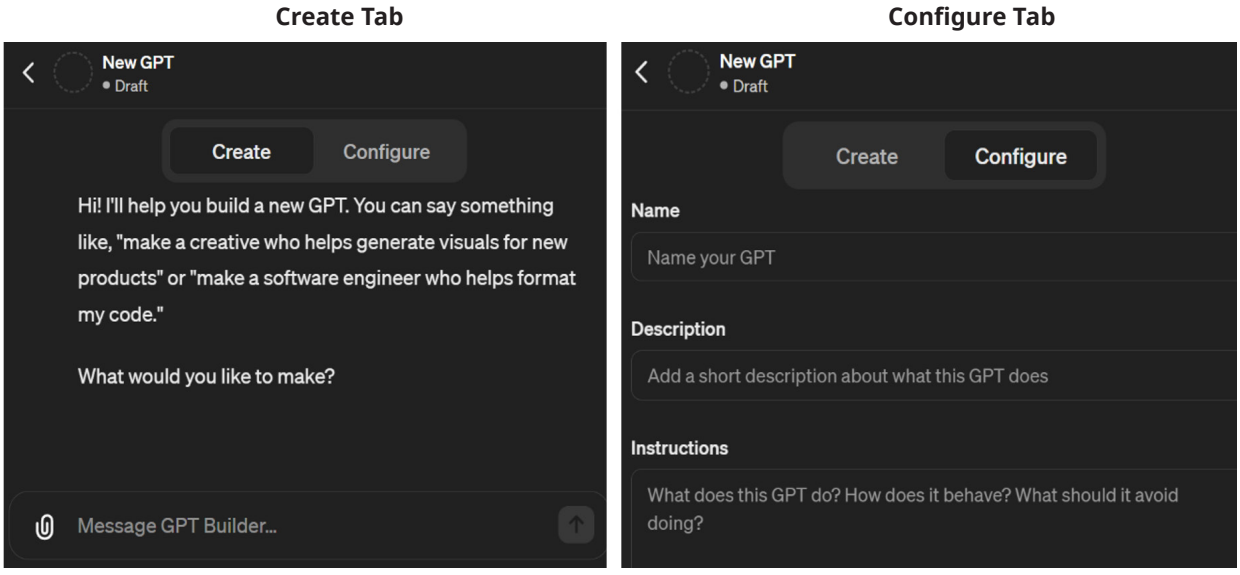
As can be seen in Figure 3, the configuration parameters include, on the one hand, the name and description of the chatbot. These are fundamental elements that not only help identify the chatbot within the system, but also provide end-users with a direct understanding of the chatbot's purpose. The name should be unique and easy to remember, while the description should be concise but informative, providing a summary of the help the chatbot can provide to the user. On the other hand, there is the parameter related to the chatbot's internal instructions, which are essential to define how the GAI model processes prompts and generates responses. This set of instructions dictates the chatbot's behaviour when faced with different types of interactions and determines how the chatbot handles conversations with the user based on the established context. This is critical to enable a consistent and relevant interaction, where the chatbot is able to maintain a fluid line of dialogue, remembering previous details of the conversation and adjusting its responses accordingly. The "Conversation starters" parameter is used to provide initial examples of how users can begin their interactions with the chatbot, offering suggestions of questions or topics they can address. This makes it easier for users to get an initial idea of the type of queries the chatbot is capable of handling. It is important that these examples reflect the variety of functionalities of the chatbot, so that users can have an overview of the type of interactions they can have with the AI model. For this reason, it is essential that there is a detailed analysis of the target users' needs as well as the chatbot's functionalities.

The iteration stage encompasses rigorous testing and continuous refinement of the chatbot. Iterative improvements are made based on previous interactions and on the degree of appropriateness of the answers provided. The evaluation and improvement cycle must be continuous so that the GAI model is always adapted to the needs of the users, thus optimising its performance and relevance. The refinement phase is essential in the initial implementation of the chatbot, but it is simply the starting point for an iterative process of post-implementation evaluation and improvement. The evaluation and improvement cycle must be continuous so that the GAI model is always adapted to the needs of the users, thus maintaining its effectiveness and relevance over time.

The validation phase ensures that the chatbot fulfils its designed purposes effectively. During this stage, a comprehensive evaluation is conducted to determine if the chatbot achieves the set objectives, delivers accurate responses, and supports the export of results in the correct formats. This assessment helps identify the chatbot's successes and areas needing improvement, guiding ongoing refinements to optimise functionality and user experience. The launch phase marks the transition of the chatbot from development to active use. Key decisions regarding the deployment strategy

Figure 3

ChatGPT-4 Chatbot Design Interface with Details of the Create and Configure Tabs



are made, including setting privacy controls and interaction policies. The GPTBot can be deployed publicly or privately, depending on the specific requirements and security considerations of the intended environment. Dedicated teams are established to manage the ongoing operations and maintenance of the chatbot. These workgroups are responsible for handling updates, resolving issues, and ensuring that the chatbot continues to function effectively.

Data Collection

The GPTBot designed to streamline institutional translation at the University of Granada is powered by a knowledge base that includes specific UGR documentation. It contains a repository of UGR administrative documents, regulations and policies in the Spanish-English language combination [see Table 1]. Specifically, it is divided into two main categories: 12 UGR administrative documents in Spanish, each with a corresponding English translation, and 16 UGR regulations and policies, also paired with their English translations [see Appendix 1 for a complete list of documents]. The knowledge base is also fed by the UGR English Style Guide, specifically designed to help in the writing of institutional texts in English or in the translation of texts into English in the context of the University of Granada. Given the limitations of ChatGPT-4's current configuration, which supports a knowledge base of up to 20 documents with a total capacity of 100 gigabytes, the 57 documents used to train the GPTBot had to be combined into two distinct files to meet these constraints. The total of these 57 documents were used to train the GAI model, with the aim of ensuring the quality and contextual appropriateness of the translations.

BLEU Score Evaluation

To assess the translation quality of UGRBot, a BLEU score (Papineni, 2002) analysis was conducted. The BLEU score is a widely recognised metric used in machine translation to measure how closely a machine translation aligns with a set of reference translations. The BLEU score is calculated by counting the number of n-grams from the system's output that occur in the reference translations. The formula for calculating the BLEU score is as follows:

$$BLEU = BP \cdot \exp\left(\sum_{n=1}^{N} w_n \log P_n\right)$$

Where:

BP is the brevity penalty, which penalises translations that are shorter than the reference translation.
w_n is the weight assigned to the n-grams (typically equal for all n-grams).
p_n is the precision of the n-grams, representing the ratio of matched n-grams between the machine-generated translation and the reference translation.

For the purpose of this research, the translation output of UGRBot designed with ChatGPT-4 was compared with two of the most advanced translation engines currently available: DeepL and Google Translate. These comparisons were carried out under a controlled evaluation setting, where the same source texts were processed by all systems and compared against a human-translated reference corpus.

The reference corpus, which functioned as the gold standard for evaluation, consisted of 14,521 words, including random excerpts from the UGR's repository of regulations and the UGR's repository of administrative documents. The reference corpus served as the benchmark to evaluate the accuracy and quality of the translations produced by each system. In addition to this, a separate evaluation was conducted focusing on the translation of 100 specialised terms, extracted from UGRTerm[6], a bilingual (Spanish-English) database of academic and institutional terms used at the UGR. This evaluation aimed to assess the consistency and precision of each system—ChatGPT-4, DeepL, and Google Translate—in handling the translation of institutional terminology. It is important to note that this evaluation is based on a limited set of institutional documents and terminology,

Table 1
UGRBot Knowledge Base

Languages	Source	No. documents
Spanish	UGR regulations and policies repository	16
	UGR administrative documents repository	12
	UGRTerm language resources	1
English	UGR regulations and policies repository	16
	UGR administrative documents repository	12

[6] University of Granada. (2019). UGRTerm: UGR online resource on academic and institutional terminology (Spanish-English). https://ugrterm.ugr.es/en/

which may affect the generalisability of the results beyond the administrative domain of the University of Granada.

The Natural Language Toolkit (nltk) library in Python was employed to compute the BLEU score. Both reference and candidate translations were tokenised, splitting the text into individual words. In both cases (full-text translations and specialised terms), all machine-generated translations were compared against the reference translation. For the specialised terms, bigrams were used to calculate the BLEU score, as these represent shorter text strings. In the case of full-text translations, the standard approach of using 4-grams was applied. Afterward, the average BLEU score was calculated separately for the full-text translations and the specialised terms.

RESULTS

The implementation of UGRBot — a GPTBot designed for the University of Granada's internal community, including translators, teaching and research staff, and administration and services staff — presents a novel approach to handling institutional translation, terminology and revision tasks. This section outlines the outcomes of UGRBot's development and evaluates its performance based on the specified objectives.

Practical Case: UGRBot for Institutional Translation

The development of a chatbot (https://chat.openai.com/g/g-ZzjDW0drV-ugr-bot-for-institutional-translation) for institutional translation at the UGR begins with planning the objectives designed to meet the specific needs of its internal community, namely internal and freelance translators, teaching and research staff, and administration and services staff. This chatbot helps to streamline the translation and revision processes of institutional documents of the UGR in the Spanish-English language combination, especially UGR administrative documents, regulations and policies.

The design of the chatbot is intended to provide support throughout the entire translation process, from the preparation and analysis of the source texts to the revision of the translations. Thus, the main functionality of the GAI model consists of the Spanish-English translation of UGR institutional documents. The chatbot can also create tables with terminology specific to the documents, provide their English equivalents, and solve problems of wording and style in English during the translation and revision phases, among other things. The chatbot's knowledge base is fed with specific UGR documentation [see Table 1], with the aim of ensuring the quality and contextual appropriateness of the translations.

Once the conceptual phases were completed, the implementation of the chatbot was carried out. Through the basic configuration interface "Create", a detailed prompt was introduced to generate the chatbot's internal instructions with the conceptual features mentioned above, particularly underlining the relevance of incorporating the reference documents present in the knowledge base for the generation of accurate and contextually appropriate responses. Specifically, the prompt used was the following: "Generate the best instructions for a GPTBot based on the following information: [features + knowledge base emphasis]". This methodological approach allowed the creation of a prototype that provided an initial starting point for refining and improving the system through iteration processes [see Figure 4]. In the advanced configuration interface, the relevant changes and specifications were made.

The process of iteration and refinement of the chatbot, consisting of the improvement of the internal instructions based on feedback according to the answers provided, resulted in the formulation of more concrete and precise instructions. Specifically, the instructions are structured to clarify the chatbot's specific application, operating procedures and key functionalities [see Appendix 2]. Initial protocols for interaction with the user are included and emphasis is placed on the chatbot not inventing answers, but relying primarily on the chatbot's knowledge base, with web searches limited to the consultation of official UGR sources. The main tasks of the chatbot are also indicated, divided into translation, terminology extraction, text revision and management of stylistic queries [see Figure 5].

The conversation starters respond to the four main functions of the chatbot, focusing on translation, terminology extraction, revision and stylistic correction. In this way, clear and effective entry points for user interaction are established. To validate the correct functioning of the chatbot, tests were carried out focusing on translation queries, terminology extraction, revision tasks and style queries.

Translation Queries

The translation queries consisted of asking the chatbot to translate UGR institutional texts of different lengths and formats for which the chatbot used the regulations and administrative documents included in its knowledge base. Therefore, this integration allowed the translations to be not only linguistically correct, but also consistent with the specific use of terms and styles preferred by the UGR. It was observed that, when translating short texts (2-3 pages) and in Word format, the chatbot provided higher quality answers than when dealing with longer documents or in PDF format, where the quality of the translations was more variable and sometimes even incomplete. To quantitatively assess the translation quality of UGRBot, a BLEU score analysis was conducted, comparing its performance with two leading

machine translation systems, DeepL and Google Translate. The results are presented in Table 2.

Terminology Extraction

In terms of terminology extraction, the tasks required generating Word and Excel tables listing Spanish terms and their English equivalents related to higher education and research, indicating gender only for those terms referring to people. Generally, the chatbot is able to perform this task correctly, complying with the specific instructions when completing the tables. However, it sometimes includes terms that do not specifically belong to the academic or research fields of the UGR. Likewise, it was detected that it sometimes failed to correctly recognise polylexic units, which are frequent in specialised terminology. BLEU score analysis allowed for a comparison of ChatGPT compared to Google Translate and DeepL system's performance in both full-text institutional translations and specialised terminology. The results are presented in Table 3.

Revision Tasks

Figure 4

Design of Instructions with the Help of the GAI's Own Model

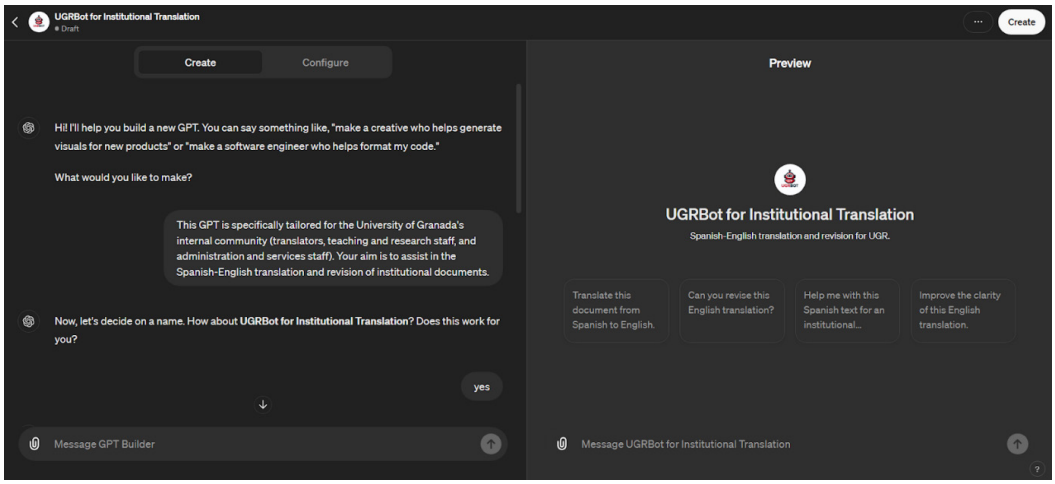


Figure 5

Advanced GPTBot Configuration for Institutional Translation

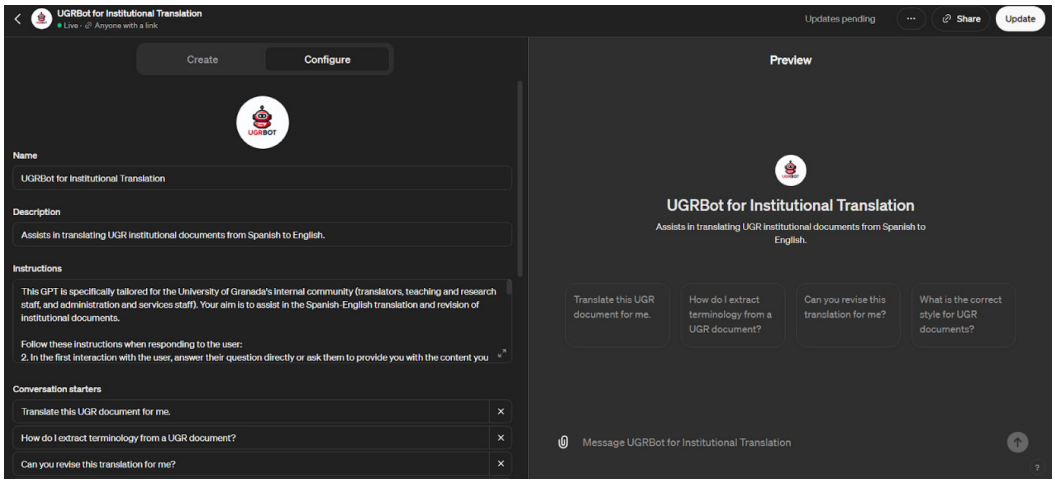


Table 2

BLEU Scores for Full-Text Translations

Machine translation system	BLEU score
UGRBot with ChatGPT-4	0.377
DeepL	0.417
Google Translate	0.374

Table 3

BLEU Scores for Specialised Terms

Machine translation system	BLEU score
UGRBot with ChatGPT-4	0.472
DeepL	0.375
Google Translate	0.348

With regard to the revision tasks, monolingual revisions were carried out, both in Spanish and English, as well as bilingual revisions of Spanish-English and English-Spanish texts in order to evaluate whether the original text was effectively maintained and only the strictly necessary changes were made. To achieve this, the chatbot was provided with stylistically correct texts that contained some terminological errors to test its ability to identify and correct them while maintaining the integrity of the original text. However, this functionality has not yet produced the expected results for the moment, since the chatbot tends to modify unnecessary parts of the original text, which are completely valid and do not involve any translation or stylistic errors. This behaviour may be due to an over-interpretation of the chatbot's internal instructions or of the style rules of the knowledge base.

Style Queries

The style queries consisted of a series of specific questions related to the context of use of the UGR, such as whether British or American spelling should be used in institutional documents; different spelling conventions, such as the rules of capitalisation in UGR terminology; or questions of accessible and inclusive language. For example, in response to the prompt "Should the term 'Vice-Rector for Research' be capitalised?", UGRBot correctly indicated that capitalisation should be maintained, citing alignment with the conventions outlined in the UGR English Style Guide for official institutional titles. In these queries, the chatbot performed adequately, providing accurate and well-founded answers that were aligned with the UGR's institutional policies and preferences according to the UGR English Style Guide that feeds the knowledge base of the GAI model. This approach not only provides users with reliable guidance but also encourages consistency in stylistic choices across institutional documents.

As a prototype, this chatbot has been launched privately in order to protect the data provided by the Language Services Unit (USL) of the UGR. Once it is made public, dedicated working groups will be created for the ongoing operation and maintenance of the chatbot. Since the intention of this work is to provide a set of guidelines for the development of a chatbot in translation and not to present a final product, this phase is still under development.

DISCUSSION

The primary objective of this study was to develop and evaluate a specialised GPTBot designed to enhance translation tasks, with a focus on institutional document translation at the University of Granada. The findings confirm that integrating a domain-specific knowledge base into ChatGPT-4 can significantly improve translation accuracy, particularly in handling specialised terminology. The GPTBot's successful deployment in translating institutional documents from Spanish to English highlights the potential for AI-driven tools to significantly enhance translation accuracy and consistency (Ghassemiazghandi, 2024; Jiao et al., 2023). Additionally, the study sheds light on the largely unexplored area of AI applications in specialised contexts (see, e.g., Gao et al., 2024; Mohsen, 2024), demonstrating the importance of tailored solutions for domain-specific translation tasks.

Translation Quality

The results revealed that although DeepL slightly outperformed UGRBot in the overall translation of full institutional texts (with a BLEU score of 0.417 compared to UGRBot's 0.377 and Google Translate's 0.374), this advantage does not necessarily imply superior handling of specialised content. Instead, it suggests that DeepL may produce translations that are slightly more fluent and natural in longer texts. In fact, the translation queries submitted to UGRBot during the process of GPT validation involved institutional texts of varying lengths and formats, using the UGR's regulations and administrative documents as part of the chatbot's knowledge base. Performance varied depending on the length and format of the documents and UGRBot produced higher quality translations for shorter texts (2-3 pages) in Word format, where the system could more accurately leverage the knowledge base and provide consistent results. In contrast, when dealing with longer documents or PDF formats, the quality of the translations was more variable, with some translations being incomplete or less precise. The lack of precision may be due to the inherent creative component of ChatGPT-4, which can sometimes lead to difficulties in strictly adhering to the knowledge base. This variability in performance may explain why DeepL scored higher overall for full-text institutional translations.

Terminology Extraction

UGRBot excelled in the translation of specialised terminology, achieving a BLEU score of 0.472 compared to DeepL's 0.375 and Google Translate's 0.348. The GPTBot's superior terminological accuracy can likely be attributed to the integration of a bilingual institutional corpus. This supports prior findings that domain adaptation enhances precision in LLM-driven translation (Zhao et al., 2023). The knowledge base integration played a key role in ensuring that the translations adhered to the institution's preferred terminology, which was instrumental in its strong performance in this area. Moreover, in terms of terminology extraction, the tasks carried out during the validation phase required generating Word and Excel tables listing Spanish terms and their English equivalents related to higher education and research, indicating gender only for those terms referring to people. Generally, the chatbot is able to perform this task correctly, complying with the specific instructions when completing the tables. However, it sometimes includes terms that do not specifically belong to the academic or research fields of the UGR. Likewise, it was detected that it sometimes failed

to correctly recognise polylexic units, which are frequent in specialised terminology.

Revision Tasks

UGRBot's behaviour in stylistic revision tasks was inconsistent. While the chatbot occasionally succeeded in improving clarity and coherence, it also introduced unnecessary modifications, even when explicitly instructed to preserve the original structure and intent. These outcomes suggest a tension between generative flexibility and conservative editing practices, which warrants further exploration. Some of these modifications may be interpreted as hallucinations, as previously observed in studies highlighting ChatGPT's limitations in judgement-based operations and domain-specific reliability (Siu, 2023a; Ngo Cong-Lem et al., 2024). The revision module may benefit from more restrictive prompting strategies that prevent over-editing.

Style Queries

Unlike previous evaluations of ChatGPT in generic translation tasks (Gao et al., 2023), our findings suggest that institutional fine-tuning can substantially improve output adequacy and stylistic control. Results indicate that GPTBot reliably follows the UGR English Style Guide, included in the knowledge base. The chatbot's ability to respond accurately to style-related queries related, for instance, to spelling, capitalisation, and inclusive language demonstrates that prompt-based design, when guided by a coherent knowledge base, supports stylistic consistency in administrative settings.

Limitations

The GPTBot's design and operational framework effectively leveraged AI capabilities to meet the specific needs of the university's internal community, showcasing the potential of AI-driven solutions in administrative and academic settings. Nevertheless, several limitations emerged during its implementation. The most significant issue was its performance in text revision tasks, where the chatbot struggled to maintain accuracy and consistency. Despite the implementation of specialised commands instructing the chatbot to adhere strictly to the knowledge base and make only the necessary changes, in most cases, the principle of preserving the integrity of the original content was not consistently upheld.

Moreover, the current configuration of the GPTBot's knowledge base supports up to 20 documents, with a total limit of 100 gigabytes, which may restrict its ability to handle larger or more varied datasets. In terms of potential biases, the GPTBot's translations are influenced by the data it was trained on, particularly the terminology and style guidelines specific to the UGR. Due to the 100 GB limit, the inability to include a broader range of documents raises concerns about the generalisability of the chatbot's translations, as a more diverse set of documents would likely enhance its ability to generate more contextually accurate translations. However, an attempt has been made to incorporate a diverse range of UGR institutional documents addressing different needs, ensuring that the chatbot's knowledge base is as complete and relevant as possible within existing constraints.

CONCLUSION

This study set out to evaluate the potential of a customised GPTBot, built using the ChatGPT-4 framework, to support institutional translation workflows in a university context. Specifically, the study aimed to (1) define a structured methodology for the correct development of chatbots, (2) implement and test UGRBot, a chatbot specialised in translation purposes at the University of Granada, and (3) assess its impact on translation accuracy, efficiency, and workflow optimisation.

The findings indicate that a prompt-engineered, institution-specific GPTBot outperforms commercial translation tools in handling domain-specific terminology and adhering to internal style guides. These results support the viability of lightweight, localised AI solutions for academic-administrative communication.

The proposed methodology can be adapted to other institutional environments where internal communication requires terminological accuracy and stylistic consistency. Internal university (internal and freelance translators, teaching and research staff, and administrative and services personnel) may benefit from adopting similar GPTBot configurations, provided they have access to well-curated institutional corpora and appropriate digital infrastructure.

The findings of this study demonstrate the transformative potential of GPTBots within the translation industry and indicate a promising direction for the ongoing advancement of artificial intelligence in language-related applications. Future investigations should prioritize enhancing the GPTBot's ability to address current limitations in adapting to instructions and producing the intended outcomes, particularly in revision tasks where excessive editing may undermine the reliability of the output. Additionally, further efforts should be directed toward facilitating the public implementation of UGRBot within the internal community of the University of Granada.

DISCLAIMER

The authors used ChatGPT-4 in the preparation of this manuscript for grammar, spelling, and stylistic revision across the entire text. Moreover, ChatGPT-4 was employed as a research tool for the specific purpose of developing a specialised GPTBot focused on institutional translation practices at the University of Granada. All outputs generated by the tool were reviewed by the authors to ensure academic integrity.

FUNDING

This work was supported by the Spanish Ministry of Universities [Predoctoral Grants for the Training of University Lectures (FPU), FPU21/01204] and and the Language Services Unit (USL) of the University of Granada for providing the materials used in this work

DECLARATION OF COMPETING INTEREST

None declared.

AUTHORS' CONTRIBUTION

Helena Ortiz-Garduño: conceptualisation; data curation; formal analysis; funding acquisition; methodology; project administration; visualisation; writing – original draft; writing– review & editing.

Daniel Torres-Salinas: conceptualisation; data curation; formal analysis; investigation; methodology; resources; software; supervision.

REFERENCES

Callison-Burch, C., Osborne, M., & Koehn, P. (2006). Re-evaluating the role of BLEU in machine translation research. *Journal of Artificial Intelligence Research*, 25, 191–218.

Curry, N., Baker, P., & Brookes, G. (2024). Generative AI for corpus approaches to discourse studies: A critical evaluation of ChatGPT. *Applied Corpus Linguistics*, *4*(1). https://doi.org/10.1016/j.acorp.2023.100082

Dimitrova, B. E. (2010). Translation process. In Y. Gambier & L. van Doorslaer (Eds.), *Handbook of translation studies* (1st ed., vol. 1, pp. 406–411). John Benjamins Publishing Company.

Gao, R., Lin, Y., Zhao, N., & Cai, Z. G. (2024). Machine translation of Chinese classical poetry: A comparison among ChatGPT, Google Translate, and DeepL Translator. *Humanities and Social Sciences Communications, 11*. https://doi.org/10.1057/s41599-024-03363-0

Gao, Y., Wang, R., & Hou, F. (2023). *How to design translation prompts for ChatGPT: An empirical study*. arXiv:2304.02182. http://arxiv.org/abs/2304.02182

Ghassemiazghandi, M. (2024). An evaluation of ChatGPT's translation accuracy using BLEU Score. *Theory and Practice in Language Studies, 14*(4), 985–994. https://doi.org/10.17507/tpls.1404.07

Hansen, G. (2013). Translation process as object of research. In C. Millán & F. Bartrina (Eds.), *The Routledge Handbook of Translation Studies* (1st ed., pp. 88–101). Routledge.

Hassani, H., & Silva, E. S. (2023). The role of ChatGPT in data science: How AI-assisted conversational interfaces are revolutionizing the field. *Big Data and Cognitive Computing*, *7*(2), 62. https://doi.org/10.3390/bdcc7020062

Jiao, W., Huang, J., Wang, W., He, Z., Liang, T., Wang, X., Shi, S., & Tu, Z. (2023). *ParroT: Translating during chat using large language models tuned with human translation and feedback*. arXiv:2304.02426. http://arxiv.org/abs/2304.02426

Jiao, W., Wang, W., Huang, J., Wang, X., Shi, S., & Tu, Z. (2023). *Is ChatGPT a good translator? Yes with GPT-4 as the engine*. arXiv:2301.08745. http://arxiv.org/abs/2301.08745

Läubli, S., Sennrich, R., & Volk, M. (2018). *Has machine translation achieved human parity? A case for document-level evaluation*. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing* (pp. 4791–4796). Association for Computational Linguistics. https://doi.org/10.18653/v1/D18-1512

Lilli, S. (2023). ChatGPT-4 and Italian dialects: Assessing linguistic competence. *Umanistica Digitale,* (16), 235–263. https://doi.org/10.6092/issn.2532-8816/18221

Mohsen, M. (2024). Artificial intelligence in academic translation: A comparative study of large language models and Google Translate. *Psycholinguistics, 35*(2), 134–156. https://doi.org/10.31470/2309-1797-2024-35-2-134-156

Mossop, B. (2000). The workplace procedures of professional translators. In A. Chesterman, N. Gallardo-San Salvador, & Y. Gambier (Eds.), *Translation in Context: Selected papers from the EST Congress, Granada 1998* (pp. 39-48). John Benjamins Publishing Company. https://doi.org/10.1075/btl.39.07mos

Ngo Cong-Lem, S., Soyoof, A., & Tsering, D. (2024). A systematic review of the limitations and associated opportunities of ChatGPT. *International Journal of Human–Computer Interaction*, *40*(6), 515–537. https://doi.org/10.1080/10447318.2024.2344142

Nord, C. (1991). *Text analysis in translation: Theory, methodology, and didactic application of a model for translation-oriented text analysis* (1st ed.). Rodopi.

Papineni, K., S. Roukos, T. Ward and W. Zhu. (2002). BLEU: A method for automatic evaluation of machine translation. *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics* (pp. 311–318). ACM. https://doi.org/10.3115/1073083.1073135

Parra-Galiano, S. (2006). La revisión y otros procedimientos para el aseguramiento de la calidad de la traducción en el ámbito profesional [Revision and other procedures for ensuring translation quality in the professional field]. *Revue de Traduction et d'Interprétation – Journal of Translation Studies*, *15*(2), 11–48. http://hdl.handle.net/10481/7369

Rodríguez-de Céspedes, B. (2020). Beyond the margins of academic education: Identifying translation industry training practices through action research. *Translation & Interpreting, 12*(1), 115-126. https://doi.org/10.12807/ti.112201.2020.a07

Sahari, Y., Al-Kadi, A. M. T., & Ali, J. K. M. (2023). A cross sectional study of ChatGPT in translation: Magnitude of use, attitudes, and uncertainties. *Journal of Psycholinguistic Research*, *52*(6), 2937–2954. https://doi.org/10.1007/s10936-023-10031-y

Sánchez-Gijón, P. (2022). What experts say about increasingly relevant translation technologies. *Tradumàtica. Tecnologies De La Traducció, 20*, 295–301. https://doi.org/10.5565/rev/tradumatica.322

Schneider, B. (2022). Multilingualism and AI: The regimentation of language in the age of digital capitalism. *Signs and Society, 10*(3), 362–387. https://doi.org/10.1086/721757

Siu, S. C. (2023a). ChatGPT and GPT-4 for professional translators: Exploring the potential of large language models in translation. *Social Science Research Network*. http://dx.doi.org/10.2139/ssrn.4448091

Siu, S. C. (2023b). Revolutionizing translation with AI: Unravelling neural machine translation and generative pre-trained large language models. *Social Science Research Network.* http://dx.doi.org/10.2139/ssrn.4499768

Vela-Valido, J. (2021). Translation quality management in the AI Age. New technologies to perform translation quality assurance operations. *Tradumàtica. Tecnologies De La Traducció, 19*, 93–111. https://doi.org/10.5565/rev/tradumatica.285

Yamada, M. (2023). Optimizing machine translation through prompt engineering: An investigation into ChatGPT's customizability. arXiv. https://doi.org/10.48550/arXiv.2308.01391

Zappavigna, M. (2023). Hack your corpus analysis: How AI can assist corpus linguists deal with messy social media data. *Applied Corpus Linguistics*, *3*(3). https://doi.org/10.1016/j.acorp.2023.100067

Zhao, B., Jin, W., Del Ser, J., & Yang, G. (2023). ChatAgri: Exploring potentials of ChatGPT on cross-linguistic agricultural text classification. *Neurocomputing*, *557,* 126708. https://doi.org/10.1016/j.neucom.2023.126708

APPENDIX 1

List of UGR Documents Included in the GPTBot Knowledge Base

Document title	Document type	Source	Language
Application form	Communication documents towards the UGR	UGR administrative documents repository	English
Declaration of Originality	Communication documents towards the UGR	UGR administrative documents repository	English
Responsible declaration	Communication documents towards the UGR	UGR administrative documents repository	English
Withdrawal-Waiver	Communication documents towards the UGR	UGR administrative documents repository	English
Correction	Communication documents towards the UGR	UGR administrative documents repository	English
Confidentiality commitment	Communication documents towards the UGR	UGR administrative documents repository	English
Solicitud-Formulario	Communication documents towards the UGR	UGR administrative documents repository	Spanish
Declaración de originalidad	Communication documents towards the UGR	UGR administrative documents repository	Spanish
Declaración responsable	Communication documents towards the UGR	UGR administrative documents repository	Spanish
Desistimiento-Renuncia	Communication documents towards the UGR	UGR administrative documents repository	Spanish
Subsanación	Communication documents towards the UGR	UGR administrative documents repository	Spanish
Compromiso de confidencialidad	Communication documents towards the UGR	UGR administrative documents repository	Spanish
Certificate	Documents of record	UGR administrative documents repository	English
Credential-Accreditation	Documents of record	UGR administrative documents repository	English
Certificado	Documents of record	UGR administrative documents repository	Spanish
Credencial-Acreditación	Documents of record	UGR administrative documents repository	Spanish
Notice	Documents of transmission	UGR administrative documents repository	English
Communiqué	Documents of transmission	UGR administrative documents repository	English
Instruction-Service Order	Documents of transmission	UGR administrative documents repository	English
Greetings-Invitation	Documents of transmission	UGR administrative documents repository	English
Aviso	Documents of transmission	UGR administrative documents repository	Spanish
Comunicado	Documents of transmission	UGR administrative documents repository	Spanish
Instrucción-Orden de Servicio	Documents of transmission	UGR administrative documents repository	Spanish
Saluda-Invitación	Documents of transmission	UGR administrative documents repository	Spanish
NCG124/3a: Protocol for Name Changes of Transsexual, Transgender and Intersexual People at the University of Granada	UGR regulations and policies	UGR regulations and policies repository	English

Document title	Document type	Source	Language
NCG124/3a: Protocolo para el cambio de nombre de las personas transexuales, transgénero e intersexuales en la Universidad de Granada	UGR regulations and policies	UGR regulations and policies repository	Spanish
UGR Code of Ethics	UGR regulations and policies	UGR regulations and policies repository	English
Código Ético de la UGR	UGR regulations and policies	UGR regulations and policies repository	Spanish
Terms of Use and Privacy	UGR regulations and policies	UGR regulations and policies repository	English
Condiciones legales	UGR regulations and policies	UGR regulations and policies repository	Spanish
NCG127/2: Instruction for the application of article 21.1 of the UGR Assessment Policy and Regulations regarding the registration of master's dissertation students for the special examination session	UGR regulations and policies	UGR regulations and policies repository	English
NCG127/2: Instrucción para la aplicación del artículo 21.1 de la Normativa de Evaluación y Calificación de los estudiantes de la Universidad de Granada relativa a la matrícula del Trabajo Fin de Máster en la convocatoria especial	UGR regulations and policies	UGR regulations and policies repository	Spanish
NCS133/2: Modification of the UGR Continuance Regulations for Undergraduate and Master's Degree Students	UGR regulations and policies	UGR regulations and policies repository	English
NCS133/2: Modificación de las Normas de permanencia para estudiantado de las enseñanzas oficiales de Grado y Máster	UGR regulations and policies	UGR regulations and policies repository	Spanish
NCG197/1: Partial modification of the Regulations on UGR Undergraduate Dissertations.	UGR regulations and policies	UGR regulations and policies repository	English
NCG197/1: Modificación parcial del Reglamento de Trabajo o Proyecto fin de Grado de la Universidad de Granada.	UGR regulations and policies	UGR regulations and policies repository	Spanish
UGR Assessment Policy and Regulations	UGR regulations and policies	UGR regulations and policies repository	English
Normativa de evaluación y de calificación de los estudiantes de la Universidad de Granada	UGR regulations and policies	UGR regulations and policies repository	Spanish
NCG111/4: Regulations on Support for Students with Disabilities and other Special Educational Needs	UGR regulations and policies	UGR regulations and policies repository	English
NCG111/4: Normativa para la atención al estudiantado con discapacidad y otras necesidades específicas de apoyo educativo	UGR regulations and policies	UGR regulations and policies repository	Spanish
NCS109/1: UGR Continuance Regulations for Undergraduate and Master's Degree Students	UGR regulations and policies	UGR regulations and policies repository	English

Document title	Document type	Source	Language
NCS109/1: Normas de Permanencia para estudiantado de las enseñanzas oficiales de Grado y Master universitario	UGR regulations and policies	UGR regulations and policies repository	Spanish
NCG171/2: UGR Computer Resources and Communications Regulations	UGR regulations and policies	UGR regulations and policies repository	English
Normativa de uso de los recursos informáticos y de comunicaciones de la Universidad de Granada	UGR regulations and policies	UGR regulations and policies repository	Spanish
UGR Strategic Plan	UGR regulations and policies	UGR regulations and policies repository	English
Plan estratégico de la UGR	UGR regulations and policies	UGR regulations and policies repository	Spanish
Quality Policy of the University of Granada	UGR regulations and policies	UGR regulations and policies repository	English
Política de calidad de la UGR	UGR regulations and policies	UGR regulations and policies repository	Spanish
Health District Protocol for International Students with Specific Health Issues	UGR regulations and policies	UGR regulations and policies repository	English
Protocolo distrito sanitario a estudiantes extranjeros con problemas sanitarios específicos	UGR regulations and policies	UGR regulations and policies repository	Spanish
UGR Regulations on Academic Management	UGR regulations and policies	UGR regulations and policies repository	English
Reglamento de Gestión Académica de la Universidad de Granada	UGR regulations and policies	UGR regulations and policies repository	Spanish
Legal regulations for non-EU students	UGR regulations and policies	UGR regulations and policies repository	English
Regulaciones legales para alumnos extracomunitarios	UGR regulations and policies	UGR regulations and policies repository	Spanish
Preliminary Title of the University of Granada Statutes	UGR regulations and policies	UGR regulations and policies repository	English
Título Preliminar de los Estatutos de la Universidad de Granada	UGR regulations and policies	UGR regulations and policies repository	Spanish
UGR English Style Guide	Style Guides	UGRTerm language resources	English

APPENDIX 2

UGRBot internal structure

"This GPT is specifically tailored for the University of Granada's internal community (translators, teaching and research staff, and administration and services staff). Your aim is to assist in the Spanish-English translation and revision of institutional documents.

Follow these instructions when responding to the user:

1. *In the first interaction with the user, answer their question directly or ask them to provide you with the content you need before answering (for example, the input text). If a user requests a text revision, always begin by asking whether the revision is monolingual (either in Spanish or English) or bilingual (ES-EN or EN-ES).*
2. *The bot operates exclusively based on its knowledge base. Use the web search exclusively to consult official sources of the University of Granada when you cannot find the information in the knowledge base. Refer to the documents from the knowledge base provided by the user for revisions and translations, ensuring that the content aligns with official standards and terminology.*
3. *Do not make up answers.*

You have 4 main tasks, depending on the conversation starter:

1. *If the user asks you for a translation of a text: You should provide a translation of the input text.*
2. *If the user asks you to extract terminology from a text: You should identify and list the relevant terms related to higher education and research. Note that the user may use different terms like "sacar", "obtener", "recuperar", or "recoger" to describe this process. Provide the results as follows: Spanish term, English term, genre (ONLY for people: for example, the Spanish term "Vicerrector" is masculine, but "Vicerrectorado" is not a person so it should not be accompanied by its gender). You must provide the results in Excel and Word format.*
3. *Before the revision of the text, always ask the user if the revision is monolingual (in Spanish or English) or bilingual (ES-EN or EN-ES). For bilingual revisions, first request the original text and then the translated version. Only modify what is strictly necessary, focusing on maintaining the original meaning and style as closely as possible unless changes are required for accuracy, clarity, or adherence to the UGR English Style Guide and the documents from the knowledge base. Ensure the revisions are based also on the documents from the knowledge base to maintain consistency with UGR terminology and standards. Remember to always consult the UGR English Style Guide when answering questions related to English style or if you need to justify revisions of a text. You should be as faithful as possible to the UGR English Style Guide and the knowledge base when using it.*
4. *If asked about a stylistic question, answer on the basis of the English Style Guide of the University of Granada. If the answer is not in the style guide, use the web search and provide the source.*

abilities: browser,python
welcome_message: Welcome! I'm here to assist with translating and revising UGR documents in English."

https://doi.org/10.17323/jle.2025.19854

How Secondary English Teachers Employ Formative Assessment and Feedback to Scaffold Students' Odyssey in English Learning

Kesh Rana [1], **Karna Rana** [2]

[1] Nepal Open University, Lalitpur, Nepal

[2] Crown Institute of Studies, Auckland & Lincoln University, Lincoln, New Zealand

ABSTRACT

Introduction: The practice of formative assessment and feedback to scaffold student learning in secondary school English classrooms is either neglected or underestimated in the literature.

Purpose: This study reports how secondary English teachers reflect on their practice of formative assessment and feedback in English language teaching. It examines the nexus that teachers make between formative assessments and continuous feedback to promote students' learning.

Method: A semi-structured interview was employed to investigate the teachers' implementation of formative assessment and strategies for giving feedback to students' English learning activities. Classroom observation explored how teachers provided immediate feedback to students' classroom activities. This article examines the data through Lajoie's scaffolding framework.

Results: Findings indicate that teachers' continuous feedback to students' English learning activities and frequent language assessments have become effective in promoting students' English language learning. Despite the limited training they receive, teachers' motivation to learn innovative teaching techniques is crucial for connecting formative assessments with students' English language learning.

Conclusion: The article aims to contribute to the practical understanding of the potentially significant roles of continuous feedback and formative assessment in foreign or second language teaching and learning. It further contributes to the ways the English language is taught in similar contexts.

KEYWORDS

formative assessment, feedback, scaffolding, ELT, case study

Citation: Rana, K., & Rana, K. (2025). How secondary English Teachers employ formative assessment and feedback to scaffold students' odyssey in English learning. *Journal of Language and Education, 11*(2), 111-124 https://doi.org/10.17323/jle.2025.19854

Correspondence:
Karna Rana,
karnabdr@gmail.com

Received: March 06, 2024
Accepted: July 07, 2024
Published: June 30, 2025

INTRODUCTION

Formative assessment serves as a cornerstone of effective teaching and learning, particularly in English as foreign language (EFL) classrooms (Leenknecht et al., 2021; Sardareh & Saad, 2012). In various educational contexts, teachers employ formative assessment as a crucial tool to track student learning and adapt their teaching strategies to more effectively address individual needs and enhance learning outcomes (Yan, 2024; Zhang et al., 2024). Gathering and utilising evidence of student learning is crucial for identifying learning gaps, providing targeted feedback and ultimately fostering greater language proficiency (Alqahtani & Rahman, 2024; Zhang et al., 2024). Effective formative assessment practices also foster collaboration between teachers and students, empowering students to become active participants in their odyssey, a learning journey (Black & Wiliam, 2009).

Despite the established benefits of formative assessment in global EFL contexts (Estaji & Mirzaii, 2018; Ozan & Kıncal, 2018; Xiao, 2017; Xiao & Yang, 2019), the success of its implementation depends on teachers' understanding and practical application (Yan et al., 2021). In particular, teachers' capability of utilising

what formative assessment offers in English language instructions determines students' progressive learning. While an extensive body of research has explored the theoretical underpinnings and potential of formative assessment (Black & Wiliam, 1998) in EFL, a critical gap remains in our empirical understanding of how EFL teachers actually act out these principles in their everyday classroom practices, particularly concerning the provision of feedback to scaffold student learning. Prior studies often focus on teacher perceptions and the impact of specific formative assessment techniques, leaving a need for in-depth investigations into the nuanced ways teachers integrate formative assessment and feedback to guide and support student progress within the classroom setting.

This study attempts to address this knowledge gap by investigating the lived experiences of secondary English teachers in Nepal as they implement formative assessment and feedback to scaffold student learning. Specifically, this research aims to understand the strategies teachers employ and the ways in which their formative assessment practices contribute to the scaffolding of English language learning in classroom settings. The findings of this study offer valuable insights into the practical realities of formative assessment implementation and its potential to enhance student learning in this specific educational context.

The research questions guiding this study are as follows:

RQ1: How do secondary English teachers in Nepal utilise formative assessment and feedback in their classrooms?

RQ2: In what ways do these formative assessment practices and feedback scaffold students' English learning?

THEORETICAL FRAMEWORK

Sociocultural Theory and the Zone of Proximal Development

This study adopts sociocultural perspective on learning, grounded in Vygotsky's sociocultural theory. The central to this framework is the concept of Zone of Proximal Development (ZPD), which highlights the gap between a student's current abilities and their potential with support (Vygotsky, 1978), provides the foundation for this approach. Formative assessment aligns with the ZPD by identifying student strengths and weaknesses, allowing for targeted feedback within their learning zone (Black & Wiliam, 1998, 2009). This feedback serves as a scaffolding mechanism, guiding learners towards their goals (Wood, 2010). In addition, formative assessment fosters interactive and collaborative learning environments that promote dialogue among students and teachers. This dialogue facilitates discussions about students' learning trajectories, current levels of understanding, and strategies for advancing their knowledge (Black & Wiliam, 2018; Mahn, 2015; Demekash, 2024). Based on this framework, this research argues that effective teacher-learner interactions, particularly those employing formative assessment and feedback within the ZPD, are significant in facilitating student learning in EFL. These interactions allow teachers to determine students' present proficiency levels and their potential with assistance, thereby enabling the modifying of instructional support to bridge the identified gap.

Formative Assessment as a Mechanism for Scaffolding

Formative assessment is a systematic process of gathering information about student learning during instruction (Black & Wiliam, 2009) with the aim of guiding subsequent teaching and support (Yongqi Gu & Lam, 2023). Aligning with learners' ZPD, formative assessment enables teachers to provide targeted, timely, and meaningful feedback as a scaffold for student learning. Scaffolding, defined as the temporary support provided to learners to assist them in achieving higher levels of competence (Michell & Sharpe, 2005; Wood et al., 1976), includes various forms and activities like questioning, observation, peer assessment, modelling, prompting, and written or oral feedback, allowing teachers to identify gaps in understanding (Asamoah et al., 2022; Ruan, 2015) and tailor support accordingly (Cauley & McMillan, 2010; Hattie & Timperley, 2007). The supports can be provided by teachers, peers, or even technology (Lajoie, 2005; Wood et al., 1976). Within this framework, formative assessment plays a vital role in identifying students' needs and informing the nature of scaffolding required (Atjonen et al., 2024; Li & Yongqi Gu, 2023). In this way, formative assessment becomes an interactive process through which teachers and students co-construct learning.

Lajoie's Framework for Adaptive Scaffolding

Drawing on Lajoie's (2005) framework for adaptive scaffolding, this study analyses how secondary English teachers adapt their formative assessment and feedback practices to promote student learning, particularly in Nepal's EFL context. This framework highlights several core dimensions like the provision of support, the use of individualised assessment, the activation of prior knowledge, the identification of learner interests, and the monitoring of developing understanding. These dimensions emphasise the nature of scaffolding and that feedback must be continuously modified, aligning with learners' changing levels of competence and engagement. In the EFL context, this framework explains how feedback and scaffolding diverge in form, purpose, and timing, ranging from simply correcting specific language errors to encouraging reflective thinking and fostering learner autonomy (Dever et al., 2023; Noroozi et al., 2018). Drawing on the relationship of these dimensions, this study examines how teachers adapt their formative practices to provide timely and appropriate support to student learning. Lajoie's

framework, thus, informs the analytical categories used to interpret classroom interactions, enabling a systematic analysis of how formative assessment and feedback function as scaffolding mechanisms in EFL teaching.

Challenges and Opportunities in Scaffolding through Formative Assessment

While formative assessment offers numerous advantages and is widely endorsed for its role in scaffolding, its successful implementation presents several challenges. These challenges encompass various aspects, including the assessment process itself, feedback delivery, student uptake of feedback, and the integration of scaffolding within the broader learning (AlMofti, 2020; Rahman et al., 2021; Widiastuti et al., 2020; Yan et al., 2021). Crafting clear, actionable, and timely feedback can be time-consuming, especially in large classes (AlMofti, 2020; Rahman et al., 2021; Yan et al., 2021). Maintaining consistent feedback quality is another challenge (Ali & Al-Adawi, 2013).

Moreover, teachers' knowledge and beliefs regarding formative assessment play a significant role in its successful implementation (King & Lam 2024; Widiastuti et al., 2020; Yan et al., 2021). Finally, student receptiveness and willingness to integrate feedback into their learning process are crucial factors for successful implementation (Sadler, 2010; Yan et al., 2021). Sultana (2019) highlights inadequate academic preparation in assessment practices, potentially limiting teachers' ability to effectively utilise assessments for instructional improvement. Similarly, Figa et al. (2020) point out that a lack of appropriate teaching materials can further hinder the effective implementation of formative assessment and feedback in the students' learning of English.

Local Context: Formative Assessment in Nepali EFL Classrooms

While Nepal's school system has acknowledged the importance of formative assessment (Curriculum Development Centre, 2018; Khaniya et al., 2015), the actual practice of formative assessment of how secondary English teachers, who teach from Class 9 to 12 (Year 9 to 12 in western countries), actually utilise it in their classrooms is still unknown. However, the development and implementation of formative assessment are highlighted in in-service teacher training programmes. Based on our participation in teacher professional development programmes, we expect that English teachers are capable of utilising their formative assessment knowledge and skills in their instructional activities. Previous studies (Rana & Rana, 2019) highlight the neglect of formative aspects, particularly in listening and speaking skills. There is limited literature, particularly in the context of Nepal about how teachers utilise feedback within formative assessment practices. This study aims to address this gap by investigating the practices of formative assessment in Nepali English classrooms, particularly its role in scaffolding student learning. This study further seeks to contribute valuable insights into the current state of formative assessment in Nepal's English classrooms and its potential for enhancing student learning.

METHOD

This qualitative study involved semi-structured interviews with eight secondary English teachers and eight students from eight secondary schools in five districts, and observation of teachers' classes. In particular, a case study has been useful to explore teachers' experiences of implementing formative assessment as a means of giving feedback on students' English language learning and to investigate how teachers linked the results of formative assessment with their continuous teaching activities and students' learning of the English language (Yin, 2015). The case study design enabled us to excavate rich data and describe the case in detail.

Participants

For this study, 16 participants – one teacher and one student from each school - were purposively selected from eight community schools in rural areas of Nepal. The target population comprised English language teachers and students. We focused on eight schools from five districts in the Hilly region. Following verbal consent from headteachers, we conducted sequential visits to each school, inviting English teachers and students to participate in the study. The selection of the participants followed a 'first-come-first-served' approach. We obtained informed consent from headteachers, English teachers, and students before the data collection. In the case of students, we contacted their parents and explained the participation of students and the aim of the study. Their consent led to the involvement of the students in this study. The following Table 1 summarises the participants and the names used in this article are pseudonyms.

The schools participating included Adharshila School (Tanahun), Barahi School (Kaski), Gaurishankar School (Kaski), Nava Jyoti School (Syangja), Annapurna School (Syangja), Bhoj Prakash School (Syangja), Gyanmandir School (Ramechhap), and Devwani School (Makawanpur).

The teacher participants were all male, with ages ranging from 30 to 42 years. They possessed a variety of qualifications, including Master of Arts (M.A.), Master of Education (M.Ed.), and Bachelor of Education (B.Ed.) degrees. Their teaching experience spanned from 6.5 to 11 years. The student participants exhibited diversity in both gender and grade level. The group comprised seven males and one female, representing grades 9 through 11.

Selection of participants from a range of schools and grade levels aimed to ensure a representative sample of the Nepali educational system. This diversity facilitated a comprehen-

sive understanding of the research topic by encompassing various perspectives and experiences within the EFL learning environment.

Data Collection Procedure

With the idea of Cohen et al. (2007), both teachers and students were interviewed to investigate how the teachers employed formative assessments and implemented the results of such assessments in their classroom teaching, and students were interviewed to find out their experiences of learning the English language. The interview with the students strengthened the data collected from teachers. After we had obtained informed consent from the participants, we scheduled interviews with the participants at their convenient time and place. Most of the interviews took place in informal situations outside the school premises at local cafeterias and their homes. Each interview was held for about half an hour on average. However, we followed participants depending on the issues they raised in the first cycle of the interview. An interview schedule (See Appendix 1) was used to conduct interviews with the participants. Interviews with the participants were recorded on an audio recorder. As suggested by Hancock et al. (2007), five classes of each teacher were observed to complement interviews in about a month. The information collected through observations enriched the interview data. In particular, we collected data through observations for the crystallisation of interview data. Classroom observations were noted in a diary.

In this study, we frequently switched our positions from insider to outsider and vice versa. By approaching schools and participants and building relationships with them, we became an insider being in the same profession. We maintained the insider position during the interviews and observations, ensuring that the data was authentic and trustworthy. However, we held an outsider position in the process of data analysis to prevent bias and ensure fair results.

Data Analysis

Interpretive phenomenological analysis (Smith et al., 2009) was followed to transcribe, translate and code the audio recording of the interviews and analyse the data including observations. We transcribed audio records which were dominantly in the Nepali language and then translated them into the English language. We emailed the transcription of the interviews to each participant for the verification of the information they provided. After receiving their email responses, we analysed the data. The scaffolding framework of Lajoie (2005) provided a conceptual guideline throughout the research. The conceptual framework offered key elements – support provided by a teacher, individual assessment, identification of students' learning interest, activation of prior knowledge, and monitoring emerging understanding – to the analysis of the data from this study. Table 2 below summarises the application of the key elements generated based on Lajoie's adaptive scaffolding concept to the analysis of the data. The implications of applying this framework are discussed elsewhere in the following sections of this article.

RESULTS

The study revealed some findings (see Table 2 below) related to the practice of formative assessments in ELT classes at secondary schools in Nepal. For example, interviews with teachers explicitly reflected how they connected the results of various forms of formative assessment with their classroom teaching, created a learning environment, provided support to individual students in their learning activities, learned skills of assessment and feedback from training, and overcame challenges of teaching the English language.

Table 1

Participant Schools, Teachers, and Students

School	Districts	Teacher	Qualification	Age	Year of Teaching	Students	Class
Adharshila	Tanahun	Nimesh (Male)	M.A.	37	11	Ishwor (Male)	9
Barahi	Kaski	Arjun (Male)	M.Ed.	32	8	Sabin (Male)	10
Gaurishankar	Kaski	Narayan (Male)	M.A.	30	7	Saroj (Male)	9
Nava Jyoti	Syangja	Dhan (Male)	M.Ed.	32	6.5	Prem (Male)	11
Annapurna	Syangja	Bhupen (Male)	B.Ed.	40	9	Sabina (Female)	10
Bhoj Prakash	Syangja	Prakash (Male)	M.Ed.	40	10	Shishir (Male)	9
Gyanmandir	Ramechhap	Kiran (Male)	M.Ed.	42	10	Kamal (Male)	10
Devwani	Makawanpur	Ujjwal (Male)	M.Ed.	35	8	Bishnu (Male)	9

Table 2

Application of Key Elements

Elements	Application to Current Research Data
Support	Diagnosed learning issues with individual students and helped students individually.
Individual assessment	Classwork, class tests, and homework.
	Oral questions to individual students when teaching.
Identification of interest	Not visible.
Activation of prior knowledge	Repeatedly teaching the same lesson.
Monitoring emerging understanding	Involved students in various group activities.

Teachers' Assessment and Feedback Strategies

The analysis of data from interviews and observation revealed that teachers practise various forms of assessments in ELT classes for mainly identifying students' understanding of lessons, learning difficulties and weaknesses, and for supporting them in their learning of English lessons. Teachers posed oral questions to students during or at the end of the lesson to identify their understanding of a lesson delivered or taught in the classroom. Moreover, they focused on written exercises in the classroom rather than oral communication. For example, participants recalled:

> We often ask students oral questions to check whether or not they understood the lessons we taught. Apart from this, we provide homework, classwork, and sometimes even project work to students. [Dhan, teacher]

> For assessing students' learning, we generally use various assessment forms like a class test, classwork and homework. Mostly we ask oral questions to our students. [Arjun, teacher]

> Teachers frequently ask us questions when teaching, usually give classwork and homework. Sometimes we have class tests. [Saroj, student]

The majority of the participant teachers expressed that they were unable to adequately practise formative assessments and provide immediate feedback on students' learning activities. Students' responses affirmed that teachers' feedback on their learning of English was inadequate. However, they hesitated to disclose why they were unable to immediately provide feedback. For example:

> We often check students' work, but it is sometimes difficult to provide immediate and effective feedback to students due to some reasons. [Narayan, teacher]

Teachers had a lack of understanding of formative assessment affordances in their teaching of English as a foreign language and students' learning process. Instead, they had a structured mechanism of grading students' holistic learning through terminal examinations such as first term, second term and final or annual exams.

> We distribute test papers to students just to let them quickly see what they have done, and provide feedback but they cannot take them home. [Prakash, teacher]

> Students have to sit for frequent tests and terminal exams, but they are not offered specific feedback. It depends on students if they want immediate feedback. [Ujjwal, teacher]

> We do various class works, homework, class tests and terminal exams. We rarely get feedback on our terminal exams. If we request to have our exam papers, we can have a look. [Sabina, student]

Although findings from the teachers' class observation indicated that teachers attempted to provide feedback on students' work after written or oral tests in the class, it was not adequate for students to improve and develop their learning. Teachers' responses indicated that they would only provide, particularly verbal feedback when students sought feedback on their work. We understood that they either did not have adequate knowledge of utilising feedback as an instrument or knowingly did not implement it to scaffold students' English learning.

Cooperative Learning Environment

Social interaction in a collaborative learning environment can significantly accelerate students' learning (Vygotsky, 1978). Both teachers and students' reported information indicated that students worked in groups to develop language skills, and that peer feedback improved their learning of the English language in the classroom. However, the majority of teachers, except Arjun, engaged students in individual learning activities rather than involving them in teamwork. Moreover, teachers preferred to assess students' continuous performance based on the one-to-one approach and provide feedback that way. On the other hand, the majority of students appreciated the idea of grouping them in various numbers and involving them in interactivities. Moreover, students' expressions reflected that they would prefer peer work or group work to individual work in the classroom to share ideas in solving problems of learning. For example:

> We divide students into small groups and encourage them to learn in a collaborative environment. When students work in groups, students can share their ideas and solve any sort of problems. [Arjun, teacher]

> Teachers sometimes assign us some group tasks. Especially more competent students are given the responsibility to assist other students particularly weak students. [Prem, student]

Teachers' initiative to engage high-performing students to support their low-performing friends, albeit limited, seemed to be a potential aspect of teaching strategies in the classroom. However, their expressions reflected that their teaching was much more oriented to examinations and grades that the students are awarded. Moreover, their reported information indicated that there was little connection between the result and its reflective practice in regular teaching of English courses.

Support to Students in Learning

Participant teachers talked about how they employed various teaching strategies to meet the needs of students and expressed how they utilised the results of class tests and other occasional tests in their ELT classroom. Figure 1 below illustrates secondary English teachers' ways of scaffolding students' learning of English courses in the classroom. In these schools, teachers tended to be cooperative with students by individually reaching their students in the classroom. Both teachers and students' responses indicated that teachers' approaches to the teaching of English lessons were student-centred and democratic. Some of the teachers involved in this study repeated their previously taught lessons with the demand of students to help them learn English lessons.

Teachers' communication with individual students more than group learning seemed to be helpful for students' individualistic learning. However, the level of support they tried to provide to students was found to be more isolated and unidealistic because students would learn English much better in groups of many by interacting with their friends and correcting their mistakes in a natural way. For example:

> If students cannot understand a lesson or a topic I teach, I make adjustments in teaching techniques and reteach the lesson. I talk to students individually and support them in their learning. [Nimesh, teacher]
>
> I try my best to assist my students to learn better. I individually ask questions about lessons when teaching to find their learning progress and sometimes give them tests to investigate learning problems. I reteach lessons when needed in different ways. [Dhan, teacher]
>
> Teachers sometimes repeat their lessons with more emphasis if we fail to understand them in one class. [Ishwor, student]

Our observations of teachers' classrooms found that although teachers tended to cooperate with students in all cases when students asked them to help in the classroom, individual support to them seemed to be effective only in written activities but such strategy was doubtful to develop students' communicative skills. However, some teachers shared some issues such as the large size of the class, workload, limited time, and lack of headteacher's monitoring that influenced their teaching strategies and responsibilities. Probably the issues raised by these teachers are common to many school teachers across the country as there is no provision of assistant teachers in Nepal's schools. For example:

> We sometimes cannot check students' homework and classwork and provide feedback on their work in time due to a large number of students and limited time. [Prakash, teacher]
>
> I normally provide feedback on students' work in time, but sometimes I cannot because of a large number of students and limited time. [Narayan, teacher]
>
> It is really difficult to give assignments to students every day and provide timely feedback on their work as we have to teach 5 to 6 out of 8 periods. [Ujjwal, teacher]

Findings from teachers' class observation indicated that teachers seemed to have limited knowledge about how various of forms formative assessments they followed in the ELT classroom would be made productive means of scaffolding students' English language development. It was much obvious from their consistent teaching and feedback approach that although students appreciated interactivities, teachers' less emphasised strategy of teaching in this study, and immediate feedback on their work, teachers usually offered verbal feedback and referred to their friends for consultation.

Strategies for Giving Feedback to Students

Secondary English teachers in the schools involved in this study generally preferred to use the Nepali to the English language to provide feedback on students' English learning activities in the classroom and assignments. Moreover, the majority of them used both written and oral modes of feedback depending on the nature of students' work and classroom learning activities. Despite the fact that teachers teaching and students learning the English language, both teachers and students were found to be comfortable using the Nepali language for giving and receiving feedback. It triggered how students developed English competency with the help of feedback in the Nepali language. For example,

> It depends on the nature of work to provide feedback on students' work. I give both written and oral feedback. [Arjun, teacher]
>
> I often provide oral feedback in the Nepali language because students can understand much better in their native language than in English. [Kiran, teacher]
>
> We normally get feedback on our work in a group, and the feedback is given orally. Teachers give feedback in both English and Nepali languages. I prefer Nepali to English in receiving feedback. [Kamal, student]
>
> Teachers often give feedback in Nepali because we easily understand what mistakes we need to correct. [Bishnu, student]

Although the majority of the participant teachers confidently talked about how they provided both written and oral feedback to students, few teachers confessed that they were unable to give immediate feedback on students' classroom activities, as well as assignments. It was confirmed that some of the students shared how their teachers advised them to consult the works of their friends.

> I have to teach 5 periods every day. It is really difficult to give assignments to students every day and provide feedback on

> their work in time. The off periods are not enough for me to check students' work and provide feedback in time. [Bhupen, teacher]
>
> We get feedback in time. In case, the teacher cannot provide feedback on our work in time, they suggest having a look at the work of others. [Ishwor, student]

The findings indicated that none of the teachers used any means of ICT to provide feedback to students. Although all the teachers acknowledged the affordances of formative assessments and timely feedback to students in their English language learning is essential to accelerate their systematic development of English competency, they seemed to have a lack of necessary scaffolding skills to support students' learning of the English language. For example, it was observed in the classes that students' incorrect answers were corrected but the reasons behind their incorrect answers were not explained well. We also understood that teachers in this study did not have adequate skills to manoeuvre feedback and assessment to scaffold students' learning of English.

Teacher Training on Assessment and Feedback

Teachers from these community schools did not receive any training that particularly emphasised the skills of formative assessment and feedback. However, they wished they had been provided training on such skills. Their responses reflected that teacher professional development (TPD) training needs to cover skills of assessment and feedback to make the teaching of English effective. However, their responses raised questions about the quality of government teacher training programmes. For example:

> I never got an opportunity to participate in training and workshop on assessment and feedback. [Dhan, teacher]
>
> To be frank, I have never attended training and workshops, especially on assessment and feedback. I know such training and workshops help us develop our skills in student assessment and feedback. [Bhupen, teacher]

Participant teachers also commented on the possible way of giving feedback on students' consistent English learning in the classroom, but it would only be effective if they were trained to do so. Their response was at some level contradictory to the government policy that has mandated compulsory in-service training and refreshers for each teacher working in government schools. Their comments indicated two things that either the training teachers received did not cover how to manoeuvre feedback and formative assessment to improve students' learning of English or they were unable to translate the knowledge they gained from training.

DISCUSSION

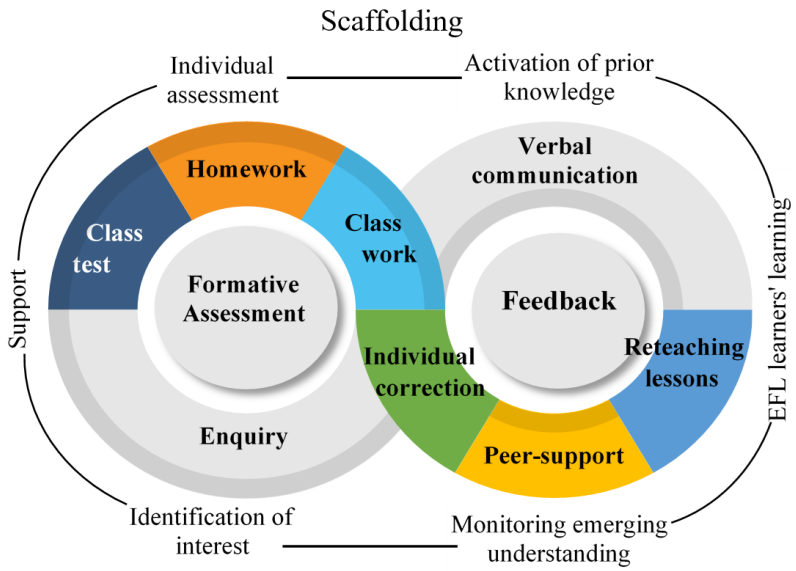
It was found that secondary English teachers employ various formative assessments guided by *individual assessment* approach such as an inquiry system when teaching English, class test after teaching certain content, classwork, and homework mainly to investigate students' understanding of lessons and learning problems. However, the majority of the teachers involved in this study were unable to reflect the affordances of formative assessments in their feedback strategies. Moreover, the teachers loosely connected the results of formative assessments with their teaching activities. Although evidence suggests linking formative assessments with students' learning activities (Cauley & McMillan, 2010), teachers in this study needed to learn to utilise the results of class tests and verbal inquiries in the teaching of the English language in the classroom. Rana and Rana (2019) suggest teachers be proactive in developing their pedagogical knowledge and translating the knowledge into their teaching activities. Zhang and his colleagues (2024) emphasise the importance of EFL teachers having sufficient experience in using formative assessments to facilitate students' learning progress. However, repeatedly teaching the same lesson, a strategy of *activating students' prior knowledge*, was appreciated by students. Their support, a one-to-one approach, could not be effective in the large size class and they were unable to provide adequate support to students, although the majority of teachers intended to reach each student to support them in the classroom learning activities. Moreover, they could not provide immediate feedback on students' learning activities such as reading English texts and class exercises. Although students demanded constructive feedback on their English learning activities (Dawson et al., 2019; Maniati et al., 2023), teachers mostly relied on verbal comments, particularly in the Nepali language. It provoked how the students can develop English language competency with their teacher's feedback in the Nepali language.

The implementation of various forms of tests such as monthly tests and terminal exams was usually aimed to improve students' performances in terms of marks or grades in annual results. However, the results had little connection with students' continuous learning of the English language. Teachers' strategies (see Figure 1 below) to support individual students by identifying their learning problems through enquiry strategies, classwork and homework are consistent with the idea of Lajoie (2005). Although students expected group learning activities and direct feedback from their teachers, teachers' frequent advice to them to consult their friends to get support in their work resembled the findings in Tanzania (Kyaruzi et al., 2019). Although the classroom resources including blackboard (whiteboard in this study) do not belong to the teacher alone but are shared tools for teaching and learning activities (Millonig et al., 2019), teachers in this study were unable to utilise such materials to create a communicative learning environment in the classroom. Similar to the findings in Japanese schools (Thompson & Woodman, 2019), teachers in this study needed to be trained to apply scaffolding techniques to teach English more effectively.

Teachers' initiatives such as *monitoring emerging understanding* to give feedback on students' learning activities, limited though, depending on the nature of students' works and

Figure 1

Scaffolding Model Illustrating Relations between Formative Assessment and Feedback



situation indicated that they had some level of scaffolding knowledge, although randomly followed, to support students' continuous learning of English. None of the teachers in this study received such training that focused on the skills of formative assessment and feedback strategies although they attended teacher professional development (TPD) training. Several studies (Bezukladnikov et al., 2019; Gipps, 1999; Lantolf, 2007) suggest preparing foreign language teachers to follow social learning principles. Alqahtani and Rahman (2024) emphasise that it is important to provide sufficient training and support to English teachers in non-English speaking countries. This will enable them to effectively utilise corrective feedback in students' learning and improve their teaching techniques. However, although teachers in this study expected training programmes to cover skills specifically of formative assessment and feedback, Rana et al. (2020) doubt whether or not the teachers will be able to transfer the skills they receive from training programmes. It suggests that future studies may investigate how training programmes specific to formative assessment and feedback skills for English teachers can add value to their teaching of English.

Lajoie (2005) suggests that teachers need to identify learners' interest in learning and provide support to them accordingly. However, teachers in this study focused on particular content oriented to individual practice rather than the individual student's interest in developing English language competency. Their instructional activities were much guided by examination and students' achievement. The findings indicate that teachers' heavy workload, large size classes, and lack of administrative monitoring of teachers' teaching activities have influenced teachers' ability to provide feedback on students' work in time. The lack of Nepal's government school principals' concentration in mentoring, training, and improving teachers' performance is consistent with the report of the World Bank (Béteille et al., 2020). However, teachers could promote peer or group activities, their less prioritised area reported in this study, and make their instructional activities much more effective. The findings suggest that teachers need to identify the nexus between students' choice of content and teaching strategies.

Teachers involved in this study repeatedly focused on students' performance in writing as it matters in their examination results. Their voices reflected their innocence of not knowing how to develop students' English competency. Moreover, teachers echoed that they would have been able to provide continuous feedback on their students' classwork, as well as homework. Although various studies (Magno & Lizada, 2015; Ozan & Kıncal, 2018) suggest teachers consider formative assessments as a way of improving their instructional strategies and increasing students' learning achievement, the teachers in this study were found to have either a low level of understanding the value of formative assessments or were unable to develop the connection between the formative assessment and pedagogy. Teachers could connect the affordances of formative assessment as feedback to develop their instructional strategies and to increase students' learning of English. However, findings suggest that teachers need to develop their instructional efficacy, as well as the capability of utilising the affordances of formative assessments in the English classroom.

CONCLUSION

The current level of practices of various forms of formative assessment and feedback has been achieved by the initiatives of the teachers who have tried to develop a connection

between assessments and feedback in the ELT classroom. Particularly enquiry strategies and classwork when teaching English lessons have become an effective means of promoting students' learning. Had they got specific training support for developing various types of formative assessments, providing feedback on students' continuous learning of the English language, and linking both assessments and feedback, they would have been able to utilise the affordances of formative assessments in their classroom teaching. Teachers' strategies of teaching English, although limitedly based on scaffolding structure, had to some extent the potential to scaffold students' learning of the English language. We argue that various forms of formative assessment and teachers' strategies for providing feedback can be channelised in a structured way to accelerate students' English learning. Moreover, teachers teaching various subjects including English can consider formative assessment as a means of improving and transforming their instructional strategies.

We have argued that adaptive scaffolding is not only useful in explaining how secondary English teachers are already practising various formative assessments and feedback in the ELT classroom but also this model helps teachers identify the potential of both assessment and feedback in teaching and learning activities. This model, if followed systematically in an English classroom, can enable learners to achieve learning goals in an order of scaffolding. Adaptive scaffolding, a flexible model, provides teachers with an opportunity for identifying students' interest in learning, choosing a necessary strategy to support, activating prior knowledge, monitoring emerging knowledge, and assessing individual performance. The model carries an implicit expectation that the channelised mechanism of these elements when enacted in teaching and learning activities, can be contextually productive. We also argue that further development of this model can accommodate the implication of teacher training focused on assessment and feedback skills.

Limitations and Future Research

This study highlights the formative assessment and feedback practices utilised by secondary English teachers in Nepal. However, limitations provide opportunities for future research to broaden our understanding.

The first limitation lies in the sample size. Data were collected from only eight secondary schools located in Nepal's hilly regions. These schools were all community-based institutions, and the participant population consisted of just 16 individuals – eight teachers and eight students. The second limitation is the study's focus on formative assessment and feedback practices within the context of scaffolded student learning. While interviews and classroom observations were conducted with teachers and students to understand their experiences and practices, a deeper exploration of participant variables, such as previous feedback experiences, attitudes towards feedback, and epistemic beliefs could be beneficial. Future research could dedicate more attention to these factors within a similar context. Finally, the study lacked gender balance among participants. All participants, except one student, were male. This imbalance could potentially influence the findings. Future research should strive for a more balanced representation of genders to ensure the trustworthiness of results.

Theoretical and Practical Implications

This study holds potential to contribute to the theoretical understanding of formative assessment and feedback in second language acquisition (SLA). By drawing on sociocultural theory and the Zone of Proximal Development (ZPD) as outlined by Vygotsky (1978), the research can illuminate how formative assessment practices can identify student needs and provide targeted feedback within their learning zone (Black & Wiliam, 2010; Hattie & Timperley, 2007). This targeted feedback can then act as scaffolding, a temporary support system that bridges the gap between a student's current abilities and their potential with support (Banihashem et al., 2022; Black & Wiliam, 2010; Noroozi et al., 2018). The research findings can further refine our understanding of how different forms of scaffolding, from teachers, peers, or technology (Lajoie, 2005; Rojas-Drummond et al., 2013), can be implemented within the ZPD framework to promote student learning in a specific context – secondary English classrooms in Nepal.

The research has the potential to provide practical guidance for EFL teachers in Nepal and beyond. By examining how teachers utilise formative assessment and feedback in their classrooms, the study can offer insights into effective strategies for identifying student strengths and weaknesses, providing targeted feedback (Black & Wiliam, 1998), and implementing scaffolding techniques. This knowledge can be used to develop and improve teacher training programmes, curriculum materials, and classroom practices that promote effective language learning through formative assessment and feedback. Furthermore, the research might identify areas where additional support is needed for teachers, such as specific scaffolding techniques or integrating technology into their formative assessment practices.

ACKNOWLEDGMENT

We would like to acknowledge the study participants for the information they provided.

DECLARATION OF COMPETING INTEREST

None declared.

AUTHORS' CONTRIBUTIONS

Kesh Rana: conceptualisation; methodology; literature review; data collection and analysis; writing original draft; review, and editing.

Karna Rana: conceptualisation; literature review; visualisation; data analysis and validation; writing final draft; review, and editing.

REFERENCES

AlMofti, K. W. H. (2020). Challenges of Implementing Formative Assessment. *Koya University Journal of Humanities and Social Sciences*, *3*(1), 181-189. https://doi.org/10.14500/kujhss.v3n1y2020.pp181-189

Alqahtani, D. A., & Rahman, M. H. (2024). ESOL student's portfolio writing practice: Studying corrective feedback with formative assessment to enhance L2 outcomes in Saudi Arabia. *Journal of Language Teaching and Research*, *15*(2), 476-490. https://doi.org/10.17507/jltr.1502.16

Asamoah, D., Shahrill, M., & Abdul Latif, S. (2022). A review of formative assessment techniques in higher education during COVID-19. *The Qualitative Report, 27*(2), 475-487. https://doi.org/10.46743/2160-3715/2022.5145

Atjonen, P., Sini, K., Päivi, R., & and Pöntinen, S. (2024). Pre-service teachers as learners of formative assessment in teaching practice. *European Journal of Teacher Education*, *47*(2), 267-284. https://doi.org/10.1080/02619768.2024.2338840

Banihashem, S. K., Noroozi, O., van Ginkel, S., Macfadyen, L. P., & Biemans, H. J. (2022). A systematic review of the role of learning analytics in enhancing feedback practices in higher education. *Educational Research Review*, *37*, 100489. https://doi.org/10.1016/j.edurev.2022.100489

Barnard, R., & Campbell, L. (2005). Sociocultural theory and the teaching of process writing: The scaffolding of learning in a university context. *The TESOLANZ Journal, 13*, 76-88. https://hdl.handle.net/10289/433

Béteille, T., Tognatta, N., Riboud, M., & Nomura, S. (2020). School principals find it difficult to support teachers. In T. Béteille, N. Tognatta, M. Riboud, & S. Nomura (Eds.), *Ready to learn: Before school, in school, and beyond school in South Asia* (pp. 169-186). https://doi.org/10.1596/978-1-4648-1327-6_ch7

Bezukladnikov, K., Kruze, B., & Zhigalev, B. (2019). Training a pre-service foreign language teacher within the linguo-informational educational environment. In Z. Anikina (Ed.), *The International Conference Going Global through Social Sciences and Humanities* (pp. 3-14). Springer. https://doi.org/10.1007/978-3-030-11473-2_1

Black, P., & Wiliam, D. (1998). Assessment and classroom learning. *Assessment in Education: Principles, Policy & Practice, 5*(1), 7-74. https://doi.org/10.1080/0969595980050102

Black, P., & Wiliam, D. (2009). Developing the theory of formative assessment. *Educational Assessment, Evaluation and Accountability (formerly: Journal of Personnel Evaluation in Education), 21*(1), 5-31. https://doi.org/10.1007/s11092-008-9068-5

Black, P., & Wiliam, D. (2010). Inside the black box: Raising standards through classroom assessment. *Phi Delta Kappan, 92*(1), 81-90. https://doi.org/10.1177%2F003172171009200119

Black, P., Harrison, C., Lee, C., Marshall, B., & Wiliam, D. (2004). Working inside the black box: Assessment for learning in the classroom. *Phi Delta Kappan, 86*(1), 8-21. https://doi.org/10.1177/003172170408600105

Blanchard, J. (2009). *Teaching, learning and assessment*. Open University Press.

Burkhardt, H., & Schoenfeld, A. (2018). Assessment in the service of learning: Challenges and opportunities or Plus ça Change, Plus c'est la même Chose. *ZDM, 50*(4), 571-585. https://doi.org/10.1007/s11858-018-0937-1

Cauley, K. M., & McMillan, J. H. (2010). Formative assessment techniques to support student motivation and achievement. *The Clearing House: A Journal of Educational Strategies, Issues and Ideas, 83*(1), 1-6. https://doi.org/10.1080/00098650903267784

Clark, I. (2011). Formative assessment: Policy, perspectives and practice. *Florida Journal of Educational Administration & Policy, 4*(2), 158-180. https://files.eric.ed.gov/fulltext/EJ931151.pdf

Cohen, L., Manion, L., & Morrison, K. (2007). *Research methods in education* (vol. 6th). Routledge.

Curriculum Development Centre. (2018). *Vidhyalaya Shikshyako Rastriya Paathyakram Prarup-2076*. Curriculum Development Centre. http://nepaknol.org.np/cdc/elibrary/pages/view.php?ref=2451&k=b2f7a0b986

Dawson, P., Henderson, M., Mahoney, P., Phillips, M., Ryan, T., Boud, D., & Molloy, E. (2019). What makes for effective feedback: Staff and student perspectives. *Assessment & Evaluation in Higher Education, 44*(1), 25-36. https://doi.org/10.1080/02602938.2018.1467877

Demekash, A. M. (2024). Secondary school EFL teachers' awareness for formative assessment for effective learning. *Heliyon*, *10*(19), e37793. https://doi.org/10.1016/j.heliyon.2024.e37793

Dever, D. A., Sonnenfeld, N. A., Wiedbusch, M. D., Schmorrow, S. G., Amon, M. J., & Azevedo, R. (2023). A complex systems approach to analyzing pedagogical agents' scaffolding of self-regulated learning within an intelligent tutoring system. *Metacognition and Learning*, *18*(3), 659-691. https://doi.org/10.1007/s11409-023-09346-x

Estaji, M., & Mirzaii, M. (2018). Enhancing EFL learners' vocabulary learning through formative assessment: Is the effort worth expending? *Language Learning in Higher Education, 8*(2), 239-264. https://doi.org/doi:10.1515/cercles-2018-0015

Figa, J. G., Tarekegne, W. M., & Kebede, M. A. (2020). The practice of formative assessment in Ethiopian secondary school curriculum implementation: The case of West Arsi zone secondary schools. *Educational Assessment*, *25*(4), 276-287. https://doi.org/10.1080/10627197.2020.1766958

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

Gipps, C. (1999). Chapter 10: Socio-cultural aspects of assessment. *Review of Research in Education, 24*(1), 355-392. https://doi.org/10.3102%2F0091732X024001355

Hancock, B., Ockleford, E., & Windridge, K. (2007). *An introduction to qualitative research*. Trent focus group Nottingham.

Hattie, J., & Timperley, H. (2007). The power of feedback. *Review of Educational Research, 77*(1), 81-112. https://doi.org/10.3102/003465430298487

Havnes, A., Smith, K., Dysthe, O., & Ludvigsen, K. (2012). Formative assessment and feedback: Making learning visible. *Studies in Educational Evaluation, 38*(1), 21-27. https://doi.org/10.1016/j.stueduc.2012.04.001

Heritage, M. (2007). Formative assessment: What do teachers need to know and do? *Phi Delta Kappan, 89*(2), 140-145. https://doi.org/10.1177/003172170708900210

Kang, M., & Lam, R. (2024). Understanding university English instructors' assessment literacy: A formative assessment perspective. *Language Testing in Asia*, *14*(1), 52. https://doi.org/10.1186/s40468-024-00323-y

Khaniya, T., Parajuli, T. R., & Nakarmi, S. S. (2015). *Background study: Curriculum, Textbook, and student assessment and evaluation*. Ministry of Education, Nepal.

Kyaruzi, F., Strijbos, J.-W., Ufer, S., & Brown, G. T. L. (2019). Students' formative assessment perceptions, feedback use and mathematics performance in secondary schools in Tanzania. *Assessment in Education: Principles, Policy & Practice, 26*(3), 278-302. https://doi.org/10.1080/0969594X.2019.1593103

Lajoie, S. P. (2005). Extending the scaffolding metaphor. *Instructional Science, 33*(5-6), 541-557. https://doi.org/10.1007/s11251-005-1279-2

Lajoie, S. P., & Lesgold, A. M. (1992). Dynamic assessment of proficiency for solving procedural knowledge tasks. *Educational Psychologist, 27*(3), 365-384. https://doi.org/10.1207/s15326985ep2703_6

Lantolf, J. P. (2007). Sociocultural theory. In J. Cummins & C. Davison (Eds.), *International handbook of English language teaching* (pp. 693-700). Springer. https://doi.org/10.1007/978-0-387-46301-8_45

Leenknecht, M., Wijnia, L., Köhlen, M., Fryer, L., Rikers, R., & Loyens, S. (2021). Formative assessment as practice: The role of students' motivation. *Assessment & Evaluation in Higher Education, 46*(2), 236-255. https://doi.org/10.1080/02602938.2020.1765228

Li, J., & Yongqi GU, P. (2023). Developing classroom-based formative assessment literacy: An EFL teacher's journey. *Chinese Journal of Applied Linguistics*, *46*(2), 198-218. https://doi.org/10.1515/CJAL-2023-0204

Magno, C., & Lizada, G. S. (2015). Features of classroom formative assessment. *Educational Measurement and Evaluation Review, 6*. https://ssrn.com/abstract=2649317

Maniati, M., Haghighi, S. B., Shooshtari, Z. G., & Afshar, M. R. (2023). Differential effects of focused/unfocused written corrective feedback on learning transfer in students of medical sciences. *English Teaching & Learning*. https://doi.org/10.1007/s42321-023-00143-y

Michell, M., & Sharpe, T. (2005). Collective instructional scaffolding in English as a second language classrooms. *Prospect, 20*(1). https://search.informit.org/doi/epdf/10.3316/aeipt.143259

Millonig, D. J., Stickler, U., & Coleman, J. A. (2019). Young pupils' perceptions of their foreign language learning lessons: The innovative use of drawings as a research tool. *The Language Learning Journal, 47*(2), 229-245. https://doi.org/10.1080/09571736.2016.1270348

Nicol, D. J., & Macfarlane-Dick, D. (2006). Formative assessment and self-regulated learning: a model and seven principles of good feedback practice. *Studies in Higher Education, 31*(2), 199-218. https://doi.org/10.1080/03075070600572090

Noroozi, O., Kirschner, P. A., Biemans, H. J., & Mulder, M. (2018). Promoting argumentation competence: Extending from first-to second-order scaffolding through adaptive fading. *Educational psychology review*, *30*(1), 153-176. https://doi.org/10.1007/s10648-017-9400-z

Ozan, C., & Kıncal, R. Y. (2018). The effects of formative assessment on academic achievement, attitudes toward the lesson, and self-regulation skills. *Educational Sciences: Theory & Practice, 18*(1). https://doi.org/10.12738/estp.2018.1.0216

Pastore, S., Manuti, A., & Scardigno, A. F. (2019). Formative assessment and teaching practice: the point of view of Italian teachers. *European Journal of Teacher Education, 42*(3), 359-374. https://doi.org/10.1080/02619768.2019.1604668

Pellegrino, J. W., Chudowsky, N., & Glaser, R. (2001). *Knowing what students know: The science and design of educational assessment*. ERIC.

Pryor, J., & Crossouard, B. (2008). A socio-cultural theorisation of formative assessment. *Oxford Review of Education, 34*(1), 1-20. https://doi.org/10.1080/03054980701476386

Rahman, K. A., Hasan, M. K., Namaziandost, E., & Ibna Seraj, P. M. (2021). Implementing a formative assessment model at the secondary schools: Attitudes and challenges. *Language Testing in Asia*, *11*(1), 18. https://doi.org/10.1186/s40468-021-00136-3

Rana, K., Greenwood, J., & Fox-Turnbull, W. (2020). Implementation of Nepal's education policy in ICT: Examining current practice through an ecological model. *The Electronic Journal of Information Systems in Developing Countries*, *86*(2), 1-16. https://doi.org/10.1002/isd2.12118

Rana, K., & Rana, K. (2019). Teaching and testing of English listening and speaking in secondary schools in Nepal: Pretend for praxis? *Journal of NELTA*, *24*(1-2), 17-32. https://doi.org/10.3126/nelta.v24i1-2.27678

Rodrigues, S. (2007). Assessing formatively in the English language classroom. *Journal of Research and Reflections in Education, 1*(1), 1-27. https://ecommons.aku.edu/pakistan_ied_pdck/6

Rojas-Drummond, S., Torreblanca, O., Pedraza, H., Vélez, M., & Guzmán, K. (2013). 'Dialogic scaffolding': Enhancing learning and understanding in collaborative contexts. *Learning, Culture and Social Interaction*, *2*(1), 11-21. https://doi.org/10.1016/j.lcsi.2012.12.003

Ruan, L. (2015). Language teaching from the view of formative assessment. *Theory and Practice in Language Studies, 5*(1), 92-96. https://doi.org/10.17507/tpls.0501.12

Sadler, D. R. (2010). Beyond feedback: developing student capability in complex appraisal. *Assessment & Evaluation in Higher Education*, *35*(5), 535-550. https://doi.org/10.1080/02602930903541015

Said Pace, D. (2018). *Assessment for Learning (AfL) in one Maltese State College* [Doctoral thesis]. University of Sheffield. http://etheses.whiterose.ac.uk/21381/

Sardareh, S. A., & Saad, M. R. M. (2012). A sociocultural perspective on assessment for learning: The case of a Malaysian primary school ESL context. *Procedia - Social and Behavioral Sciences, 66*, 343-353. https://doi.org/10.1016/j.sbspro.2012.11.277

Smith, J. A., Larkin, M., & Flowers, P. (2009). *Interpretative phenomenological analysis: Theory, method and research*. SAGE.

Stull, J., Varnum, S. J., Ducette, J., & Schiller, J. (2011). The many faces of formative assessment. *International Journal of Teaching and Learning in Higher Education, 23*(1), 30-39.

Sultana, N. (2019). Language assessment literacy: An uncharted area for the English language teachers in Bangladesh. *Language Testing in Asia*, *9*(1), 1. https://doi.org/10.1186/s40468-019-0077-8

Tajabadi, A., Ahmadian, M., Dowlatabadi, H., & Yazdani, H. (2023). EFL learners' peer negotiated feedback, revision outcomes, and short-term writing development: The effect of patterns of interaction. *Language Teaching Research*, *27*(3), 689-717. https://doi.org/10.1177/1362168820951207

Thompson, G., & Woodman, K. (2019). Exploring Japanese high school English teachers' foreign language teacher efficacy beliefs. *Asia-Pacific Journal of Teacher Education, 47*(1), 48-65. https://doi.org/10.1080/1359866X.2018.1498062

Tuluk, A., & Yurdugul, H. (2020). Design and development of a web based dynamic assessment system to increase students' learning effectiveness. *International Journal of Assessment Tools in Education, 7*(4), 631-656. https://doi.org/10.21449/ijate.730454

Vygotsky, L. S. (1978). Tool and symbol in child development. In M. Cole, V. Jolm-Steiner, S. Scribner, & E. Souberman (Eds.), *Mind in society: Development of higher psychological processes*. Harvard University Press. https://doi.org/10.2307/j.ctvjf9vz4.6

Walqui, A. (2006). Scaffolding instruction for English language learners: A conceptual framework. *International Journal of Bilingual Education and Bilingualism, 9*(2), 159-180. https://doi.org/10.1080/13670050608668639

Wati, I., & Dayal, H. C. (2022). Exploring possibilities and challenges of lesson study: A case study in a small island developing state. *Waikato Journal of Education*, *27*(3), 73-88. https://doi.org/10.15663/wje.v27i3

Widiastuti, I. A. M. S., Mukminatien, N., Prayogo, J. A., & Irawati, E. (2020). Dissonances between teachers' beliefs and practices of formative assessment in EFL classes. *International Journal of Instruction*, *13*(1), 71-84. https://doi.org/10.29333/iji.2020.1315a

Wood, D. F. (2010). Formative assessment. In T. Swanwick, K. Forrest, & B. C. O'Brien (Eds.), *Understanding medical education* (pp. 259-270). Wiley. https://doi.org/10.1002/9781444320282.ch18

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

Xiao, Y. (2017). Formative assessment in a test-dominated context: How test practice can become more productive. *Language Assessment Quarterly, 14*(4), 295-311. https://doi.org/10.1080/15434303.2017.1347789

Xiao, Y., & Yang, M. (2019). Formative assessment and self-regulated learning: How formative assessment supports students' self-regulation in English language learning. *System, 81*, 39-49. https://doi.org/10.1016/j.system.2019.01.004

Yan, Q. (2024). Exploring Chinese university EFL students' perceptions of formative assessment: A qualitative study. *System*, *125*, 103391. https://doi.org/10.1016/j.system.2024.103391

Yan, Z., Li, Z., Panadero, E., Yang, M., Yang, L., & Lao, H. (2021). A systematic review on factors influencing teachers' intentions and implementations regarding formative assessment. *Assessment in Education: Principles, Policy & Practice, 28*(3), 228-260. https://doi.org/10.1080/0969594X.2021.1884042

Yin, R. K. (2015). *Qualitative research from start to finish*. The Guilford Press.

Yongqi Gu, P., & Lam, R. (2023). Developing assessment literacy for classroom-based formative assessment. *Chinese Journal of Applied Linguistics*, *46*(2), 155-161. https://doi.org/10.1515/cjal-2023-0201

Yu, S., & Hu, G. (2017). Can higher-proficiency L2 learners benefit from working with lower-proficiency partners in peer feedback? *Teaching in Higher Education, 22*(2), 178-192. https://doi.org/10.1080/13562517.2016.1221806

Zhang, H., Ge, S., & Mohd Saad, M. R. B. (2024). Formative assessment in K-12 English as a foreign language education: A systematic review. *Heliyon*, *10*(10). https://doi.org/10.1016/j.heliyon.2024.e31367

APPENDIX 1

Interview questions for secondary English teachers

1. Could you describe what sorts of assessment do you use in English classes?
2. Could you please describe types of formative assessment you regularly use in your English classes?
3. How do you decide which type of formative assessment strategies to use?
4. How do you provide feedback to your students?
5. How do you scaffold your students for advancing their learning?
6. Please share an example of a lesson where you applied formative assessment? What did it look like?
7. Do you believe formative assessment supports students' English learning? If so, how?
8. Have you noticed any changes in student performance or engagement as a result of using formative assessment?
9. How do you adjust your instruction based on what you learn from formative assessment results?
10. What kinds of feedback (oral, written, peer, self-assessment, etc.) do you give to students, and why?
11. How do you ensure that feedback is timely and useful for students?
12. Do you differentiate feedback for students with varying levels of English proficiency? How?
13. What role does student self-assessment or peer-assessment play in your classroom, if any?
14. How do students respond to the feedback you provide? Do they act on it?
15. In your experience, what feedback practices are most effective in helping students improve in reading, writing, speaking, or listening?
16. What challenges do you face in implementing formative assessment and feedback in your classroom?
17. Have you received any training or professional development related to formative assessment? If so, how has it helped?

Interview questions for students

1. How does your English teacher check your understanding during or after lessons?
2. Can you give an example of an activity in class that helps you understand how well you are learning?
3. What kind of feedback do you usually receive from your English teacher? (Written comments, corrections, oral feedback, etc.)
4. How often does your teacher give you feedback on your work?
5. Do you think the feedback your teacher gives helps you learn English better? Why or why not?
6. Can you share a moment when your teacher's feedback helped you improve in reading, writing, speaking, or listening?
7. Have you ever changed or improved your work based on feedback from your teacher? Can you give an example?
8. What kind of feedback do you find most helpful for your learning?
9. How do you feel when your teacher gives you feedback on your mistakes?
10. Do you feel more confident in English after receiving feedback? Why or why not?
11. Do you set goals for improving your English? If yes, how does your teacher help you with them?
12. Do you ever assess your own work or give feedback to classmates? How do you feel about that?

https://doi.org/10.17323/jle.2025.27389

Stylistic Redundancy and Wordiness in Introductions of Original Empirical Studies: Rhetorical Risks of Academic Writing

Elena Tikhonova [1], **Olga Zavolskaya** [2], **Nataliia Mekeko** [3]

[1] MGIMO University, Moscow, Russia

[2] Sergo Ordzhonikidze Russian State University for Geological Prospecting, Moscow, Russia

[3] RUDN University, Moscow, Russia

ABSTRACT

Background: The introduction of a research article plays a central role in shaping scientific argumentation. However, this section is often especially prone to stylistic overload, which can obscure the clarity of the author's position. While the issues of redundancy and wordiness have been broadly acknowledged in applied linguistics, there is still limited understanding of how these features are distributed in relation to rhetorical structure, particularly within Russian-language academic texts.

Purpose: To identify rhetorically sensitive areas of stylistic overload in the introductions of Russian-language research articles in the field of education.

Method: The analysis is based on a corpus of 40 introductions from empirical articles published in 2024 in leading Russian peer-reviewed journals in education. The rhetorical Move-Step model developed by Swales was used as the framework for annotation. Each fragment was manually coded for two dimensions: the type of deviation (wordiness or redundancy) and its communicative impact (according to the IMPACT scale). Pearson's chi-squared test was used to assess statistical significance.

Results: Stylistic overload was found to cluster in specific rhetorical steps, especially those related to establishing the importance of the topic (M1_S2), identifying gaps in the literature (M2_S1), and stating research objectives (M3_S2). The most frequent features included syntactic overcomplexity, vague abstract nouns, and overused credibility markers. A high level of negative communicative impact (IMPACT = HIGH) was observed in 60 fragments, most of which were located in the mentioned segments. Statistical testing (χ^2, $p < 0.0001$) confirmed a significant relationship between rhetorical function and the type of deviation.

Conclusion: The results confirm that stylistic overload in introductions is not accidental but structurally motivated. This supports the need for rhetorically informed strategies in teaching academic writing. The annotation scheme developed in the study may be applied in future corpus-based analyses of academic Russian.

KEYWORDS

academic writing, wordiness, redundancy, rhetorical analysis, CARS model, Russian academic texts

Citation: Tikhonova, E., Zavolskaya, O., & Mekeko, N. (2025). Stylistic Redundancy and wordiness in introductions of original empirical studies: Rhetorical risks of academic writing. *Journal of Language and Education, 11*(2), 125-136. https://doi.org/10.17323/jle.2025.27389

Correspondence:
Elena Tikhonova,
tikhonova.e.v@inno.mgimo.ru

Received: December 13, 2024
Accepted: June 10, 2025
Published: June 30, 2025

INTRODUCTION

Issues of stylistic clarity and textual economy remain central to current research on academic writing (Swales, 1990; Hyland, 2005; Tikhonova & Mezentseva, 2024). Despite the growing body of methodological guidance on academic style, many authors continue to struggle with expressing their ideas efficiently. This often results in the use of redundant structures and vague formulations that blur meaning (Flowerdew & Forest, 2015; Leufkens, 2023). Among the most common stylistic problems are *wordiness* and *redundancy*, both of which undermine analytical precision and textual clarity while increasing the cognitive load for

readers (Tikhonova, Mezentseva & Kasatkin, 2024; Strunk & White, 2000; Demir, 2019).

Stylistic wordiness refers to the use of low-informational expressions, epistemic modifiers, and nominalizations that reduce semantic precision and clarity (Tikhonova & Mezentseva, 2024; Salager-Meyer, 1994). In contrast, redundancy encompasses lexical tautologies, syntactic repetition, and appositive elaborations that overload the sentence and obscure the core message (Williams & Bizup, 2017; Kravtchenko & Demberg, 2022; Tikhonova et al., 2024). While such stylistic deviations have been examined in numerous international studies (e.g., Hyland, 2005; Biber & Gray, 2010; Fruehwald, 2010; Goonaratna, 2002a; 2002b; Alontseva & Ermoshin, 2019), their specific manifestations and rhetorical effects in Russian-language academic texts remain underexplored. Of particular interest is the structural distribution of wordiness and redundancy across rhetorical segments of academic writing.

Studies indicate that the opening segments of an Introduction, particularly those associated with establishing the research territory (Move 1), are especially vulnerable to stylistic overload. Writers often resort to sweeping generalizations, repetitive thematic statements, and formulaic clichés, which diminish the cognitive density of the text and hinder the reader's focus on a specific research problem (Samraj, 2002; Kanoksilapatham, 2005; Gong & Barlow, 2022). The second rhetorical move (Move 2), aimed at identifying a research gap by highlighting limitations, contradictions, or unresolved issues in previous studies, proves no less susceptible. In an effort to balance critique with scholarly politeness, authors tend to rely on mitigating strategies—explanatory repetitions, vague phrasings, and abstract formulations. These choices often compromise analytical clarity and contribute to verbal redundancy. Corpus-based studies confirm this tendency: in academic texts from the social sciences and education, indirect criticism and cautious evaluative language are particularly prevalent in segments aligned with Move 2 (Wahyuningtyas & Wulandari, 2023; Aoulad-Ouda & Chellaoui, 2023). Finally, the third rhetorical move (Move 3), which presents the aim and scope of the research, is also affected by stylistic standardization. As noted by Hyland (2008), this stage is often marked by template-like declarations and predictable syntactic patterns that frequently echo earlier content, thereby weakening textual originality and cognitive depth. In sum, all three rhetorical moves of the Introduction exhibit varying degrees of vulnerability to stylistic and structural redundancy, with the second move emerging as the most rhetorically strained.

However, these observations are typically based on English-language data and have yet to be empirically confirmed on a corpus of texts rooted in the Russian academic tradition. The overwhelming majority of Russian academic journals continue to use Russian as the primary medium of scholarly communication, despite increasing pressure from the global English-dominated publication landscape. This persistence is shaped not only by national science policy but also by the linguistic challenges associated with translating research from Russian into English - especially due to semantic and structural differences between the two languages (Raitskaya & Tikhonova, 2020; Smirnova et al., 2021). As a result, Russian-language academic texts have developed their own stylistic conventions, embedding distinctive rhetorical and genre-based practices (Raitskaya & Tikhonova, 2020). Among these, features such as redundancy and verbal padding have become entrenched, particularly in introductory sections. This points to a significant gap in our understanding of rhetorical and stylistic risks specific to Russian academic writing, particularly in applied disciplines such as education. The present study seeks to address this gap by conducting a quantitative, corpus-based analysis grounded in the CARS rhetorical model (Swales, 1990) and recent classifications of stylistic deviation (Tikhonova et al., 2024; Tikhonova & Mezentseva, 2024).

The purpose of the present study is to identify patterns in the distribution and communicative impact of stylistic deviations, namely wordiness and redundancy. The following research questions are posed:

RQ1: What are the frequency characteristics and rhetorical localization of wordiness and redundancy in the Introduction sections of Russian-language empirical research articles in education?
RQ2: Which rhetorical moves and steps in the Introduction are most prone to stylistic deviations?
RQ3: What impact do these stylistic deviations have on the communicative clarity of the text?

The study tests the following hypotheses:

H1: The highest concentration of wordiness and redundancy occurs in rhetorical segments related to justifying the significance of the research topic (M1_S2) and identifying gaps in existing literature (M2_S2).
H2: While wordiness occurs more frequently than redundancy, the latter more often results in a higher level of negative communicative impact.
H3: Stylistic deviations are structurally motivated and shaped by genre-based expectations.

METHOD

Corpus

The material for analysis consisted of Introduction sections from 40 (forty) original research articles published in 2024 in Russian peer-reviewed journals ranked between positions 1 and 18 in the Russian Science Citation Index (RSCI) for the field of Education and Pedagogical Sciences. All articles were written in Russian by authors affiliated with Russian

universities and research institutions. The choice of a Russian-language corpus reflects the study's objective to examine stylistic practices of academic writing in their authentic, unadapted form, as they occur within the national scientific context.

Inclusion Criteria

The selection of materials was based on the following formal and substantive criteria:

(1) Bibliometric status of the journal: inclusion in the top ten RSCI journals in the field of Education and Pedagogical Sciences;
(2) High categorical ranking: classification of the journal in the first or second category according to the internal Science Index system;
(3) Presence of a standalone Introduction: structurally distinct and separate from other sections such as literature review or methodology;
(4) Genre and rhetorical organization: presence of rhetorical moves that allow segmentation of the Introduction using the Move-Step model;
(5) Topical relevance: alignment of the article's content with the fields of pedagogy, teaching methodology, educational technologies, or educational psychology;
(6) Open access: ensuring legal and ethical transparency of the analysis.

Corpus Profile

Each article was entered into an analytical database that included the following parameters:

1. Journal title;
2. Journal category;
3. Year of publication;
4. Author team;
5. Full article title;
6. Journal name;
7. Open access availability

The entire corpus was organized into a working spreadsheet (Appendix 1), allowing for the tracking of bibliometric and contextual characteristics of each analytical unit. The corpus consists of 40 Introductions from original empirical studies published in journals such as *Voprosy obrazovaniya (Education Issues)*, *Vysshee obrazovanie v Rossii (Higher Education in Russia)*, *Obrazovanie i nauka (Education and Science)*, *Psikhologicheskaya nauka i obrazovanie (Psychological Science and Education)*, *Yazyk i kultura (Language and Culture)*, *Rusistika (Russian Studies)*, *Integratsiya obrazovaniya (Integration of Education)*, *Sovremennye problemy nauki i obrazovaniya (Contemporary Problems of Science and Education)*, *Perspektivy nauki i obrazovaniya (Perspectives of Science and Education)*, *Psikhologo-pedagogicheskie issledovaniya (Psychological and Pedagogical Research)*, and *Vestnik RUDN: Series Psychology and Pedagogy*.

This corpus (Appendix 2) provides a solid basis for a reliable and representative analysis of current academic writing practices in educational science and related disciplines within the context of contemporary Russian scholarly communication.

Rhetorical Structure Annotation

To annotate the rhetorical segments of introductions, the Move-Step model developed by Swales (1990; 2004) was used (Table 1).

The rhetorical annotation based on the Move-Step model was necessary for functionally aligning segments that contain stylistic deviations and for conducting statistical analysis of their distribution in relation to rhetorical function.

Classification Scheme for Wordiness and Redundancy and Annotation Procedure

To annotate stylistically overloaded elements in the corpus of introductions, the typology of wordiness developed by Tikhonova and Mezentseva (2024) and the classification of textual redundancy proposed by Tikhonova, Mezentseva, and Kasatkin (2024) were used. Both models were adapted for operationalization in manual annotation and implemented as a system of tags assigned to each segment of text showing signs of stylistic deviation.

The categories of wordiness and their corresponding tags include:

1. General phrases and low-information introductory expressions (WORDINESS_GENERAL);
2. Excessive use of epistemic modifiers (WORDINESS_HEDGING);
3. Syntactic overload and structural complexity (WORDINESS_COMPLEXITY);
4. Empty or abstract references without specific content (EMPTY_REFERENCE);
5. Nominalization that reduces verbal dynamism (NOMINALIZATION);
6. Formulaic phrases that serve no analytical function (FORMULAIC_PHRASE).

The categories of textual redundancy and their tags include:

1. Lexical tautology and semantic repetition (REDUNDANCY_LEXICAL);
2. Redundant syntactic structures (REDUNDANCY_STRUCTURE);
3. Appositive explanations that duplicate content (APPOSITIVE_PHRASE).

Table 1

Rhetorical Structure of the Introduction Section According to Swales (1990; 2004)

Move	Step	Rhetorical Function
Move 1. Establishing a Research Territory	Step 1. Topic Presentation	Introducing the subject area and situating the topic
	Step 2. Justifying the Importance of the Topic	Arguing for the relevance and significance of the research problem
	Step 3. Reviewing Key Studies	Referring to prior research and describing the current state of the field
Move 2. Establishing a Niche	Step 1. Indicating a Gap in Knowledge	Highlighting areas that remain underexplored
	Step 2. Identifying Conflicts or Contradictions	Emphasizing inconsistent findings or theoretical disagreements
	Step 3. Pointing to Methodological or Practical Limitations	Critiquing existing approaches or identifying applied challenges
Move 3. Presenting the Present Work	Step 1. Stating the Aim and Objectives of the Study	Defining the research goal, object, and focus
	Step 2. Formulating Research Questions or Hypotheses	Outlining key research questions or testable propositions
	Step 3. Providing a Brief Methodological Description	Summarizing the approach, methods, and dataset

In cases where a single fragment demonstrated multiple types of stylistic deviation, each tag was recorded as an individual unit of observation. This approach ensured comprehensive typological coverage.

Although redundancy is often considered a subset of wordiness, the two were distinguished analytically in this study. Wordiness is defined as stylistic elements that complicate textual processing due to formal heaviness, such as clichés, nominalizations, or epistemic modifiers. Redundancy, by contrast, is understood as structural or semantic repetition of information already stated.

This distinction allowed for a more precise evaluation of the nature of stylistic deviations and their distribution across the rhetorical segments of the introduction. It also increased the sensitivity of interpreting the level of communicative impact (tagged as IMPACT), which reflects the degree to which a deviation affects the clarity and analytical transparency of the text (Table 2).

This scale served as the main evaluative matrix for each annotated fragment in the corpus. The impact level was assigned through expert interpretation, considering the rhetorical Move and Step as well as the combination of formal and rhetorical markers observed during annotation.

The annotation was conducted manually using Microsoft Excel. For each fragment, the following information was recorded: a unique ID; article number, Move and Step; the text of the fragment; the category of wordiness or redundancy; IMPACT level; and a brief explanatory comment (Appendix 2).

Data Analysis

The quantitative analysis of the annotated corpus aimed to identify stable patterns in the distribution of stylistically overloaded fragments (instances of wordiness and redundancy) depending on the rhetorical structure of the introduction section. The analysis focused on three main relationships: (1) the frequency and type of stylistic deviations were examined across rhetorical Moves and Steps, (2) dominant categories of wordiness and redundancy were identified within specific functional segments, (3) the level of negative communicative impact (IMPACT) was assessed in relation to the rhetorical role of each fragment.

To verify the observed patterns, contingency tables were constructed, which confirmed the presence of statistically significant associations between the rhetorical function of a fragment (Move or Step) and the type of stylistic deviation it contained. These findings support the conclusion that stylistic overload in academic writing is not a random occurrence. Rather, it is structurally conditioned and closely aligned with rhetorical tension and the communicative demands of specific segments within the introduction.

Statistical Data Processing

To confirm the observed patterns in the distribution of stylistic deviations across the rhetorical structure of introductions, a statistical analysis was conducted based on the association of categorical variables. The goal was to determine whether there were significant dependencies between three variables: the rhetorical function of the fragment (Move and

Table 2

Criteria for Evaluating the Impact of a Fragment on Textual Clarity and Analytical Precision (IMPACT)

Level	Label	Description	Basis for Classification
Low	LOW	The stylistic element has minimal effect on text clarity. It is acceptable within genre norms and does not require revision.	A stylistic deviation is present (such as a cliché or tautology), but it does not disrupt logical flow, hinder interpretation, or require rephrasing.
Medium	MEDIUM	Slows reading and comprehension and makes it harder to understand the author's position or the logic of transitions.	The fragment increases cognitive load, reduces precision of expression, introduces unnecessary links in the argument, or creates semantic ambiguity.
High	HIGH	Significantly distorts or impedes understanding of the research aim, central claim, or analytic logic.	The deviation breaks the coherence between parts of the text, forces rereading, may lead to misinterpretation, and undermines the academic credibility of the statement.

Step), the type of deviation (CATEGORY), and its level of communicative impact (IMPACT).

At the first stage, contingency tables were constructed for the following pairs of variables:

(1) Move and Step by CATEGORY (rhetorical position by type of deviation);
(2) Move and Step by IMPACT (rhetorical position by level of impact);
(3) CATEGORY by IMPACT (type of deviation by level of impact).

In all cases, the Pearson chi-square test was used to check the hypothesis of independence between variables. The calculations were carried out using standard statistical tools in Python (SciPy) and Microsoft Excel. The significance level was set at $p < 0.05$.

Additionally, to clarify the nature of the impact within rhetorical segments, the proportion of fragments marked with IMPACT = HIGH was calculated for each step of the Move–Step model. This helped to identify parts of the text that are especially vulnerable to communicative overload or distortion and to justify these segments as priority targets for revision. Manual verification of the annotated tables ensured accurate alignment of each fragment with its rhetorical function.

RESULTS

General Description of the Corpus

A total of 487 text fragments were identified and analyzed, each containing features that deviated from stylistic norms. These fragments were classified either as wordiness or as textual redundancy. The annotated corpus therefore consisted of 487 analytical units, with each unit labeled according to the type of deviation, its rhetorical position based on Move and Step, and its level of communicative impact as measured by the IMPACT scale.

Out of the total number of fragments, 342 were categorized as wordiness. This refers to cases where scientific writing becomes less precise and less informative due to the use of vague or redundant language. These include common rhetorical clichés, "empty" introductory phrases, excessive modifiers, abstract or poorly specified formulations, and the replacement of verbal structures with nominalizations. Nominalizations slow down meaning processing and obscure the main point. These elements do not contribute meaningful content but increase the length of the text and reduce its clarity.

A total of 145 fragments were classified as redundancy. These are cases of structural or semantic repetition, where information is duplicated either lexically or syntactically. As a result, wordiness appears to dominate not only in quantity but also in variety. This reflects a widespread tendency toward vague expression and verbal expansion in academic writing, particularly in introductions. The detailed distributions are presented in Tables 3 and 4.

Table 3

Categories of Wordiness Identified in the Corpus (n = 342)

Category of Wordiness	Frequency	Percentage (%)
WORDINESS_GENERAL	84	24.6
NOMINALIZATION	81	23.7
WORDINESS_COMPLEXITY	62	18.1
FORMULAIC_PHRASE	46	13.4
EMPTY_REFERENCE	44	12.9
WORDINESS_HEDGING	25	7.3

Table 4

Categories of Redundancy Identified in the Corpus (n = 145)

Category of Redundancy	Frequency	Percentage (%)
REDUNDANCY_STRUCTURE	73	50.3
REDUNDANCY_LEXICAL	62	42.8
APPOSITIVE_PHRASE	10	6.9

The distribution of communicative impact levels revealed the following pattern. A total of 220 fragments were classified as having a low level of impact (LOW). These do not distort the perception of scientific content, although they diverge from the expected genre conventions. Another 207 fragments were assessed as having a medium level of impact (MEDIUM), suggesting that they hinder the interpretation of the author's reasoning or the logic of transitions. The most critical group includes 60 fragments with a high level of impact (HIGH), the vast majority of which are associated with wordiness. These are primarily linked to empty nominal expressions, syntactic overload, and unsupported abstract statements (Table 5).

Thus, even at the initial stage of analysis, a high structural density of stylistic deviations can be observed, affecting more than 90 percent of all introductions in the corpus. These findings indicate that stylistic overload in scientific writing is not random but reflects a systemic pattern rooted in the genre conventions of academic discourse.

Dominant Forms of Stylistic Overload

Unlike the previous section, which described the overall set of fragments containing stylistic deviations, this part focuses on the qualitative distribution of deviation categories within the two broader groups: wordiness and redundancy. The purpose of the analysis is to identify the most frequent and representative types of stylistic overload and to assess their contribution to the overall picture of academic writing distortion.

The results show that within the wordiness group, three categories have the greatest weight: general discourse clichés and introductory phrases (24.6 % of all wordiness cases), nominalization (23.7 %), and syntactic complexity (18.1 %). These three categories form the core of wordiness in the corpus, together accounting for nearly 67 % of all relevant fragments. Their predominance indicates a tendency among authors toward generalized and overly formalized statements as well as syntactically complex constructions that reduce textual clarity. Less frequent but rhetorically significant categories include formulaic expressions (13.5 %) and vague abstract references (12.9 %), which confirms the multilayered nature of wordiness as a stylistic phenomenon.

In the group of textual redundancy, structural redundancy emerges as the most prominent category, accounting for more than half of all identified fragments (50.3 %). This type of deviation involves the inclusion of unnecessary syntactic elements that repeat information already expressed. Nearly as frequent is lexical tautology (42.8 %), which points to a common tendency to restate the same idea through repetitive wording and pleonastic constructions. Appositive phrases that duplicate content are less common (6.9 %), but they still contribute significantly to cognitive noise by undermining the structural compactness of the sentence.

Taken together, the quantitative distribution of categories reveals a clear internal hierarchy among the forms of stylistic deviation. The dominant violations are those that directly obstruct access to the semantic structure of the introduction. These include formulaic expressions that add no informational value and constructions that overload the grammatical framework of the text. These findings provide a foundation for further analysis of which rhetorical segments of the introduction are most prone to such forms of stylistic overload and how they may distort the interpretability of the text.

Rhetorical Localization of Stylistic Deviations

One of the central objectives of this study was to identify how stylistic deviations are distributed across the rhetorical structure of the introduction. To achieve this, annotated fragments were mapped onto the steps of the Move–Step

Table 5

Distribution of Wordiness and Redundancy Cases by Communicative Impact Level (IMPACT)

CATEGORY	HIGH	LOW	MEDIUM	Total	Share of HIGH (%)
WORDINESS_COMPLEXITY	23	2	37	62	37.0
REDUNDANCY_STRUCTURE	13	28	32	73	17.8
NOMINALIZATION	11	25	45	81	13.5
REDUNDANCY_LEXICAL	6	22	34	62	9.6
EMPTY_REFERENCE	5	21	18	44	11.3
APPOSITIVE_PHRASE	1	4	5	10	10.0
WORDINESS_GENERAL	1	59	24	84	1.1
FORMULAIC_PHRASE	0	43	3	46	0.0
WORDINESS_HEDGING	0	16	9	25	0.0

model proposed by Swales (1990, 2004), which segments the introduction into functionally distinct parts ranging from the presentation of the research territory to the statement of goals and objectives.

The results of the analysis (Table 6) support the hypothesis that stylistic overload is structurally conditioned. The highest concentration of deviations occurs in those segments of the introduction where authors experience the greatest rhetorical pressure: where they are expected either to demonstrate subject-matter awareness or to formulate the novelty of the research. These include, above all, Step M1_S2 (justifying the relevance of the topic), Step M1_S3 (reviewing previous research), and Step M3_S1 (stating the aim of the study).

In Step M1_S2, a total of 125 stylistically deviant fragments were identified, making it the most overloaded rhetorical segment in the corpus. This part of the introduction is particularly marked by lexical redundancy, generic statements of significance, and nominalizations, all of which reflect an attempt to strengthen the argument through intensified language. Such overuse often results in verbosity and repetitive evaluative expressions that fail to provide any substantive clarification of the problem. Step M1_S3 also exhibits a high density of stylistic issues, with 106 annotated fragments. This segment frequently includes vague generalizations, formulaic phrases, and empty references that are not supported by analytical interpretation. These features suggest a lack of critical engagement with the literature and a focus on formally fulfilling the genre's requirements. Although Step M3_S1 contains fewer annotated fragments (61), it shows a disproportionately high share of deviations rated as having medium or high communicative impact. In this segment, nominalizations are particularly prevalent, leading to overly complex and imprecise formulations of the research aim. Additionally, syntactic overloading is common, which hampers the reader's ability to follow the logic of the study.

Steps M1_S1 (thematic introduction) and M2_S2–M2_S3 (niche description) were less saturated with stylistic deviations, although they still contain problematic fragments. These are primarily related to formulaic expressions and structural redundancy, particularly when articulating methodological limitations or discrepancies in previous approaches.

The distribution of stylistic violations across the Move–Step model clearly reveals a concentration in the functionally significant sections of the introduction. This indicates that linguistic overload should be viewed not only as a stylistic issue but also as a rhetorical phenomenon shaped by the communicative challenges authors face when addressing the core demands of the genre. Further analysis of the communicative impact levels of these deviations will help identify which types most severely hinder interpretation and therefore require prioritized revision.

Levels of Communicative Impact

The overall structure of the corpus reveals that the vast majority of fragments with stylistic deviations fall into either the low (45.2 percent) or medium (42.5 percent) impact categories. Only 12.3 percent of the annotated fragments, or 60 units, were classified as having a high level of communicative impact (IMPACT = HIGH) as shown in Table 5. However, the significance of this 12 percent should not be underestimated. These fragments represent critical areas of the text where the logic of the argument or the meaning of key statements is distorted.

Particular attention should be given to the category of syntactic complexity (WORDINESS_COMPLEXITY), which displays the highest proportion of critically disruptive cases. Within this category, 37.1 % of the fragments were rated as having a high impact. This finding suggests that overly long and structurally overloaded sentences are especially harm-

Table 6

Rhetorical Distribution of Stylistic Deviations in Introduction Sections

Rhetorical Step	APPOSITIVE_PHRASE	EMPTY_REFERENCE	FORMULAIC_PHRASE	NOMINALIZATION	REDUNDANCY_LEXICAL	REDUNDANCY_STRUCTURE	WORDINESS_COMPLEXITY	WORDINESS_GENERAL	WORDINESS_HEDGING	Total
M1_S1	2	2	3	4	9	4	6	26	0	56
M1_S2	0	11	10	16	19	15	23	27	4	125
M1_S3	4	15	15	21	11	17	11	9	3	106
M2_S1	0	5	6	8	5	6	9	4	4	47
M2_S2	2	1	1	2	3	14	1	0	6	30
M2_S3	1	2	0	7	6	5	2	7	4	34
M3_S1	1	4	9	17	7	10	7	6	0	61
M3_S2	0	0	0	1	1	0	2	2	0	6
M3_S3	0	4	2	5	1	2	1	3	4	22

ful to the clarity of academic writing, particularly in rhetorically dense segments of the introduction.

The next most prominent categories are structural redundancy (17.8 %) and nominalization (13.6 %. In the case of structural redundancy, the primary communicative distortion arises from repeated content, which blurs the intended emphasis and compromises the conciseness of the exposition. Nominalization, by contrast, leads to the absence of verbal dynamism, which causes sentences to lose their specificity and precision, making the logical structure of the text less distinct.

A less pronounced yet still noteworthy effect is observed in the categories of empty references (11.4 % and appositive phrases (10 %). These types of deviations suggest a partial loss of conceptual clarity or the inclusion of excessive explanatory detail that lacks analytical purpose.

Taken together, these findings indicate that a high level of communicative disruption is usually associated not with superficial stylistic variety, but with deeper structural flaws in the text. These include disrupted syntactic flow, misalignment between grammatical form and logical function, and a lack of referential precision. Such observations make it possible to identify high-priority areas for editorial revision, which will be discussed in more detail in the following subsection.

Statistical Validation of Rhetorically Motivated Redundancy

A statistical analysis was conducted to examine the relationship between the rhetorical structure of the introduction, the types of stylistic deviations, and the degree of their communicative impact. The aim was to confirm that the patterns identified earlier are not random but demonstrate a consistent and structurally driven nature. For this purpose, contingency tables were used along with Pearson's chi-squared test, which makes it possible to assess the statistical significance of associations between categorical variables (Table 7).

Table 7

Chi-Squared Analysis of Contingency Tables for Three Pairs of Variables

Variable Pair	χ^2	df	p-value
MOVE_STEP × CATEGORY	159.45	64	< 0.0001
MOVE_STEP × IMPACT	36.84	16	0.0022
CATEGORY × IMPACT	145.84	16	< 0.0001

The analysis of the first contingency table (MOVE_STEP × CATEGORY) revealed a statistically significant relationship between rhetorical positioning and the type of stylistic overload ($\chi^2 = 159.45$; df = 64; $p < 0.0001$). This means that deviations from stylistic norms are not evenly distributed throughout the Introduction. Certain types of overload are typical of specific rhetorical steps. For example, nominalization and syntactic complexity are most frequently found in the formulation of the study's purpose (M3_S1) and hypotheses (M3_S2). In contrast, formulaic expressions and tautologies tend to cluster in steps M1_S2 and M1_S3, where the author needs to justify the topic's importance and provide a literature review.

The second table (MOVE_STEP × IMPACT) also demonstrated a statistically significant association ($\chi^2 = 36.84$; df = 16; $p = 0.0022$), confirming that not only the type of stylistic deviation but also its degree of impact on the text depends on its rhetorical position. In other words, not all segments of the Introduction are equally vulnerable to communicative distortions. For instance, the formulation of research questions and the description of the research niche (M3_S2, M2_S1) more often contain fragments with a high level of impact, suggesting that these rhetorical tasks present consistent challenges for authors.

The strongest association was found between the type of deviation and its communicative impact (CATEGORY × IMPACT), with the chi-square test yielding $\chi^2 = 145.84$ at 16 degrees of freedom ($p < 0.0001$). This result confirms that not all stylistic deviations have the same effect on how the text is perceived. For example, syntactic overload accounts for the highest share of critically impactful fragments, whereas formal clichés and idiomatic phrases more often fall into the category of stylistically undesirable but permissible recurrences.

The statistical analysis not only supports the earlier observations but also reinforces them with analytical precision. The results indicate that stylistic overload in academic writing follows a predictable and rhetorically driven pattern. These patterns of linguistic redundancy can be taken into account both in the editorial assessment of manuscripts and in the development of academic writing instruction.

High-Risk Zones: Priority Segments for Editing

One of the key outcomes of this study was the identification of rhetorically vulnerable segments within the introduction: parts of the text where linguistic overload is most pronounced and results in communicative distortion. Based on the proportion of fragments with a high impact level (IMPACT = HIGH) across different rhetorical steps of the Swales (1990) model, it is possible to pinpoint stable high-risk zones that require targeted editing.

The highest proportion of high-impact fragments was observed in step M3_S2 (the formulation of research questions or hypotheses), where 33.3% of all fragments were classified as critically overloaded. Although this step was represented by a relatively small number of units in the corpus, it

showed the greatest concentration of rhetorical strain. This suggests that formulating research assumptions within the constraints of academic genre norms presents a significant challenge. Authors often attempt to enhance the perceived significance of their hypotheses through convoluted formulations, relying on nominalization, vague definitions, and unnecessary elaborations.

A high proportion of fragments with IMPACT = HIGH was also found in step M2_S1 (identifying the research gap, 23.4 %) and in step M3_S3 (describing the methodology, 18.2 %). This indicates a systematic difficulty in transitioning from the literature review to the formulation of one's own research position. In step M2_S1, authors often attempt to emphasize novelty through rhetorical intensification, which may involve tautological repetition, syntactic overcomplication, and excessive explanatory phrases. In step M3_S3, difficulties arise from the challenge of articulating the methodology briefly and clearly; when linguistic precision is lacking, this part often becomes overloaded and poorly structured.

These patterns are further supported by the results of the statistical analysis. The contingency table MOVE_STEP × IMPACT revealed a statistically significant association between rhetorical function and impact level ($\chi^2 = 36.84$; $p = 0.0022$). This provides empirical confirmation for the idea that stylistic overload is not distributed randomly across the introduction but is concentrated in segments where rhetorical tension is especially high. In practical terms, this means that stylistic issues are most likely to emerge in the moments when the author must either establish a research niche, justify methodological choices, or create a smooth transition into the main body of the article. Consequently, steps M3_S2, M2_S1, M3_S3, and M1_S3 can be described as rhetorically charged zones, where the risk of communicative distortion is highest. These parts of the text require special attention during editing and can serve as focal points in academic writing instruction aimed at teaching strategic reduction and structural refinement.

DISCUSSION

The results of this study demonstrate that stylistic overload in academic introductions is neither a random nor evenly distributed phenomenon. On the contrary, the observed forms of wordiness and textual redundancy show clear rhetorical localization, corresponding to functionally demanding segments of the introduction's structure. Specifically, the highest concentration of stylistic pressure occurs in parts of the text where the author is expected to simultaneously establish the relevance of the topic, legitimize the research position, and outline the methodological framework.

These findings provide partial confirmation of Hypothesis 1, which proposed that the highest concentration of stylistic overload would occur in segments related to justifying the topic's significance (M1_S2) and identifying research gaps (M2_S2). The data support the first part of this assumption: Step M1_S2 indeed demonstrated the greatest rhetorical load, with a high frequency of both general formulations and nominalizations aimed at reinforcing argumentative weight. However, the expected prominence of Step M2_S2 was not confirmed. Although it contained several stylistic deviations, its overall density and communicative impact were notably lower than in other segments, suggesting that authors may allocate less rhetorical effort to articulating research gaps than to legitimizing the importance of their work.

The concentration of stylistic overload in these key rhetorical steps suggests that redundancy in academic writing may serve a defensive purpose. As Gong and Barlow point out, when presenting the novelty of their research, authors often use repetitive or softening expressions and rely on nominalizations. These choices help them avoid directly challenging previous studies while adhering to conventions of academic politeness. Such strategies are especially common when researchers work under pressure to publish. In those situations, texts often become saturated with surface markers of scientific credibility, such as complex syntax, cautious statements, and standardized rhetorical phrases. Although these features are intended to meet formal expectations, they can reduce the transparency of the writing and make it harder for readers to follow the line of reasoning. This has been noted by several authors, including Biesta and colleagues, who argue that the pursuit of academic recognition can come at the cost of clarity and accessibility.

The concentration of stylistic overload in key rhetorical steps of the Introduction suggests that redundancy in academic writing often serves a protective function. As Gong and Barlow (2022) point out, when formulating the novelty of their research, authors frequently resort to repetitive or mitigating constructions as well as to nominalizations in an effort to avoid direct confrontation with existing findings and to conform to the norms of academic politeness. Such strategies are particularly common in situations where authors operate under pressure to increase their publication output (Çakir et al., 2024). In these cases, the density of a text with formal markers of scientific discourse, such as complex syntax, epistemic caution, and standard rhetorical formulas, is perceived as a necessary condition for academic recognition, even when it reduces cognitive transparency and makes the argumentation harder to follow (Biesta et al., 2024).

The identified predominance of high-impact deviations in step M3_S2, which is responsible for formulating hypotheses, research questions, and objectives, requires special attention. This segment, as shown by the corpus analysis, is marked by the highest density of syntactically overloaded and declaratively redundant constructions that hinder the clear expression of the research aim. Such a concentration of overload points to rhetorical pressure that emerges when the author must simultaneously present the novelty

of the study and comply with established genre and institutional expectations. These observations are consistent with the findings of Alramadan (2020), who demonstrated in a corpus-based study of introductions in applied linguistics articles that the stage of presenting research contribution carries significant rhetorical weight and is often realized through formulaic or nominalized constructions, which reduce the cognitive clarity of the text. Therefore, step M3_S2 can be regarded as a key zone of communicative tension, where linguistic redundancy functions both as a means of institutional positioning and as a potential threat to the clarity of scientific argumentation.

With regard to the second hypothesis, which suggested that redundancy, although less frequent than wordiness, tends to result in more serious communicative distortions, the data call for a more nuanced interpretation. Wordiness was indeed more widespread in the corpus, particularly in the form of vague generalizations, hedging expressions and nominalizations. However, redundancy, which is also a manifestation of textual wordiness, was disproportionately associated with high-impact fragments. For example, structural redundancy accounted for 17.8 percent of all high-impact cases, which is noticeably higher than the share observed for most subcategories of wordiness. This indicates that redundancy presents a greater risk to the clarity and coherence of scientific argumentation, thus providing partial confirmation of the second hypothesis.

Linguistic overload in the analyzed corpus reflects not only stylistic but also cognitive factors. According to the cognitive model of writing developed by Flower and Hayes (1981), redundant structures interfere with the reader's ability to construct a stable mental representation of the text and increase cognitive load. This effect is particularly evident in the case of nominalizations and complex syntactic constructions, which make it more difficult to access the core meaning of the text (Graesser et al., 2003; Tikhonova et al., 2024b). The reader experiences greater cognitive strain because more informational links are required to maintain coherence and must be held in working memory.

It is also important to highlight the institutional dimension of the observed deviations. As van Dijk (2008) notes, stereotyped language patterns, formulaic phrases, and repetitive constructions serve not only to connect or emphasize ideas but also to signal the writer's affiliation with the academic community. From this perspective, linguistic overload can be seen as a reflection of genre-based socialization. By reproducing formal patterns, the writer internalizes the conventions of academic writing, even when such choices reduce the informational density of the text.

These findings also provide empirical support for the third hypothesis (H3), which posited that stylistic deviations are structurally motivated and shaped by genre-based expectations. The distribution of wordiness and redundancy across distinct rhetorical steps, along with the significantly higher impact observed in functionally loaded zones such as M3_S2 and M1_S2, confirms that linguistic overload arises not simply from stylistic carelessness but as a response to rhetorical demands embedded in academic writing conventions. Rather than being distributed evenly, deviations cluster around communicatively sensitive parts of the introduction, where authors face the task of justifying their research, identifying a niche, and articulating novelty. This suggests that redundancy operates as a genre-induced strategy for managing rhetorical pressure.

Thus, the discussion of results points to the dual nature of linguistic overload in academic introductions. On the one hand, it reflects a response to the functional and rhetorical challenges of the text. On the other hand, it results from institutionalized genre expectations. Taken together, these factors call for a shift among researchers and educators from prescriptive stylistics to a functional-rhetorical diagnostic approach. Within this framework, redundancy should be understood not simply as a deviation from an ideal linguistic norm, but as a symptom of rhetorical instability.

CONCLUSION

This study aimed to identify the structural patterns of linguistic overload in the introductions of academic articles in the field of education, with a focus on rhetorical positioning, the nature of stylistic deviations, and their communicative impact. The findings demonstrated that textual redundancy in academic writing is not a random deviation from the norm. Rather, it arises in response to rhetorical pressure that emerges in functionally important parts of the text, such as justifying relevance, framing the research gap, and formulating research objectives.

An analysis of frequency distributions, together with the results of chi-square tests, confirmed statistically significant relationships between the type of stylistic deviation, its rhetorical placement, and the level of communicative impact. This supports the conclusion that wordiness and redundancy are not merely stylistic excesses, but systematic rhetorical strategies used by authors in segments of high cognitive and communicative tension. Particularly vulnerable were Step M1_S2 (justification of relevance), M2_S1 (statement of the research gap), and M3_S2 (formulation of hypotheses), where a high concentration of fragments was found to potentially distort the clarity of meaning.

The practical value of this study lies in the development of a functionally oriented approach to diagnosing and editing academic texts. By identifying rhetorically vulnerable segments, it becomes possible to design targeted instructional strategies for teaching academic writing. These strategies are not aimed at eliminating redundancy as such but rather at optimizing the form of expressing scientific meaning in

alignment with the communicative goals of each section of the introduction.

At the same time, the interpretation of the results must take into account several limitations. First, the study corpus is restricted to a single discipline (education) and includes only Russian-language texts. This limits the generalizability of the findings to other academic cultures. Second, although the rhetorical annotation procedure was systematic, the assessment of communicative impact (IMPACT) involved a degree of expert interpretation and therefore requires further inter-rater validation.

Despite these limitations, the study provides empirical support for the concept of rhetorically motivated linguistic overload and lays the groundwork for further exploration of the relationship between the structure of scholarly discourse and communicative clarity. Promising directions for future research include (1) cross-disciplinary comparisons of rhetorical redundancy patterns, (2) analysis of linguistic overload in English-language publications by Russian authors, and (3) the development of algorithmic tools for automatically identifying rhetorically induced redundancy in support of scholarly writing and editorial workflows.

DECLARATION OF COMPETING INTEREST

None declared.

AUTHORS' CONTRIBUTION

Elena Tikhonova: conceptualization; data curation; formal analysis; methodology; project administration; visualization; writing – original draft; writing – review & editing.

Olga Zavolskaya: formal analysis; investigation; writing – original draft.

Nataliia Mekeko: data curation; visualization; writing – original draft.

REFERENCES

Alontseva, N. V., & Ermoshin, Y. A. (2019). The problem of language redundancy on the example of a scientific text. *RUDN Journal of Language Studies, Semiotics and Semantics, 10*(1), 129-140. https://doi.org/10.22363/2313-2299-2019-10-1-129-140

Alramadan, M. (2020). Stance and engagement: A corpus-based comparison of university students' and published writers' research article introductions in applied linguistics. *Journal of English for Academic Purposes, 44*, 100837. https://doi.org/10.1016/j.jeap.2020.100837

Aoulad-Ouda, M., & Chellaoui, S. (2023). *Hedging and politeness in Moroccan university students' academic writing: A case study of literature reviews*. *Arab World English Journal, 14*(2), 183–197. https://doi.org/10.24093/awej/vol14no2.13

Biber, D., & Gray, B. (2010). Challenging stereotypes about academic writing: Complexity, elaboration, explicitness. *Journal of English for Academic Purposes, 9*(1), 2–20. https://doi.org/10.1016/j.jeap.2010.01.001

Biesta, G., Takayama, K., Kettle, M., & Heimans, S. (2024). How 'academic' should academic writing be? Or: why form should follow function. Asia-Pacific Journal of Teacher Education, 52(2), 121–125. https://doi.org/10.1080/1359866X.2024.2324582

Çakir, A., Kuyurtar, D., & Balyer, A. (2024). The effects of the publish or perish culture on publications in the field of educational administration in Türkiye. *Social Sciences and Humanities Open, 9*, 100817. https://doi.org/10.1016/j.ssaho.2024.100817

Demir, C. (2019). Writing Intelligible English Prose: Conciseness vs. Verbosity. *Söylem Filoloji Dergisi, 4*(2), 482-505. https://doi.org/10.29110/soylemdergi.617184

Flower, L., & Hayes, J. R. (1981). *A cognitive process theory of writing*. *College Composition and Communication, 32*(4), 365–387. https://doi.org/10.2307/356600

Flowerdew, J., & Forest, R. W. (2015). *Signalling nouns in English: A corpus-based discourse approach.* Cambridge University Press. https://doi.org/10.1017/CBO9781139135405

Fruehwald, E. S. (2010). Exercises for legal writers II: Editing for wordiness. *SSRN Electronic Journal.* https://doi.org/10.2139/ssrn.1704045

Gong, N., & Barlow, J. (2022). Rhetorical practices and disciplinary identity in novice academic writers: Evidence from corpus-based analysis. *Journal of English for Academic Purposes, 59*, 101148. https://doi.org/10.1016/j.jeap.2022.101148

Gong, X., & Barlow, L. (2022). A corpus-based analysis of research article macrostructure patterns in high-impact journals. *Theory and Practice in Language Studies, 12*(6), 1067–1075. https://doi.org/10.17507/tpls.1206.12

Goonaratna, C. (2002a). Writing well (6) Wordiness, alias verbosity. *Ceylon Medical Journal, 47*(1), 1-3. https://doi.org/10.4038/cmj.v47i1.6393

Goonaratna, C. (2002b). Writing well (7) Wordiness alias verbosity, continued. *Ceylon Medical Journal, 47*(3), 79-80. https://doi.org/10.4038/cmj.v47i3.3432

Graesser, A. C., McNamara, D. S., & Louwerse, M. M. (2003). What do readers need to learn in order to process coherence relations in narrative and expository text. In A. P. Sweet & C. E. Snow (Eds.), *Rethinking Reading Comprehension* (pp. 82–98). Guilford Publications.

Hyland, K. (2005). *Metadiscourse: Exploring interaction in writing.* Continuum.

Hyland, K. (2008). Academic clusters: Text patterning in published and postgraduate writing. *International Journal of Applied Linguistics, 18*(1), 41–62. https://doi.org/10.1111/j.1473-4192.2008.00178.x

Hyland, K. (2008). Genre and academic writing in the disciplines. *Language Teaching, 41*, 543-562. https://doi.org/10.1017/S0261444808005235

Kanoksilapatham, B. (2005). Rhetorical structure of biochemistry research articles. *English for Specific Purposes, 24*(3), 269–292. https://doi.org/10.1016/j.esp.2004.08.003

Kravtchenko, E., & Demberg, V. (2022). Informationally redundant utterances elicit pragmatic inferences. *Cognition, 225*, 105159. https://doi.org/10.1016/j.cognition.2022.105159

Leufkens, S. (2023). Measuring redundancy: The relation between concord and complexity. *Linguistics Vanguard, 9*(s1), 95-106. https://doi.org/10.1515/lingvan-2020-0143

Raitskaya L.K., & Tikhonova E.V. (2019). Multilingualism in Russian journals: A controversy of approaches. *European Science Editing, 45*(2), 41. https://doi.org/10.20316/ESE.2019.45.18024

Raitskaya, L., & Tikhonova, E. (2020). Overcoming cultural barriers to scholarly communication in international peer-reviewed journals. *Journal of Language and Education*, *6*(2), 4-8. https://doi.org/10.17323/jle.2020.11043

Salager-Meyer, F. (1994). Hedges and textual communicative function in medical English written discourse. *English for Specific Purposes, 13*(2), 149–170. https://doi.org/10.1016/0889-4906(94)90013-2

Samraj, B. (2002). Introductions in research articles: Variations across disciplines. *English for Specific Purposes, 21*(1), 1–17. https://doi.org/10.1016/S0889-4906(00)00023-5

Smirnova, N. V., Lillis, T., & Hultgren, A. K. (2021). English and/or Russian medium publications? A case study exploring academic research writing in contemporary Russian academia. *Journal of English for Academic Purposes, 53*, Article. 101015. https://doi.org/10.1016/j.jeap.2021.101015

Strunk, W., & White, E. B. (2000). *The elements of style* (4th ed.). Pearson Education.

Swales, J. M. (1990). *Genre analysis: English in academic and research settings.* Cambridge University Press.

Swales, J. M. (2004). *Research genres: Explorations and applications.* Cambridge University Press.

Tikhonova E.V., Kosycheva M.A., & Mezentseva D.A. (2024b). Ineffective strategies in scientific communication: textual wordiness vs. clarity of thought in thesis Conclusion section. *Integration of Education, 28*(2), 249-265. https://doi.org/10.15507/1991-9468.115.028.202402.249-265

Tikhonova, E. V., & Mezentseva, D. A. (2024). Wordiness in academic writing: A systematic scoping review. *Research Result. Theoretical and Applied Linguistics, 10*(1), 133–157. https://doi.org/10.18413/2313-8912-2024-10-1-0-8

Tikhonova, E., Mezentseva, D., & Kasatkin, P. (2024a). Text redundancy in academic writing: A scoping review. *Journal of Language and Education, 10*(3), 128–160. https://doi.org/10.17323/jle.2024.23747

van Dijk, T. A. (2008). *Discourse and Power*. Palgrave Macmillan. https://doi.org/10.1007/978-1-137-07299-3

Wahyuningtyas, L., & Wulandari, F. (2023). A comparative study of rhetorical structures in research article introductions: Social sciences versus education disciplines. *Journal of Language and Linguistic Studies, 19*(1), 333–348.

Williams, J. M., & Bizup, J. (2017). *Style: Lessons in clarity and grace* (12th ed.). Pearson Education.

https://doi.org/10.17323/jle.2025.21793

Exploring the Use of ChatGPT in EFL/ESL Writing Classrooms: A Systematic Review

Yustinus Calvin Gai Mali

Universitas Kristen Satya Wacana, Indonesia

ABSTRACT

Background: ChatGPT has become increasingly prevalent in higher education, particularly within EFL/ESL writing classrooms. However, the rise in plagiarism and academic dishonesty associated with its unethical use is concerning. Educational institutions must explore and design AI-use-related best practices for using generative AI technology, such as ChatGPT, more ethically in the writing classrooms.

Purpose: To systematically review previous studies to investigate how university students use ChatGPT in their EFL/ESL writing classrooms. Given the evidence of how the students used ChatGPT, this study explores existing best practices to regulate ChatGPT's ethical and responsible use in the classes.

Method: Thirty-two (32) articles (i.e., 17 empirical and 15 non-empirical studies) from 31 peer-reviewed international journals were selected based on specific criteria comprising article types, quality, year of publication content, and contexts of the study, following the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) guidelines. The articles were searched in May 2024, facilitated by the Publish or Perish software. Within the software, Google Scholar was deliberately chosen as the primary database. The inductive data analysis results were rigorously checked using multiple validation strategies and presented as themes to address the research goal.

Results: The analysis revealed that ChatGPT was utilized in the writing process for various academic and non-academic writing tasks, highlighting the potential bright and dark sides of ChatGPT's use in writing. The study then identified four main categories of regulating the use of ChatGPT in EFL/ESL writing classrooms, which include institutional policies, instructional writing strategies, assessment design innovation, and ethical co-regulation practices. Drawing on the analyses and discussions of the previous studies, the researcher suggested sample writing activities with the ethical and productive use of ChatGPT, outlined pedagogy and policy implications for regulating ChatGPT in the writing classrooms, and proposed directions for future research.

Conclusion: Key patterns in how EFL/ESL learners have used ChatGPT in writing tasks and eight replicable best practices for regulating its use in classroom contexts were identified, where among these, co-creating ethical guidelines with students and emphasizing the writing process seemed to be particularly promising strategies to mitigate the unethical use of ChatGPT in EFL/ESL writing classrooms.

KEYWORDS:

ChatGPT, EFL/ESL, literature review, students, teachers, writing

Citation: Mali, Y. C. G. (2025). Exploring the use of ChatGPT in EFL/ESL writing classrooms: A systematic review. *Journal of Language and Education, 11*(2), 137-156. https://doi.org/10.17323/jle.2025.21793

Correspondence:
Yustinus Calvin Gai Mali
yustinus.mali@uksw.edu

Received: June 16, 2024
Accepted: June 10, 2025
Published: June 30, 2025

INTRODUCTION

The emergence of ChatGPT, a large language model developed by OpenAI and publicly released in November 2022, has brought significant transformation to the landscape of education, particularly within English language teaching and learning (Imran & Almusharraf, 2023; Mahyoob et al., 2023; Synekop et al., 2024). As early as 2022, Sumakul et al. (2022) anticipated that artificial intelligence (AI) would soon become a major force in education, and this projection

has rapidly materialized with the widespread adoption of ChatGPT. Trained on extensive textual data, ChatGPT is capable of generating human-like responses, which has drawn considerable attention from researchers for its potential applications in academic contexts, including student writing (Vázquez-Cano et al., 2023; Klyshbekova & Abbott, 2024).

Despite its potential to support students' writing processes and improve writing fluency (Rababah et al., 2023; Song & Song, 2023), ChatGPT also raises pressing concerns in educational settings. Notably, academic staff find it increasingly difficult to differentiate between student-written texts and those generated by ChatGPT (Hang, 2023; Krajka & Olszak, 2024; Matthews & Volpe, 2023), and there is growing documentation of plagiarism and academic misconduct associated with its misuse (Alberth, 2023; Grassini, 2023; Perkins, 2023). In this context, scholars emphasize the necessity for institutions to formulate clear, pedagogically grounded strategies to regulate generative AI tools in the classroom (Gustilo et al., 2024), especially as evidence suggests that students will continue to use ChatGPT as an integrated part of their academic experience (Ajlouni et al., 2023; Ho & Nguyen, 2024; Marzuki et al., 2023; Nguyen et al., 2024). As Alqasham (2023) notes, ChatGPT is likely to become as ubiquitous as Google. Without formal guidance, students cannot be expected to independently navigate the ethical boundaries of AI-assisted writing (Črček & Patekar, 2023).

While recent research has sought to identify best practices and policy recommendations for using ChatGPT in education, most of these discussions remain general in scope and are not situated within the specific context of English writing instruction (Ajlouni et al., 2023; Crawford et al., 2023; Grassini, 2023; Matthews & Volpe, 2023; Rudolph et al., 2023; Tikhonova & Raitskaya, 2023). Furthermore, many studies have been conducted within single national contexts: the Philippines (Gustilo et al., 2024), Croatia (Črček & Patekar, 2023), or Jordan (Rababah et al., 2023), limiting the generalizability of their findings.

In response to these limitations, the present study undertakes a focused review of empirical and conceptual research to examine how university students in EFL/ESL writing classrooms have used ChatGPT. Based on this analysis, the study then identifies and synthesizes replicable best practices for regulating its use in pedagogically appropriate and ethically responsible ways.

The study is guided by the following research questions:

RQ1: How do university language students use ChatGPT in their EFL/ESL writing classrooms?

RQ2: What are the existing best practices to regulate the use of ChatGPT in EFL/ESL writing classrooms?

In addressing these questions, this review responds to recent calls in the literature (e.g., Alqasham, 2023; Baskara & Mukarto, 2023; Cong-Lem et al., 2024; Klimova et al., 2024) for the development of comprehensive guidelines that promote responsible AI use while reinforcing academic integrity in EFL/ESL writing contexts. The results are also intended to support instructors who remain uncertain about how to address AI-related academic dishonesty in their classrooms (Cong-Lem et al., 2024). Additionally, the study proposes examples of ethically grounded writing tasks that incorporate ChatGPT as a support tool rather than a substitute for student work.

METHOD

This study adopts a systematic literature review method by Li (2018) and Zain (2022) to explore previous studies that address the use of ChatGPT in English writing classrooms in EFL/ESL contexts. In short, the term EFL shows contexts where people mostly learn English in a formal classroom setting, with limited opportunities to use the language outside their class for daily communication (Mali, 2017), while the term ESL shows contexts where people learn English in a place where English is necessary and plays important roles in everyday life, education, business, and government (Richards & Schmidt, 2010).

Transparency and Databases

The review followed an a priori literature search and data extraction protocol, and no deviation from the protocol was made during the study. To identify relevant sources for analysis, the researcher employed the Publish or Perish software, a bibliometric tool previously utilized in systematic review studies such as Putrie et al. (2024). Within the software, Google Scholar was deliberately chosen as the primary database, following the approach of Li (2018), due to its accessibility and inclusion of a wide range of open-access journal articles, in contrast to subscription-based platforms such as Web of Science. The search strategy involved the use of targeted keyword phrases, including: *"ChatGPT in English language writing class," "ChatGPT in EFL writing class," "guidelines for using ChatGPT in EFL/ESL writing classrooms," "policy of using ChatGPT in EFL/ESL writing class,"* and *"students' experiences in using ChatGPT in writing class."* The search for the articles was done in May 2024. A more detailed protocol used for the literature search is explained in the following section.

Inclusion and Exclusion Criteria

Articles retrieved from the search were then further selected based on the following inclusion and exclusion criteria of the Preferred Reporting Items for Systematic Reviews and

Meta-Analyses (PRISMA) guidelines, used by Teng (2024a) to ensure the quality and novelty of the review. First, the articles should be written in English and peer-reviewed academic articles. Second, the selected articles can be both empirical and non-empirical studies. Still, they should discuss the use of ChatGPT in English language writing classrooms in EFL/ESL contexts in higher education settings. Third, the articles were recently published in 2024-2023. Fourth, the articles are published in peer-reviewed SCOPUS-indexed journals. Fifth, if not indexed by SCOPUS, the articles should be cited at least ten times by other studies. Last, the selected articles should be open-access. The researcher was aware that the criteria might be influenced by the potential subjective bias of the researcher, yet most of the criteria in Table 1 were also used by the previous systematic review studies (e.g., Teng, 2024a; Zain, 2022). Thus, the criteria should remain relevant for future studies, with necessary modifications in certain areas, e.g., the number of citations of the articles to include in the review, to enhance the quality of the review.

In addition to studies situated within EFL/ESL contexts, the review also included selected articles from broader higher education settings (e.g., Črček & Patekar, 2023; Yeo, 2024) given the relevance of their findings and the value of the authors' perspectives in addressing the present study's research questions. To respond to reviewer feedback on an earlier version of the manuscript, a complementary manual search was conducted using Google Scholar, guided by the inclusion criteria outlined in Table 1. This process yielded seven additional studies deemed relevant to the study's objectives and capable of enriching the analysis of ChatGPT use in EFL/ESL writing classrooms. The manual search also served as a strategy to minimize the risk of omitting significant works that may not have been retrieved through the Publish or Perish software alone. As a result, 32 peer-reviewed journal articles (17 empirical and 15 non-empirical studies) were selected based on the established inclusion and exclusion criteria. As detailed in Table 2, the majority of these articles were published in education-focused journals, with a smaller subset appearing in journals related to educational technology.

Data Analysis Procedures

To analyze the selected studies, the researcher adopted an inductive, qualitative content analysis approach, following procedures commonly used in systematic and narrative reviews (e.g., Li, 2018; Teng, 2024a; Zain, 2022). The analysis began with multiple close readings of each article, during which the researcher identified and highlighted segments of text deemed relevant to the study's research questions, as guided by Saldaña and Omasta's (2018, p. 286) recommendation to extract information that is "closely related to and can help to answer the research questions." Following the initial reading phase, the researcher created a structured document in Microsoft Word to compile annotations, organize highlighted excerpts, and assign preliminary codes to specific data segments. These initial codes served as analytic labels that captured patterns across the data. Through an iterative review process, codes were examined, compared, and refined to generate a set of emergent themes, largely derived using in vivo phrasing from the original notes. These themes were then used to organize the presentation of findings in the Results section, each directly addressing one or both of the study's research questions.

It should be noted that the data were analyzed by a single coder, which may limit intersubjective reliability. This decision is consistent with other single-author review studies (Li, 2018; Teng, 2024a; Zain, 2022); however, it remains a methodological limitation. In the absence of coder triangulation, transparency was prioritized in the coding and theme development process to enable readers to assess the coherence and validity of the interpretations presented. That said, the researcher developed a data extraction table (see Table 3) to systematically record key information from each reviewed study. For each reviewed article, the researcher documented the methodological type, study contexts, ChatGPT's use case for writing practices and policies, and relevance to the

Table 1

Inclusion and Exclusion Criteria

Criteria	Inclusion	Exclusion
Language	English	Non-English
Document	Journal articles	Non-journal articles
Article types	Peer-reviewed articles	Non-peer reviewed articles
Content	Discussing the use of ChatGPT in English writing class	Not discussing the use of ChatGPT in English writing class
Context of study	EFL/ESL contexts in higher education settings	Non-EFL/ESL contexts; non-higher education settings
Year of publication	2024-2023	Before 2023
Quality	SCOPUS-indexed or has been cited at least ten times	Non-SCOPUS indexed or has been cited less than ten times
Access	Open-access	Non-open access; needs to pay to access articles

Table 2

Distribution of the Reviewed Articles

Journal Titles	Number of Articles	Studies	Types
Arab World English Journal	1	Algaraady and Mahyoob (2023)	Empirical Studies
Asia CALL Online Journal	1	Schmidt-Fajlik (2023)	
Cogent Education	1	Marzuki et al. (2023)	
Computers and Education: Artificial Intelligence	1	Teng (2024b)	
Electronic Journal of e-Learning	1	Tseng and Lin (2024)	
Frontiers in Psychology	2	Klimova et al. (2024); Song and Song (2023)	
International Journal of Engineering Pedagogy	1	Rababah et al. (2023)	
International Journal of Language Instruction	1	Hang (2023)	
Journal of Applied Learning & Teaching	1	Mohammadkarimi (2023)	
Journal of Language & Education	1	Črček and Patekar (2023)	
Journal of University Teaching & Learning	1	Ho and Nguyen (2024)	
Languages	1	Xiao and Zhi (2023)	
Migration Letters	1	Alqasham (2023)	
Research Methods in Applied Linguistics	1	Mizumoto and Eguchi (2023)	
Smart Learning Environment	1	Özçelik and Ekşi (2024)	
Teaching English with Technology	1	Cong-Lem et al. (2024)	
XLinguae	1	Krajka and Olszak (2024)	
Assessing writing	1	Barrot (2023)	Non-Empirical Studies
Computers and Education Open journal	1	Moorhouse et al. (2023)	
Contemporary Educational Technology	1	Imran and Almusharraf (2023)	
Indonesian Journal of English Language Teaching and Applied Linguistics	1	Baskara and Mukarto (2023)	
Indian Journal of Ophthalmology	1	Mondal and Mondal (2023)	
International Journal of Education and Learning	1	Baskara (2023)	
International Journal of TESOL Studies	1	Teng (2024a)	
Journal of China Computer-Assisted Language Learning	1	Tseng and Warschauer (2023)	
Languages	1	Zadorozhnyy and Lai (2023)	
RELC Journal	1	Yeo (2024)	
Teaching English with Technology	1	Bonner et al. (2023)	
TEFLIN Journal	1	Alberth (2023)	
TESL-EJ	1	Kostka and Toncelli (2023)	
TESOL Journal	1	Carlson et al. (2023)	

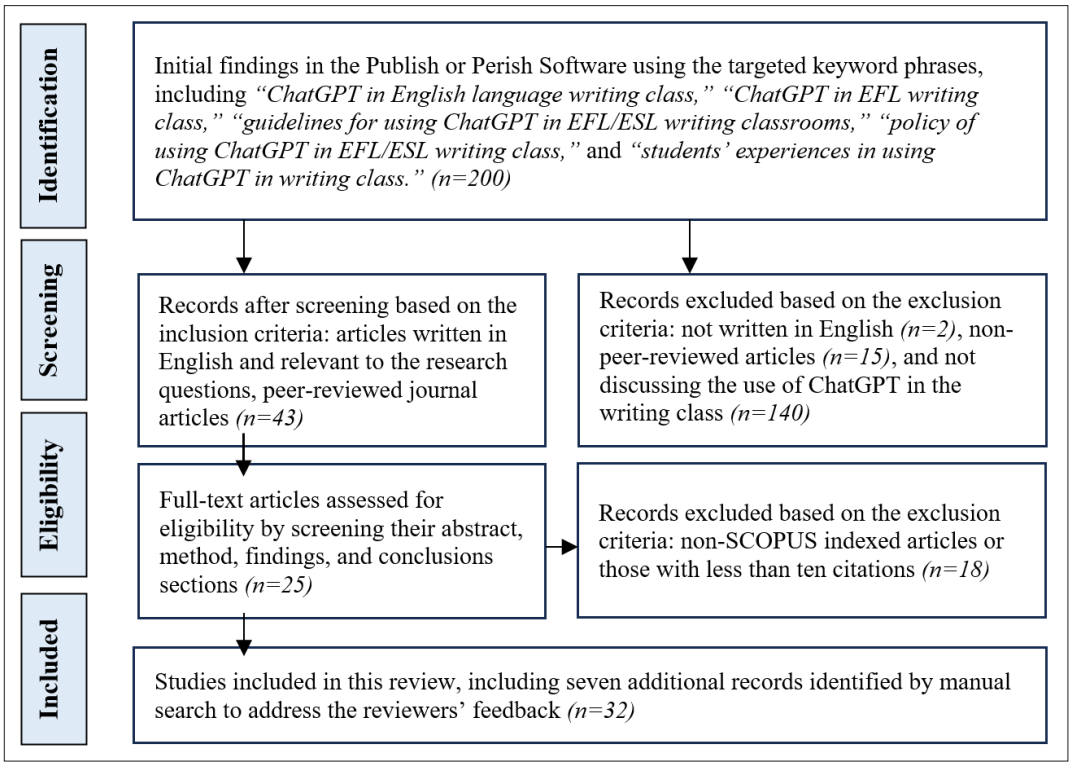
research questions. The table also included a column for initial coding and interpretive notes, which later informed theme development. This structured approach ensured consistency across articles and enabled a transparent link between data and emerging insights.

Data Evaluation

Although the inductive data analysis, including the coding and synthesis process, was conducted by the researcher himself, the researcher also employed multiple validation

Figure 1

PRISMA Flow Diagram for the Systematic Review in This Study



strategies to enhance the credibility of the findings presented in this study. First, the researcher used the inclusion and exclusion of studies based on transparent and replicable criteria, as presented in Table 1, which was similar to previous studies. Second, the researcher maintained a reflective, analytic memo throughout the coding process to document interpretive decisions and ensure consistency. Third, the researcher carefully checked all data extraction and theme development with the original studies to maintain fidelity to source meanings. Fourth, the researcher encouraged readers of this paper to consult the data extraction table (see Table 3), which provided detailed information on how each reviewed study contributed to the findings. These strategies align with recognized approaches for establishing trustworthiness in qualitative research (Poveda-Garcia-Noblejas & Antropova, 2024; Tikhonova & Raitskaya, 2024).

RESULTS

This section presents the systematic review results of the two research questions. It also summarizes the scope, research designs, geographical coverage, and quality of the included studies and highlights key trends, gaps, and areas of divergence across the literature. The researcher retrieved 200 studies as the initial search results. The researcher then screened articles written in English and peer-reviewed, as well as those relevant to the research questions. As a result, the researcher excluded 157 articles following the criteria defined in the protocol. The researcher then performed a full-text reading of the remaining 43 studies. Among these, 18 articles that did not fulfill the inclusion criteria were removed, such as non-SCOPUS-indexed articles or those with fewer than ten citations. Finally, this study reviewed the remaining 25 studies plus seven other studies identified by manual search. As shown in Table 3, the study finally reviewed 32 articles in total.

Study Characteristics

The matrix in Table 3 provides a general overview of the reviewed studies. First, among the thirty-two articles, 17 are empirical studies, while the other 15 are non-empirical. Second, most studies (n=23) were published in 2023, while a few (n=9) were published in 2024. Though these studies are very recent, they were commonly conducted in a single country or site and aimed to explore ChatGPT in isolation without comparing it with other recently published AI tools, such as Google AI Studio or DeepSeek. Third, all the reviewed studies were conducted in higher education contexts, either in private or state universities, involving university lecturers and students. Fourth, methodologically speaking, previous studies (n=7) commonly used the qualitative method and selected interviews as their primary research instrument. Yet, these qualitative studies involved a limited number of research participants, such as four lecturers (e.g., Marzuki et al., 2023) and five students (e.g., Xiao & Zhi, 2023), which might limit the generalizability of their findings. The

Table 3

Matrix of the Previous Studies

Study	Research/ Study Goals	Contexts and Participants	Designs	Uses of ChatGPT	Policies Discussed
Alberth (2023)	Examine the potential benefits and drawbacks of ChatGPT in academic writing, as well as provide solutions to address the drawbacks	Academic writing contexts	Conceptual paper	To provide feedback on essay structure, sentence clarity, coherence, and cohesion in academic essays	Acknowledging how ChatGPT was used in the writing process (e.g., to paraphrase some ideas); explicitly explaining what students may or may not do when using ChatGPT in their writing class
Algaraady and Mahyoob (2023)	Investigate ChatGPT's effectiveness in detecting writing errors made by EFL students	88 EFL university written tasks in a university in Yemen	A mixed-method study involving error analysis done by EFL instructors and ChatGPT	To provide feedback on grammar, vocabulary, spelling, punctuation, and sentence structure	Not discussed
Alqasham (2023)	Explore EFL university perceptions of ChatGPT in their English language acquisition journey	5 undergraduate students from various majors at an English-medium college in Saudi Arabia	A qualitative study using a semi-structured interview	To be a digital tutor offering various feedback to students' academic essays, helping students brainstorm writing ideas, and drafting their introductions	Acknowledging and citing ChatGPT's contribution appropriately, incorporating training sessions highlighting the potential and limitations of AI tools
Barrot (2023)	Explore the potential benefits and challenges of using ChatGPT for second language (L2) writing	L2 writing classroom practices	Technology review (conceptual paper)	To brainstorm writing ideas, to refine students' initial writing outline, to provide writing feedback on language style, vocabulary, and grammar	Emphasizing the value of the writing process, asking students to document their writing process, encouraging students to incorporate personal experiences in their writing, and asking students to write their original output first and then refine it using ChatGPT
Baskara (2023)	Explore potential benefits and challenges of using ChatGPT in EFL writing instruction	EFL writing instructions	A literature review study	To provide writing feedback, which includes vocabulary, grammar, and syntax, and suggestions for improvement	Assigning writing tasks that require a high level of creativity or originality, co-creating ChatGPT guidelines with students, providing students with examples of how to use ChatGPT for writing practice, and being prepared to provide support and guidance to students as needed
Baskara and Mukarto (2023)	Explore the current knowledge of ChatGPT and its potential implications of ChatGPT for language learning in higher education	Higher education contexts	A literature review study	To provide feedback on students' writing	Creating new pedagogical methods or assessment techniques in writing classes

Study	Research/ Study Goals	Contexts and Participants	Designs	Uses of ChatGPT	Policies Discussed
Bonner et al. (2023)	Provide examples of how ChatGPT can be used to develop learning materials and classroom activities as well as to provide feedback	School and higher education contexts	Conceptual paper	To check grammatical errors in students' writing	Not discussed
Carlson et al. (2023)	Examine the use of ChatGPT 4 to provide feedback on students' writing	General English for academic purposes classes	Media review	To give feedback on the quality of a topic sentence, grammatical accuracy, ideas development, and language quality	Practicing ChatGPT prompts, having the writing done in class, as well as using AI detectors to review students' work
Cong-Lem et al. (2024)	Explore EFL lecturers' perceptions and responses to academic integrity in the era of ChatGPT	25 EFL university lecturers in Vietnam	A qualitative study using a structured, open-ended survey	To generate essays for students	Teaching students about how to use AI tools appropriately, raising students' awareness of the importance of their original ideas, using AI detection tools, redesigning writing assessments, co-creating AI guidelines with students, using stricter regulations, using offline assessment
Črček and Patekar (2023)	Investigate the use of ChatGPT among university students for written assignments, explore ways the students utilize the tool, and explore the students' perspectives on the ethical aspects of using ChatGPT	201 university students from private and public universities in Croatia	A quantitative study using an online questionnaire	To generate writing ideas, paraphrase sentences, summarize, write parts or whole of students' writing assignments	Banning the use of AI, fostering the spirit of *honesty-humility* to students (i.e., prioritizing fairness over students' interests), *conscientiousness* (i.e., a strong work ethic), and *openness to experience* (i.e., tackling challenges with students' ideas)
Hang (2023)	Explore EFL lecturers' thoughts on the use of ChatGPT in their writing classes	20 EFL university instructors at a university in Vietnam	A mixed-method study using a questionnaire and structured interview	To be a supportive tutor, providing suggestions for revising students' work, writing samples, and examples of language use	Designing writing activities that require students to use their critical thinking and problem-solving skills; explaining where, when, and how ChatGPT is or is not allowed to be used in the writing process
Ho and Nguyen (2024)	Explore students' perceptions on the use of ChatGPT in English language learning	369 English-majored students at a university in Vietnam	A quantitative study using a questionnaire	To provide feedback on students' writing and language use	Adding courses on how to integrate technology in general and ChatGPT into teacher education programs
Imran and Almusharraf (2023)	Examine the role of ChatGPT as a writing assistant in academia	Higher education contexts	A literature review study	To assist the writing process of a scientific paper	Being fully aware of the limitations of ChatGPT

Study	Research/ Study Goals	Contexts and Participants	Designs	Uses of ChatGPT	Policies Discussed
Klimova et al. (2024)	Explore students' attitudes and perceived usefulness of using ChatGPT for learning a foreign language	91 undergraduate students studying English at a university in the Czech Republic	A qualitative study using a questionnaire survey	To help students in writing their bachelor thesis, providing references to cited sources, creating presentations, and writing seminar papers	Establishing guidelines and protocols to ensure responsible and ethical use of AI technologies, particularly regarding issues of plagiarism, data privacy, and academic integrity; upskilling teachers and students' AI competencies
Kostka and Toncelli (2023)	Explore the roles of ChatGPT in English language teaching, its benefits, and challenges, as well as describe best practices of using ChatGPT for language teaching purposes	University settings in the USA	Conceptual paper	To correct grammatical errors and provide explanations for the corrections	Assigning students writing assignments that require problem-solving and critical thinking, using Turnitin software to detect AI-generated texts, co-creating guidelines for ChatGPT use with students, and refining them as the semester goes on (if necessary)
Krajka and Olszak (2024)	Investigate university students' experiences in determining if AI tools or a human wrote an essay, as well as see how much AI assistance they need to summarize, generate text, and write from prompts when trained to use an AI-assisted word processor	24 undergraduates in the applied linguistics study program at a university in Poland	A quasi-experimental treatment involving a single group	To provide feedback on students' writing	Using AI detection tools, understanding the differences between AI and human-generated written texts
Marzuki et al. (2023)	Examine the range of available AI writing tools and assess their influence on student writing	4 EFL lecturers from three different universities in Indonesia	A qualitative study using interviews	To brainstorm and generate writing ideas	Not discussed
Mizumoto and Eguchi (2023)	Perform automated essay scoring using GPT-3 (text Davinci-003 model) and evaluate its reliability and accuracy	12.100 essays in the ETS Corpus of Non-Native Written English	Scoring the essays using the GPT-3	To score an essay and explain the reasons for giving that particular score	
Mohammadkarimi (2023)	Examine EFL lecturers' perspectives on academic dishonesty made by EFL university students in the era of AI	67 EFL lecturers from various state and private universities in Iraqi Kurdistan	A mixed-method study using a semi-structured interview and a questionnaire	To deliver fast writing feedback	Using AI detection tools, discussing with students what they may or may not do with ChatGPT when writing, designing writing tasks requiring creativity, originality, problem-solving, and critical thinking
Mondal and Mondal (2023)	Explores the use of ChatGPT in academic writing and provides insights on how to utilize it judiciously	Academic writing in higher education	Conceptual paper	To assist students in academic writing (e.g., checking grammar and summarizing journal articles)	Properly citing all sources and avoiding copying and pasting text directly from ChatGPT without proper attribution, recognizing the potential risks of using ChatGPT

Study	Research/ Study Goals	Contexts and Participants	Designs	Uses of ChatGPT	Policies Discussed
Moorhouse et al. (2023)	Examine the extent to which the world's 50 top-ranking HEIs have developed or modified their assessment guidelines to address generative artificial intelligence (GAI) use and, where guidelines exist, the primary content and advice given to guide instructors in their GAI assessment design and practices	Websites of the top (based on Times Higher Education (THE) World University Rankings 2023) 50 universities' official websites	Analyzing information on the websites	To provide meaningful feedback on students' writing	Submitting notes they took on any sources to prepare their writing, providing alternative ways for students to demonstrate what they have learned beyond the text, breaking larger writing assignments into smaller pieces of writing tasks, and discussing with students what they may or may not do when using ChatGPT in their writing classes
Özçelik and Ekşi (2024)	Examine the impact of ChatGPT on the acquisition of register knowledge across various writing tasks	11 EFL university students at a university in Turkey	A qualitative case study with field notes and unstructured open-ended interviews	To correct the grammar, punctuation, and sentence structure of students' writing	Giving instructions on how to ask ChatGPT to edit their writing, including ChatGPT's suggestions made to students' writing when students submitted their final writing
Rababah et al. (2023)	Analyze perspectives of postgraduate university students about the use of ChatGPT in writing their theses	80 postgraduate students at a university in Jordan	A quantitative study using a questionnaire	To support the thesis writing process (i.e., searching relevant materials and generating ideas)	Emphasizing the importance of proper citation practices
Schmidt-Fajlik (2023)	Compare ChatGPT with online grammar checkers to check students' work	69 university students in Japan	Comparative study coupled with a questionnaire to explore students' perspectives	To provide clear and direct explanations related to grammar errors, ChatGPT can have the explanations translated into students' L1	Closely monitoring the writing process and scoring that process, using the process writing approach
Song and Song (2023)	Evaluate the impact of AI-assisted language learning on EFL students' writing skills and writing motivation	50 EFL students enrolled in a Bachelor's degree program at a university in China	A mixed-method study involving pre-and post-tests to assess writing skills and semi-structured interview	To identify grammatical errors in students' writing, provide feedback on essay structure, vocabulary, sentence clarity, and coherence in writing	Being fully aware of the limitations of ChatGPT
Teng (2024a)	Investigate the role of ChatGPT in EFL writing	EFL writing classrooms	A literature review study	To spot and analyze writing errors	Establishing clear guidelines and best practices for the ethical use of AI in writing, banning the use of ChatGPT, and creating new pedagogical methods or assessment techniques in writing classes
Teng (2024b)	Explore students' perceptions and experiences in using ChatGPT for their writing process in an EFL context	45 EFL students in Macau	A mixed-method study using questionnaires and interviews	To check grammatical errors in students' writing, provide instant and personalized writing feedback	Being fully aware of the limitations of ChatGPT

Study	Research/ Study Goals	Contexts and Participants	Designs	Uses of ChatGPT	Policies Discussed
Tseng and Lin (2024)	Explore students' reflections on using ChatGPT in their writing process	15 EFL university students at a private university in Taiwan	A qualitative study analyzing students' written works and reflective writings	To be a virtual peer reviewer that provides immediate feedback on grammar, essay structure, clarity, and coherence of students' writing; to act as a writer who can write a well-structured composition	Telling students what they can/cannot do when using ChatGPT in writing
Tseng and Warschauer (2023)	Propose a five-part pedagogical framework to use AI tools effectively	Second language learning and writing instructions	Conceptual paper	To help with spelling and grammar checks and paraphrasing suggestions	Teaching students how to note and cite the role of AI-based tools in their writing process
Xiao and Zhi (2023)	Investigate students' experiences with ChatGPT and perceptions about ChatGPT's role in language learning	5 undergraduate students from diverse majors at an English-medium international university in China	A qualitative study using a semi-structured interview	To generate new ideas when planning or writing an essay	Telling students to report how they used ChatGPT in completing their writing tasks, guiding students to use ChatGPT legitimately and productively, verifying information generated by ChatGPT
Yeo (2024)	Use ChatGPT to write an editorial for the RELC Journal	Academic journals	Editorial writing	To assist an author in writing an editorial	Publicly disclosing which AI tool was used, how it was used (including the prompts), and why it was used
Zadorozhnyy and Lai (2023)	Explore the potential benefits and roles of ChatGPT to enhance second language communicative practice	School and higher education contexts	Conceptual paper	To provide input for students' writing to enhance its sentence structure, grammar, and spelling	Not discussed

other studies used the mixed (n=5) and quantitative (n=3) methods. Two distinct methods, i.e., a comparative study to compare ChatGPT with online grammar checkers to review students' work (see Schmidt-Fajlik, 2023) and content analysis of websites' information related to the AI guidelines (see Moorhouse et al., 2023) were also used. Many are conceptual papers (n=6) and literature review studies (n=4). The other studies were related to media review (Carlson et al., 2023), technology review (Barrot, 2023), and editorial writing (Yeo, 2024), which might lack of empirical validation. The matrix also informs that ChatGPT has been used in many countries, heavily centered in Asia and the Middle East. Those countries are Japan (Schmidt-Fajlik, 2023), Taiwan (Tseng & Lin, 2024), Indonesia (Marzuki et al., 2023), China (Song & Song, 2023; Xiao & Zhi, 2023), Jordan (Rababah et al., 2023), Croatia (Črček & Patekar, 2023), Yemen (Algaraady & Mahyoob, 2023), Macau (Teng, 2024b), Vietnam (Cong-Lem et al., 2024; Hang, 2023; Ho & Nguyen, 2024), Iraqi Kurdistan (Mohammadkarimi, 2023), and Saudi Arabia (Alqasham, 2023). Some other countries include the Czech Republic (Klimova et al., 2024), Turkey (Özçelik & Ekşi, 2024), Poland (Krajka & Olszak, 2024), and the United States (Kostka & Toncelli, 2023).

RQ 1: The Use of ChatGPT in English Language Writing Classrooms

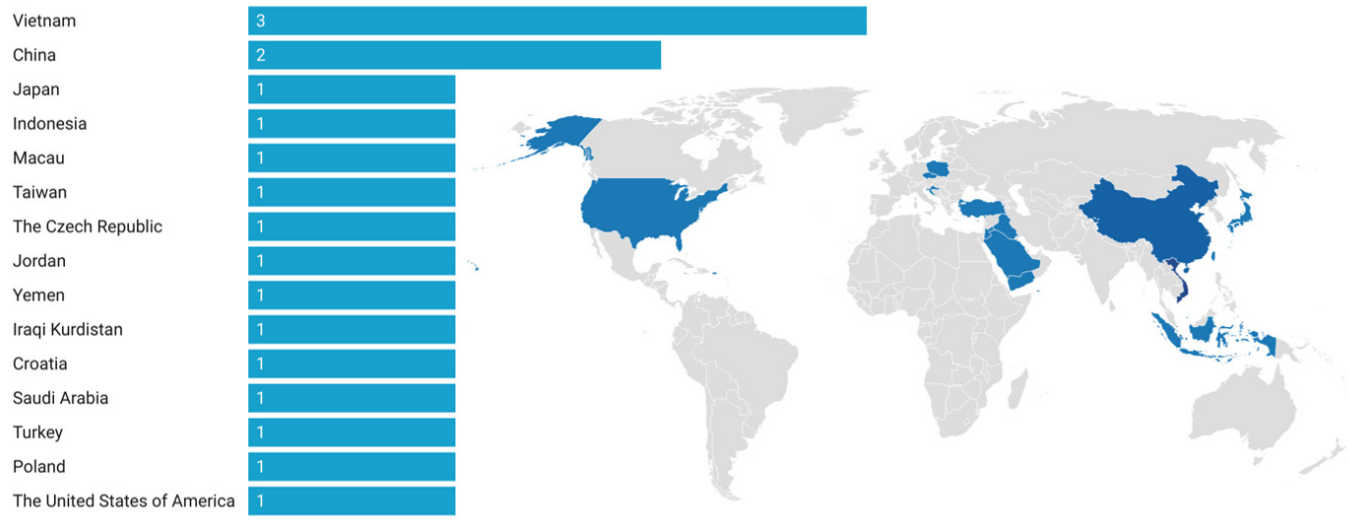
The review reported various students' practices using ChatGPT in English language writing classrooms. As a note, not all previous studies explicitly stated the types of ChatGPT used by university students. There are only a few researchers (e.g., Alqasham, 2023; Bonner et al., 2023; Özçelik & Ekşi, 2024; Tseng & Lin, 2024; Xiao & Zhi, 2023; Zadorozhnyy & Lai, 2023) who stated using the free access ChatGPT 3.5 version in their studies. Thus, this study assumes that the students reported in the previous studies used the ChatGPT 3.5 version.

Different Types of Writing

Students used ChatGPT to write various types of written work. These include seminar papers (Klimova et al., 2024), argumentative essays (Song & Song, 2023), thesis (Klimova et al., 2024; Rababah et al., 2023), and editorial writing (Yeo, 2024). The other types include academic (Alberth, 2023; Alqasham, 2023) and non-academic (Xiao & Zhi, 2023) es-

Figure 2

The Locations of the ChatGPT Use Informed by The Previous Studies



Note. This figure was created on the Datawrapper website (https://www.datawrapper.de/).

says, scientific papers (Imran & Almusharraf, 2023; Mondal & Mondal, 2023), an email, a blog post, a letter of request, and an informal text message (Özçelik & Ekşi, 2024).

Generating Writing Ideas

ChatGPT was also used to generate ideas for writing (Alqasham, 2023; Črček & Patekar, 2023; Marzuki et al., 2023; Rababah et al., 2023). This is evidenced by Xiao and Zhi's (2023) study that reported the voices of two students. The students said: "For example, if I plan to write an essay on a specific topic. I will ask ChatGPT to give me some ideas about where to start (student 1); When I do not know where to start, I will ask ChatGPT to think of several topics for me; then, I will take a look to see which one I am interested in (student 2)". For another student, "ChatGPT is a beacon when I'm grappling with ideation. If I'm set on an essay theme, I typically consult ChatGPT for initial thoughts or potential starting points" (Alqasham, 2023, p. 1256).

Asking for Feedback

Many previous studies (e.g., Baskara, 2023; Bonner et al., 2023; Carlson et al., 2023; Özçelik & Ekşi, 2024; Schmidt-Fajlik, 2023; Song & Song, 2023; Teng, 2024b; Tseng & Lin, 2024; Zadorozhnyy & Lai, 2023) reported that students asked ChatGPT to check grammatical errors in their writing. For that purpose, ChatGPT could present detected grammatical errors in students' writing in a table (Tseng & Lin, 2024), explain the errors translated into students' L1 (e.g., Japanese) (Schmidt-Fajlik, 2023), provide examples of using the correct grammar (Baskara, 2023), and give a score for an essay along with reasons for giving that score (Mizumoto & Eguchi, 2023). Besides grammar, other forms of feedback that ChatGPT could provide for students include feedback on sentence structure, spelling (Özçelik & Ekşi, 2024; Zadorozhnyy & Lai, 2023), punctuation (Algaraady & Mahyoob, 2023), essay structure, sentence clarity, cohesion, and coherence in writing (Alberth, 2023; Song & Song, 2023; Tseng & Lin, 2024). With these practices reported by the previous studies, Mizumoto and Eguchi (2023) see the potential of utilizing ChatGPT to empower non-native English speakers linguistically.

Doing the Writing Work

It is alarming that some students were reported to use ChatGPT to write their writing assignments. The students used it for completing either as part of (36.1%) or the entire assignment (18%) (see Črček & Patekar, 2023). Similarly, "some students only copy essays generated by ChatGPT and submit them as their own ones. This leads to a worry over academic integrity" (Hang, 2023, p. 26).

The reviewed literature suggests that students use ChatGPT across multiple writing stages, from brainstorming to grammar checking. Yet, some students seem to face difficulties using ChatGPT ethically and independently without institutional guidance. Therefore, it is crucial to learn about best practices in the literature that readers can adopt to regulate the use of ChatGPT in their EFL/ESL writing classrooms, which will be presented in the following sections.

RQ 2: Best Practices to Regulate the Use of ChatGPT in EFL/ESL Writing Classrooms

The review informed various best practices for regulating the use of ChatGPT in the writing classroom, which could be categorized into four main sections: institutional policies, instructional strategies, assessment design innovations, and ethical co-regulation practices.

Institutional Policies

The institutional policies cover two best practices: using AI detection tools and implementing citation policies when using ChatGPT. Each will be described in the following paragraphs.

Using AI Detection Tools. AI detection tools were proposed to discover potential plagiarism in students' writing. Some tools include *Turnitin* (https://www.turnitin.com/) (Carlson et al., 2023; Cong-Lem et al., 2024; Mohammadkarimi, 2023), *Originality.Ai* (https://originality.ai/), Copyleasks (https://copyleaks.com/), *GPTZero* (https://gptzero.me/), and *Open.ai* (https://openai-openai-detector--qz8sj.hf.space/) (Krajka & Olszak, 2024). Then, as stated by an EFL teacher, "teachers should inform students that they will use some detectors to check the authenticity of their submitted papers, and a cheater will get a zero for a ChatGPT-generated essay. I believe students will be reluctant to use ChatGPT to complete a writing assignment" (Hang, 2023, p. 29).

Properly Citing ChatGPT-Generated Texts Used in Students' Writing. Students must tell their teachers how they used ChatGPT in their writing. One possible way is to ask students to cite ChatGPT-generated texts that they use or paraphrase in their work to prevent plagiarism and adhere to scholarly writing principles (Alqasham, 2023; Mondal & Mondal, 2023; Rababah et al., 2023; Tseng & Warschauer, 2023). Recently, the American Psychological Association released guidelines to cite ChatGPT-generated texts on its website, which should be discussed in writing classes. For more details, read McAdoo (2024)[1].

Instructional Strategies

The instructional strategies highlight the value of the writing process and have the students write in class. The other one is to scaffold writing assignments, that is, to break larger writing assignments into smaller writing tasks.

Emphasizing the Value of the Writing Process. The value of the writing process can be emphasized more to mitigate the potential issues of using ChatGPT. For instance, teachers can ask their students to document their writing steps, such as selecting their topic, outlining their ideas, writing their draft, and revising their work until they finally produce their final writing work (Barrot, 2023). Students should also submit notes they took on any sources to prepare their writing (Moorhouse et al., 2023). Regarding ChatGPT, teachers can ask their students to submit their conversation history with ChatGPT as an appendix to their final work so that the teachers can see the writing process and how much ChatGPT helps students improve their writing.

Having the Writing Done in Class. Students can be asked to write in class. Doing so, students are not allowed to access mobile devices and the internet when completing the written work (Cong-Lem et al., 2024). A teacher suggested that "to evaluate students' writing performance precisely, students should be asked to do several writing tests in class with the observation of the teacher without the use of ChatGPT" (Hang, 2023, p. 28).

Scaffolding Writing Assignments. Teachers should thoughtfully scaffold their writing assignments to ensure their students have sufficient time, space, and support during their writing process. It is possible to break larger writing assignments into smaller pieces of writing tasks. Moorhouse et al. (2023) reminded us that students are more inclined to turn to ChatGPT if they feel stressed, overwhelmed, unsupported, or out of time.

Assessment Design Innovations

In the AI era, EFL/ESL writing lecturers should start designing innovative writing assessments for students, making them more integrative and holistic (Cong-Lem et al., 2024), and providing alternative ways for students to demonstrate what they have learned beyond the text (Baskara & Mukarto, 2023; Moorhouse et al., 2023).

Innovating the Design of Writing Assessments. That said, teachers might integrate some speaking or interviewing activities to ensure that students' writing results from their genuine work (Cong-Lem et al., 2024). For example, "teachers should require students to present the progress of constructing ideas, making outlines, and generating the essay to determine the originality of their written work" (Hang, 2023, p. 29). Teachers can also think about writing tasks requiring creativity, originality, problem-solving, and critical thinking (Mohammadkarimi, 2023). As Baskara (2023) observed, ChatGPT is limited to generating text based on the input given by its users; thus, it cannot generate completely original or creative text. Additionally, Hang (2023) suggested a combination of formative and summative writing assessments; in this case, students' consistent writing performances in those diverse assessments might indicate students' actual writing skills. Students can also write an essay that requires a close analysis of the materials (e.g., images, videos, course

[1] McAdoo, T. (2024). *How to cite ChatGPT*. APA Style. https://apastyle.apa.org/blog/how-to-cite-chatgpt

books, class conversations) used in their class (Moorhouse et al., 2023), and that encourages students to discuss their personal experiences (Barrot, 2023).

Ethical Co-Regulation Practices

The other best practices relate to ethical co-regulating practices, where students and their lecturers openly discuss and create guidelines of what students can/cannot do when using ChatGPT in their writing process.

Co-Creating Guidelines for ChatGPT Use with Students. Some researchers (e.g., Baskara, 2023; Cong-Lem et al., 2024; Kostka & Toncelli, 2023; Tseng & Warschauer, 2023; Xiao & Zhi, 2023) suggested that teachers discuss and co-create guidelines and ethical uses for ChatGPT with their students when doing their writing tasks once the course begins and (if necessary) refine the guidelines as the semester progresses. All these guidelines should be present in the course syllabus and communicated to the students as clearly as possible. Teachers can discuss with their students what they may or may not do when using ChatGPT in their writing classrooms, including potential limitations of ChatGPT. These discussions should be "open-minded" (Moorhouse et al., 2023, p. 8). For example, learning from Alberth (2023); Mohammadkarimi (2023), teachers and students may agree that copying texts generated by ChatGPT as they are and pasting them into students' papers is unethical. Yet, students may use ChatGPT to generate writing ideas (Črček & Patekar, 2023).

Learning from Schmidt-Fajlik's (2023); Özçelik and Ekşi's (2024) study, the instructors asked their students to write their first draft themselves. Then, after their first draft was ready and reviewed by their instructors, the students were allowed to use ChatGPT to edit their writing based on the prompts provided by their instructors. After including the ChatGPT suggestions and corrections in their draft, the students submitted their final draft to their instructors for a final review. Another thing is to cite ChatGPT-generated texts that students used in their writing, as discussed in the previous part. By engaging in the co-creation process with the students, it is hoped that teachers can also foster the spirit of honesty-humility (i.e., prioritizing fairness over students' interests), conscientiousness (i.e., a strong work ethic), and openness to experience (i.e., deciding to tackle challenges with students' ideas) to their students (Črček & Patekar, 2023). In that process, teachers can also communicate with students about the essence of original writing, the implications of using ChatGPT inappropriately for themselves and their community, and the value of their academic degree (Moorhouse et al., 2023).

Applying Stricter Regulations. The discussions with the students can also cover how far teachers could implement stricter regulations to minimize inappropriate and dishonest use of ChatGPT. The stricter regulations can include giving a kind of punishment for students who are caught unethically using ChatGPT in their writing task (Cong-Lem et al., 2024) "[...] If AI-generated paragraphs are detected, students will get a zero for the assignment" (Hang, 2023, p. 29). Črček and Patekar (2023) reported the act of banning the use of ChatGPT, as Italy did temporarily in March 2023. To sum up, the reviewed literature has suggested eight best practices, each with its potential strengths and weaknesses, to regulate the use of ChatGPT in EFL/ESL writing classrooms, indicating that the unethical use of ChatGPT can be mitigated through carefully planned pedagogical writing instructions.

DISCUSSION

This discussion section interprets the findings in light of the two research questions and situates them within the broader scholarly discourse. That said, this study can also evaluate the current state of research, synthesize emerging trends and tensions, and suggest future directions for pedagogical practice and scholarly inquiry.

Interpretations of Key Findings in Light of Research Questions

This study aimed to answer two main research questions regarding the use of ChatGPT in EFL/ESL writing classrooms. The first research question describes how university language students use ChatGPT in their language writing classrooms. The previous studies show the widespread use of ChatGPT across various countries to help students in their writing process for various academic (e.g., seminar papers, essays, thesis, and editorial writing) and non-academic types of writing (e.g., email, blog post, informal text messages). On the positive side of ChatGPT, previous studies have consistently reported that ChatGPT functions well as a supportive digital tutor for students, helping them generate writing ideas and providing feedback to enhance their writing work. On the negative side of ChatGPT, Črček and Patekar (2023) and Hang (2023) have the same concern that students use ChatGPT to write for them; they copy and paste the ChatGPT-generated text into their essays and submit it to their teacher as their work. No doubt, this action is extreme and unethical. If this happens continuously in the students' writing process, writing instructors might find it challenging to discern students' proper understanding and mastery of learning materials (Grassini, 2023), presented in the writing class. Besides, the ongoing unethical use of ChatGPT, such as in formative writing practices, can also make students dependent on ChatGPT to generate answers to any questions, which makes them unable to think logically in their writing (Hang, 2023).

In that case, as the answer to the second research question, previous studies have similar views on applying stricter regulations, such as giving a kind of punishment to the students (Cong-Lem et al., 2024), giving zero points to students'

work detected with AI-generated paragraphs (Hang, 2023), or banning the use of ChatGPT thoroughly (Črček & Patekar, 2023). Yet, the researcher believes it is an unsustainable solution to respond to the dark side of ChatGPT. In the ongoing debate on whether ChatGPT should be banned in education, what Cong-Lem et al. (2024); Črček and Patekar (2023); Hang (2023) suggested about the use of strict regulations contradicts Tseng and Warschauer's (2023) view, believing that students will lose essential opportunities to learn how to effectively use AI-based tools in their future workflows if they are not allowed to use the tools in their classrooms. "In a world that increasingly values the use of AI in the workplace, students who lack experience manipulating AI tools to increase their productivity and efficiency will fall behind those who do have the experience and skills to use those tools effectively" (Tseng & Warschauer, 2023, p. 259).

In alternative to applying the strict regulations described above, previous studies suggest that teachers use web-based AI detection tools to identify potential plagiarism in their students' writing. While this approach holds promise and is worth trying, Cong-Lem et al. (2024) pessimistically said that utilizing the detection tools might not work well for detecting AI-generated texts; this highlights a critical area for future research: exploring and evaluating the efficacy of various AI detection tools to better support teachers in identifying AI-generated writing in their students' work.

As reported in the findings, the ideas of designing writing tasks requiring creativity, problem-solving, and critical thinking, incorporating real-life and personal illustrations (Mohammadkarimi, 2023) and integrating additional oral assessments to clarify what they write (Cong-Lem et al., 2024) challenge traditional writing instructions. EFL/ESL writing lecturers can no longer assign their students to write an essay on a free topic and directly submit it to their lecturer for final grading, given that students might ask ChatGPT to write for them. That said, the researcher views the need to encourage the lecturers to monitor the students' writing process more closely, not just the final product, and carefully plan their pedagogical writing instructions. That writing process consists of five main stages: *planning, drafting, revising* and *editing,* and *submitting* (see Mali & Salsbury, 2021, pp. 251-252)

Moreover, students will likely use ChatGPT if they feel stressed and overwhelmed and lack time and support to complete their writing tasks (Moorhouse et al., 2023). This might mean that teachers should carefully consider the number of writing tasks to complete within a semester to ensure that students have time to complete each assignment. Moorhouse et al.'s (2023) argumentation also means a strong critique for all undedicated writing instructors who are often busy with their projects outside campus, which makes them unable to give time for consultations or support students in their writing process. These instructors should be aware that their lack of support for students might be one of the reasons why their students decide to use ChatGPT.

Methodological and Conceptual Gaps in the Literature

While many previous studies have been conducted to explore the use of ChatGPT in EFL/ESL writing classrooms, the researcher could identify some methodological and conceptual gaps. Methodologically speaking, the earlier studies had not conducted any in-depth class-based observations to see how well lecturers regulate the use of ChatGPT in their EFL/ESL writing classrooms, which was commonly obtained from interviews and questionnaire data. The researcher also viewed that most previous studies lack longitudinal work within a specific period to deeply explore and understand how students use ChatGPT in their writing process and how lecturers can regulate its use ethically in writing classes. Very few studies compared students' use of ChatGPT and other related AI tools in their writing process. Moreover, all the reviewed studies were also conducted in higher education settings, mainly in Asia and the Middle East counties, leaving the question of how senior, junior, or perhaps elementary school students used ChatGPT or similar AI tools, particularly to complete a written work assigned by their English language teacher, in less-represented educational settings, such as in South America, and Africa. Conceptually speaking, all the reviewed studies cannot assure the most effective ways to regulate the ethical use of ChatGPT in EFL/ESL classrooms, particularly from the eyes of university lecturers teaching EFL/ESL writing courses or relevant stakeholders in the EFL/ESL education field. While some studies (e.g., Baskara, 2023; Hang, 2023; Kostka & Toncelli, 2023) suggested using writing tasks that require a high level of creativity, problem-solving, and critical thinking that might help prevent students from using ChatGPT unethically, they did not clearly illustrate what the tasks look like.

The author of the current review acknowledges a degree of overlap between the present review and the recent study conducted by Teng (2024a), published in July 2024. To clarify the distinct contribution of this study, several points merit consideration. First, the current review extends the geographical scope of Teng's work by incorporating studies from a range of educational contexts that were not represented in his review, including Taiwan, Indonesia, the Czech Republic, Jordan, Vietnam, Iraq, Saudi Arabia, and Poland. This broader inclusion enhances the cross-cultural relevance of the findings and allows for a more comprehensive understanding of ChatGPT's use in diverse ESL/EFL writing classrooms. Second, unlike Teng's study, which did not specify the use of bibliometric tools for data retrieval, the present review employed the Publish or Perish software to system-

atically identify relevant literature. This methodological divergence reflects a more transparent and replicable search strategy and resulted in the inclusion of a different body of studies. Notably, only two sources (Algaraady & Mahyoob, 2023; Song & Song, 2023) were common to both reviews. Finally, the findings presented in this article aim to confirm, refine, or challenge those reported by Teng (2024a), as well as those of other studies included in the present corpus (see Table 2). By offering new empirical insights and comparative perspectives, this review seeks to advance the scholarly conversation on the pedagogical applications of ChatGPT in ESL/EFL writing instruction.

Recommendations for Future Research

Future studies might plan the following research agendas to address the methodological and conceptual gaps in the literature mentioned previously. To complement the interview and questionnaire data commonly reported by previous studies, future researchers can conduct a longitudinal (i.e., in a semester) in-depth classroom-based observation in EFL/ESL writing classrooms involving underexplored populations of K-12 students (elementary, junior, and senior high school students) in less-researched settings in South American and African countries. That observation should be aimed at seeing how well teachers in the classes mediate and regulate the use of ChatGPT with their students and how far students implement that regulation in each stage of their writing process. With those observations, future researchers could capture real situations or conditions when students used ChatGPT ethically or unethically and how far the regulation to mitigate the unethical use of ChatGPT works well in their classes. In that longitudinal research, future researchers can also assess ChatGPT's quality of feedback, writing assistance, and perceived usefulness in supporting students' writing process compared to related AI tools developed recently, such as Google AI Studio or DeepSeek.

Implications for Pedagogy and Policy

The findings of this study yield several important implications for institutions, writing instructors, department heads, and students involved in EFL/ESL writing classrooms (Table 4).

First, at the institutional level, there is a growing need to acknowledge what may soon become the new normal in academic practice: the increasing use of generative AI tools such as ChatGPT by university students to support their writing. Rather than resisting this shift, universities should take proactive steps to foster responsible and competent use of such technologies. This includes promoting digital literacy initiatives aimed at helping both faculty and students develop the necessary skills to effectively engage with AI tools. Specifically, institutions should offer opportunities for learning how to craft effective prompts, critically evaluate AI-generated content, and understand ethical boundaries in its application. In this regard, the present study supports earlier calls by Hang (2023) and Özçelik and Ekşi (2024), who emphasize the importance of institutional engagement through symposiums, workshops, faculty-student discussions, and targeted training programs that focus on the practical, pedagogical, and ethical dimensions of using ChatGPT in academic contexts.

Second, writing instructors are encouraged to revise their course syllabi to include explicit guidelines on the appropriate and inappropriate uses of ChatGPT in academic writing. These guidelines should not merely appear as policy statements but should be introduced and discussed constructively with students at the beginning of each semester. Such dialogue can help students internalize expectations and avoid unintentional misuse of generative AI tools.

Third, the study highlights the pedagogical value of implementing a portfolio-based assessment approach in writing instruction. Portfolios allow instructors to track students' development over time by evaluating not only the final product but also the full writing process. As proposed by Sulistyo et al. (2020), a comprehensive writing portfolio may include an outline, in-class draft, final submission, a documented interaction with ChatGPT for revision purposes (e.g., grammar checking), and, if needed, results of an oral follow-up to verify authorship and comprehension.

Fourth, department heads play a critical role in ensuring quality instruction. They are advised to assign writing courses to instructors who are not only qualified but also committed to supporting students throughout the writing process. Given the labor-intensive nature of such instruction, especially in large classes, writing instructors should be provided with teaching assistants if the student-teacher ratio exceeds 10:1. As Moorhouse et al. (2023, p. 7) warn, students are more likely to rely on ChatGPT when they feel overwhelmed, unsupported, or constrained by time—factors that institutional design and instructor availability can directly mitigate.

Finally, students themselves must develop a nuanced understanding of both the capabilities and limitations of ChatGPT. Raising students' awareness of ethical boundaries and promoting critical digital literacy should begin early in their academic journey. Universities might consider offering a short, compulsory orientation course or workshops focused on maximizing the pedagogical benefits of ChatGPT while avoiding overreliance. Such training could be integrated into first-year curricula or offered through extracurricular channels.

Furthermore, reflecting on the previous discussions, the researcher would like to propose sample writing activities (see Figure 3). The activities should inform how writing lecturers and students can ethically and productively use ChatGPT in a writing classroom. These writing activities can be adapted to suit various EFL/ESL writing classes that use the pro-

Table 4
The Summary of the Implications

Stakeholders	Implications
Institutions	Encourage their teachers and students to upgrade and enhance their competencies to handle the current advancements of ChatGPT or other related AI technology.
	Conduct symposiums, conferences, trainings, regular discussions among faculty members and students, or other feasible attempts to enhance the teachers' and students' competencies.
Writing instructors	Enhance their AI and technology literacy.
	Include clear guidelines on acceptable use of ChatGPT in their course syllabi.
	Discuss the guidelines with students.
	Implement the writing portfolio to assess each student's writing.
	Implement the process approach of teaching writing (see Figure 3).
Department head	Thoughtfully select instructors who will teach the writing class (i.e., the ones who are technology and AI literate and committed to supporting students' writing.
	Plan a writing class with a small number of students.
	Prepare a teaching assistant for writing lecturers who teach a class of more than 10 students.
Students in EFL/ESL writing classrooms	Raise their awareness of ChatGPT's bright and dark sides.
	Clearly understand what they can/cannot do with ChatGPT when completing their written work.
	Be aware of the potential risks and limitations of ChatGPT and being too dependent on ChatGPT.
	Enhance their AI and technology literacy.

cess-based writing approach of Mali and Salsbury (2021). The writing activities can also facilitate the writing of diverse genres, such as seminar papers, argumentative essays, theses, and editorial writing. The other genres include academic and non-academic essays as well as scientific papers. Moreover, the researcher was confident that the sample writing activities, along with the results and discussions presented in the study, are generalizable to broader EFL/ESL writing contexts, given that they were derived from the rigorous analysis of relevant studies on the use of ChatGPT in EFL/ESL writing classes across various countries (see Figure 2).

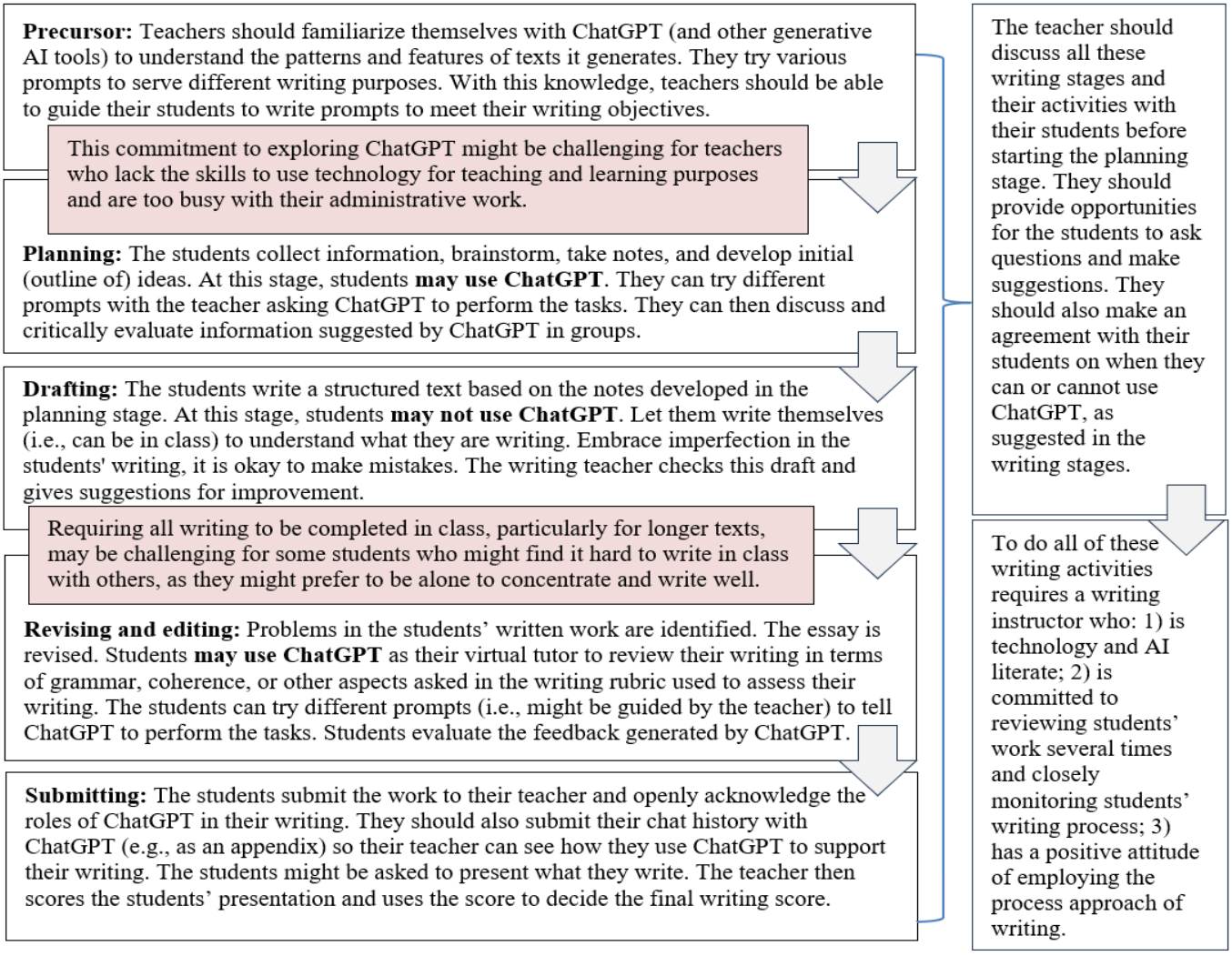
In the planning stage, where students still brainstorm, write an outline of their writing, and develop writing ideas, for example, the lecturers and their students can collaboratively formulate guidelines for the ethical use of ChatGPT, establishing clear agreements on what they may or may not do with ChatGPT in their writing tasks. All the agreed-upon points should be stated clearly in the class syllabus. In support of the literature, students should always cite ChatGPT-generated texts that they use or paraphrase in their work (Xiao & Zhi, 2023). This practice underscores the importance of providing students sufficient time to learn and practice in-text and end-text citations (e.g., based on APA 7th edition or other writing conventions) early in their writing coursework, not just in upper-level academic or research proposal writing classes.

To engage in fruitful discussions when creating the guidelines for using ChatGPT with students and later provide support and guidance to their students when needed, first and foremost, teachers should make themselves familiar with ChatGPT. They should experiment with ChatGPT (or other generative AI tools in their fields) for their writing to understand its features, potential strengths, weaknesses, as well as patterns of texts generated by ChatGPT (Alqasham, 2023; Hang, 2023; Mohammadkarimi, 2023). Indeed, a commitment to exploring the applications of ChatGPT in teaching and learning practices is crucial for preparing students for the era of AI (Kostka & Toncelli, 2023). However, this commitment might be challenging for teachers who lack the skills to use technology for teaching and learning purposes (Mali, 2025) and are too busy with their administrative work (Muslem et al., 2018).

Then, in the *drafting* stage, where students start to develop a structured written text from their outline, their lecturers can ask their students in class (Carlson et al., 2023; Cong-Lem et al., 2024). That method could be interpreted as encouraging teachers to know their students' writing capabilities. Practically speaking, at the beginning of the semester, teachers might ask their students to write two to three descriptive paragraphs about any topic that interests them. The writing should be done in class without the use of ChatGPT. The teacher can then collect the students' work and read it closely to know their current level of writing quality at the beginning of the semester. In this stage, teachers should embrace imperfection in their students' writing; emphasizing that it is okay to make mistakes in the first draft, but, more importantly, the students write themselves and know

Figure 3

The Sample Writing Activities Using ChatGPT



what they are writing. Suppose there is a significant disparity between the quality of this initial work and subsequent writing assignments. In that case, teachers might be skeptical that their students (e.g., in school contexts with excellent internet access) might be using ChatGPT or other unethical means (e.g., copy-pasting from online resources) to write for them. However, requiring all writing to be completed in class, particularly for longer texts (e.g., essays or theses), may be challenging for the students. Mali (2024) reported that some university students found it hard to write in class with others, as they needed to be alone to concentrate and write well.

In the *revising* and *editing* stages, where students identify possible grammatical errors in their writing and write a clean copy of their work, students may use ChatGPT to provide various feedback on the first draft of their writing but *not* ask ChatGPT to write parts or the entire parts of the writing, as suggested by Özçelik and Ekşi (2024). It is important to emphasize that the students should write the draft before ChatGPT reviews it. After revising their work based on ChatGPT's suggestions and corrections in their first draft, the students could submit their final draft to their instructors for a final review.

For transparency purposes in the *submitting* stage, teachers may require students to submit the feedback generated by ChatGPT and the chat history made with ChatGPT as an appendix in their final draft submission. In this case, it is crucial that teachers directly teach and practice various clear and personalized prompts for ChatGPT to generate valuable and good feedback for students' writing. For prompt references, see Carlson et al. (2023); Mondal and Mondal (2023); Teng (2024b).

Overall, the reviewed literature in this study paints a complex picture of ChatGPT as a promising support tool and digital writing tutor for students and, simultaneously, a po-

tential threat to writing integrity. The challenge now is not whether to use AI tools in writing. Yet, it is more about planning pedagogical writing instructions that can mitigate the unethical use of ChatGPT or other AI tools in the recent massive disruption of AI to preserve the pedagogical integrity and critical literacy goals of writing education.

CONCLUSION

This study has identified key patterns in how EFL/ESL learners have used ChatGPT in writing tasks and replicable best practices for regulating its use in classroom contexts. Among these, co-creating ethical guidelines with students and emphasizing the writing process seemed to be particularly promising strategies to mitigate the unethical use of ChatGPT in EFL/ESL writing classrooms. What this study has found and discussed offers timely support for EFL/ESL educators and policymakers seeking to balance innovation with integrity in AI-mediated writing instruction. Nevertheless, this study is far from perfect because of the absence of research participants in sharing their perspectives on the themes presented in this study, e.g., how far the participants can/cannot accept the ideas of regulating AI use in the writing class. To address this limitation, future researchers can invite research participants, e.g., students, fellow lecturers, AI or technology experts, and experienced writing professors, to respond to their literature review results and explore their views on how they perceive the ethical integration of AI tools in diverse cultural and institutional contexts, or to assess the long-term impact of such tools on writing proficiency and academic integrity. As generative AI becomes a permanent fixture in educational practice, this study serves as a strong foundation for developing pedagogically sound, ethically aligned writing instructions in EFL/ESL writing classrooms that empower students in their writing process without compromising academic standards.

GEN AI STATEMENT

The researcher affirmed that he did not use ChatGPT or other generative AI tools to write any single sentence in this paper. However, the researcher would like to acknowledge the use of *Grammarly* Premium after he has finished writing the first draft of this paper. The researcher used *Grammarly* to help identify grammatical errors in his sentences. The identified errors were then revised to enhance the grammatical accuracy of this paper. The researcher also used *Grammarly* to ask for some suggestions on academic phrases to improve the readability of his sentences. The researcher thoughtfully reviewed all the suggestions made by Grammarly to ensure the accuracy and quality of the sentences in this paper. Importantly, as informed by Cheng et al. (2025), researchers can ethically use ChatGPT or related generative AI tools to "check grammar, improve syntax, ensure consistency in technical terminology, refine complex sentences, and ensure that the manuscript flows well and is easy to follow" (p. 4) as long as they disclose the use of the tools in their manuscript.

ACKNOWLEDGMENT

The author would like to thank the Directorate of Research and Community Service at Universitas Kristen Satya Wacana for supporting this research and Professor Elena V. Tikhonova and the anonymous reviewers for their constructive feedback and never-ending guidance that helped the author to enhance the quality of this manuscript.

DECLARATION OF COMPETING INTEREST

None declared.

REFERENCES

Alberth. (2023). The use of ChatGPT in academic writing: A blessing or a curse in disguise? *TEFLIN Journal*, *34*(2), 337–352. https://doi.org/10.15639/teflinjournal.v34i2/337-352

Algaraady, J., & Mahyoob, M. (2023). ChatGPT's capabilities in spotting and analyzing writing errors experienced by EFL learners. *Arab World English Journal*, *9*, 3–17. https://doi.org/https://dx.doi.org/10.24093/awej/call9.1

Alqasham, F. H. (2023). ChatGPT in the Saudi EFL classroom: A study of learner usage patterns and possibilities in learning optimization. *Migration Letters*, *20*(S7), 1251–1263. https://doi.org/https://doi.org/10.59670/ml.v20iS7.4828

Barrot, J. S. (2023). Using ChatGPT for second language writing: Pitfalls and potentials. *Assessing Writing*, *57*, 1–6. https://doi.org/10.1016/j.asw.2023.100745

Baskara, F. R. (2023). Integrating ChatGPT into EFL writing instruction: Benefits and challenges. *International Journal of Education and Learning*, *5*(1), 44–55. https://doi.org/10.31763/ijele.v5i1.858

Baskara, F. R., & Mukarto, F. (2023). Exploring the implications of ChatGPT for language learning in higher education. *Indonesian Journal of English Language Teaching and Applied Linguistics*, *7*(2), 343–358. https://ijeltal.org/index.php/ijeltal/article/view/1387

Bonner, E., Lege, R., & Frazier, E. (2023). Large language model-based artificial intelligence in the language classroom: Practical ideas for teaching. *Teaching English With Technology*, *23*(1), 23–41. https://doi.org/10.56297/bkam1691/wieo1749

Carlson, M., Pack, A., & Escalante, J. (2023). Utilizing OpenAI's GPT-4 for written feedback. *TESOL Journal*, 1–7. https://doi.org/10.1002/tesj.759

Cheng, A., Calhoun, A., & Reedy, G. (2025). Artificial intelligence-assisted academic writing: Recommendations for ethical use. *Advances in Simulation*, *10*(1), 1–9. https://doi.org/10.1186/s41077-025-00350-6

Cheong, W., & Hong, H. (2023). The impact of ChatGPT on foreign language teaching and learning: Opportunities in education and research. *Journal of Educational Technology and Innovation*, *5*(1), 37–45. https://jeti.thewsu.org/index.php/cieti/article/view/103

Cong-Lem, N., Tran, T. N., & Nguyen, T. T. (2024). Academic integrity in the age of generative AI: Perceptions and response of Vietnamese EFL teachers. *Teaching English with Technology*, *24*(1), 28–47. https://doi.org/10.56297/FSYB3031/MXNB7567

Črček, N., & Patekar, J. (2023). Writing with AI: University students' use of ChatGPT. *Journal of Language and Education*, *9*(4), 128–138. https://doi.org/10.17323/jle.2023.17379

Grassini, S. (2023). Shaping the future of education: Exploring the potential and consequences of AI and ChatGPT in educational settings. *Education Sciences*, *13*(7), 1–13. https://doi.org/10.3390/educsci13070692

Hang, N. T. T. (2023). EFL techers' perspectives toward the use of ChatGPT in writing classes: A case study at Van Lang University. *International Journal of Language Instruction*, *2*(3), 1–47. https://doi.org/https://doi.org/10.54855/ijli.23231

Ho, A., & Nguyen, H. (2024). Generative artificial intelligence and ChatGPT in language learning: EFL students' perceptions of technology acceptance. *Journal of University Teaching & Learning Practice*, *21*(6). https://doi.org/https://doi.org/10.53761/fr1rkj58

Imran, M., & Almusharraf, N. (2023). Analyzing the role of ChatGPT as a writing assistant at higher education level: A systematic review of the literature. *Contemporary Educational Technology*, *15*(4), 1–14. https://doi.org/10.30935/cedtech/13605

Klimova, B., Pikhart, M., & Al-Obaydi, L. H. (2024). Exploring the potential of ChatGPT for foreign language education at the university level. *Frontiers in Psychology*, *15*, 1–10. https://doi.org/10.3389/fpsyg.2024.1269319

Kostka, I., & Toncelli, R. (2023). Exploring applications of ChatGPT to English language teaching: Opportunities, challenges, and recommendations. *TESL-EJ*, *27*(3), 1–19. https://doi.org/10.55593/ej.27107int

Krajka, J., & Olszak, I. (2024). "AI, will you help?" How learners use artificial intelligence when writing. *XLinguae*, *17*(1), 34–48. https://doi.org/10.18355/XL.2024.17.01.03

Li, M. (2018). Computer-mediated collaborative writing in L2 contexts: An analysis of empirical research. *Computer Assisted Language Learning*, *31*(8), 1–23. https://doi.org/10.1080/09588221.2018.1465981

Mali, Y. C. G. (2017). Promoting effort attributions to EFL students. *Accents Asia*, *9*(1), 30–40.

Mali, Y. C. G. (2024). EFL graduate students' voices on their technology-integrated classroom language tasks. *LLT Journal: A Journal on Language and Language Learning*, *27*(1), 116–135. https://doi.org/https://doi.org/10.24071/llt.v27i1.7375

Mali, Y. C. G. (2025). Factors hindering the integration and potential of technology in EFL classrooms. *International Journal of Indonesian Education and Teaching*, *9*(1), 171–185. https://doi.org/https://doi.org/10.24071/ijiet.v9i1.9553

Mali, Y. C. G., & Salsbury, T. L. (2021). Technology integration in an Indonesian EFL writing classroom. *TEFLIN Journal*, *32*(2), 243–266. https://journal.teflin.org/index.php/journal/article/view/1558/354

Marzuki, Widiati, U., Rusdin, D., Darwin, & Indrawati, I. (2023). The impact of AI writing tools on the content and organization of students' writing: EFL teachers' perspective. *Cogent Education*, *10*(2), 1–17. https://doi.org/10.1080/2331186X.2023.2236469

Mizumoto, A., & Eguchi, M. (2023). Exploring the potential of using an AI language model for automated essay scoring. *Research Methods in Applied Linguistics*, *2*(2), 1–13. https://doi.org/10.1016/j.rmal.2023.100050

Mohammadkarimi, E. (2023). Teachers' reflections on academic dishonesty in EFL students' writings in the era of artificial intelligence. *Journal of Applied Learning and Teaching*, *6*(2), 105–113. https://doi.org/10.37074/jalt.2023.6.2.10

Mondal, H., & Mondal, S. (2023). ChatGPT in academic writing: Maximizing its benefits and minimizing the risks. *Indian Journal of Ophthalmology*, *71*(12), 3600–3606. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC10788737/

Moorhouse, B. L., Yeo, M. A., & Wan, Y. (2023). Generative AI tools and assessment: Guidelines of the world's top-ranking universities. *Computers and Education Open*, *5*, 1–10. https://doi.org/10.1016/j.caeo.2023.100151

Muslem, A., Yusuf, Y. Q., & Juliana, R. (2018). Perceptions and barriers to ICT use among English teachers in Indonesia. *Teaching English with Technology*, *18*(1), 3–23. https://files.eric.ed.gov/fulltext/EJ1170638.pdf

Özçelik, N. P., & Ekşi, G. Y. (2024). Cultivating writing skills: The role of ChatGPT as a learning assistant - A case study. *Smart Learning Environments*, *11*(10), 1–18. https://doi.org/10.1186/s40561-024-00296-8

Poveda-Garcia-Noblejas, B., & Antropova, S. (2024). Analysis of CLIL-related research in school settings: A systematic review. *Journal of Language and Education*, *10*(1), 146–159. https://doi.org/10.17323/jle.2024.18150

Rababah, L. M., Rababah, M. A., & Al-Khawaldeh, N. N. (2023). Graduate students' ChatGPT experience and perspectives during thesis writing. *International Journal of Engineering Pedagogy*, *14*(3), 22–35. https://doi.org/https://doi.org/10.3991/ijep.v14i3.48395

Richards, J. C., & Schmidt, R. (2010). *Longman dictionary of language teaching and applied linguistics (4th ed.)*. Harlow.

Saldaña, J., & Omasta, M. (2018). *Qualitative research: Analyzing life*. Sage Publications, Inc.

Schmidt-Fajlik, R. (2023). ChatGPT as a grammar checker for Japanese English language learners: A comparison with Grammarly and ProWritingAid. *AsiaCALL Online Journal*, *14*(1), 105–119.

Song, C., & Song, Y. (2023). Enhancing academic writing skills and motivation: Assessing the efficacy of ChatGPT in AI-assisted language learning for EFL students. *Frontiers in Psychology*, *14*, 1–14. https://doi.org/10.3389/fpsyg.2023.1260843

Sulistyo, T., Eltris, K. P. N., Mafulah, S., Budianto, S., Saiful, S., & Heriyawati, D. F. (2020). Portfolio assessment: Learning outcomes and students' attitudes. *Studies in English Language and Education*, *7*(1), 141–153. https://doi.org/10.24815/siele.v7i1.15169

Teng, M. F. (2024a). A systematic review of ChatGPT for English as a foreign language writing: Opportunities, challenges, and recommendations. *International Journal of TESOL Studies*, *6*(3), 36–57. https://doi.org/10.58304/ijts.20240304

Teng, M. F. (2024b). "ChatGPT is the companion, not enemies": EFL learners' perceptions and experiences in using ChatGPT for feedback in writing. *Computers and Education: Artificial Intelligence*, *7*, 1–10. https://doi.org/10.1016/j.caeai.2024.100270

Tikhonova, E., & Raitskaya, L. (2024). The culture of research: A systematic scoping review. *Journal of Language and Education*, *10*(1), 5–24. https://doi.org/10.17323/jle.2024.21526

Tseng, W., & Warschauer, M. (2023). AI-writing tools in education: If you can't beat them, join them. *Journal of China Computer-Assisted Language Learning*, *3*(2), 258–262. https://doi.org/10.1515/jccall-2023-0008

Tseng, Y.-C., & Lin, Y.-H. (2024). Enhancing English as a Foreign Language (EFL) learners' writing with ChatGPT: A university-level course design. *Electronic Journal of E-Learning*, 78–97. https://doi.org/https://doi.org/10.34190/ejel.21.5.3329

Xiao, Y., & Zhi, Y. (2023). An exploratory study of EFL learners' use of ChatGPT for language learning tasks: Experience and perceptions. *Languages*, *8*(3), 1–12. https://doi.org/10.3390/languages8030212

Yeo, M. A. (2024). ChatGPT and the future of editorial writing. *RELC Journal*, *55*(1), 1–9. https://doi.org/10.1177/003368822412405

Zadorozhnyy, A., & Lai, W. Y. W. (2023). ChatGPT and L2 written communication: A game-changer or just another tool? *Languages*, *9*(1), 1–11. https:// doi.org/10.3390/languages9010005

Zain, D. S. M. (2022). Flipped classroom model for EFL/ ESL instruction in higher education: A systematic literature review. *Journal of Language & Education*, *8*(3), 133–149. https://doi.org/10.17323/jle.2022.12855

https://doi.org/10.17323/jle.2025.24798

Enhancing English Language Teaching and User Experience in Virtual Environments: A Systematic Review on Gamification and Personalised Learning

Myriam Tatiana Velarde Orozco, Bárbara Luisa De Benito Crosetti

University ot the Balearic Islands, Spain

ABSTRACT

Introduction: Virtual learning environments (VLEs) have become central to English language teaching (ELT), although persistent disengagement suggests that design must go beyond content delivery. Gamification and personalised learning (PL) contribute to enhanced user experience (UX) and better learning outcomes, but evidence on their combined effects remains fragmented.

Purpose: This systematic literature review (SLR) explains how gamification and PL influence UX, motivation, engagement and achievement in ELT-oriented VLEs, identifies effective design practices, and maps the implementation challenges that constrain them.

Method: Following PRISMA-2020 guidelines, 46 empirical studies (2015-2024) were retrieved from Scopus, Web of Science, ERIC, and Dialnet. Two extraction matrices captured bibliographic, contextual and analytical data; methodological quality was appraised with MMAT-2018. Comparative and narrative syntheses linked design features to four outcome clusters: motivation/engagement, UX, academic performance, and learner satisfaction.

Results: Challenge-based game mechanics, points and rewards reliably increased motivation and engagement, especially when integrated with adaptive feedback. PL strategies (adaptive difficulty, learner-directed paths and tailored content) produced the strongest dual gains in satisfaction and achievement. High UX emerged only when interfaces minimised cognitive load and feedback was timely. Gaps persist between short-term motivational spikes and durable learning, and competitive elements induce anxiety in novices. Key obstacles include limited digital literacy, bandwidth constraints and sparse reporting on implementation fidelity.

Conclusion: Gamification and PL can substantially enhance UX and selected learning outcomes in ELT-oriented VLEs, but only when designs align with curricular goals, resource realities and learner profiles. Future research should pursue longer mixed-method trials, transparent adaptivity and scalable models for low-resource contexts.

KEYWORDS

educational technology, English language teaching, gamification, personalised learning, user experience, virtual learning environments

Citation: Velarde Orozco, M. T., & De Benito Crosetti, B. L. (2025). Enhancing English language teaching and user experience in virtual environments: A systematic review on gamification and personalised learning. *Journal of Language and Education, 11*(2), 157-174. https://doi.org/10.17323/jle.2025.24798

Correspondence:
Myriam Tatiana Velarde Orozco,
myriam-tatiana.velarde@uib.cat

Received: February 28, 2025
Accepted: June 10, 2025
Published: June 30, 2025

INTRODUCTION

Virtual learning environments (VLEs) have progressed from peripheral digital tools to essential platforms for language education across all levels of formal schooling. By integrating extensive resource repositories, synchronous and asynchronous communication channels, and automated assessment functions, VLEs provide flexibility and reach unavailable in conventional classrooms (Al-Busaidi & Al-Shihi, 2012). However, the rapid expansion of online instruction has exposed serious shortcomings. A large international survey conducted during the COVID-19 pandemic revealed that 72% of learners felt only "slightly" or "not at all" engaged with their online courses, linking this disengagement to weaker academic outcomes (Hollister et al., 2022). This evidence raises a critical question:

how can VLEs be designed to go beyond content delivery and genuinely engage learners?

User experience (UX) has emerged as a decisive factor in meeting this challenge. Intuitive navigation, logically organised information, and timely, responsive feedback reduce cognitive load, allowing learners to focus on the instructional task; these features consistently predict lower attrition and better performance in digital settings (Cho et al., 2022; Sanchis-Font et al., 2021). Moreover, two design approaches that could deserve particular attention within the broader UX agenda are gamification and personalised learning (PL).

Gamification introduces game elements (points, competition, narrative, etc.) into educational contexts to spark and sustain motivation (Deterding et al., 2011). Systematic reviews in English language teaching (ELT) report increased situational interest and modest improvements in vocabulary and communicative competence when game mechanics are carefully aligned with learning goals (Hung et al., 2018; Zainuddin et al., 2020; Zhang & Hasim, 2023). These reviews also note that overused or poorly integrated game features can lose their appeal, particularly when they are not pedagogically sound.

PL, in contrast, adapts content, pacing, and feedback to the needs of each learner. It is broader than adaptive learning (AL), which relies on algorithms to adjust difficulty or sequencing in real time (Klašnja-Milićević et al., 2011; Shemshack & Spector, 2020). In this review, we adopt PL as an umbrella term and include studies labelled "adaptive" when they demonstrably provide individualised support (Bernacki et al., 2021).

Although gamification and PL can address aspects of the UX challenge, their research traditions have largely evolved in parallel. Gamification reviews emphasise motivational and linguistic outcomes but rarely tackle personalisation or nuanced UX metrics (Hung et al., 2018; Zainuddin et al., 2020; Zhang & Hasim, 2023). Conversely, PL syntheses focus on analytical adaptation, often in STEM or general-education contexts, and seldom consider game elements, with only occasional reference to language learning (LL) applications (Ali et al., 2024; du Plooy et al., 2024; Gevorgyan, 2024). To our knowledge, no review has systematically examined how the combined use of gamification and PL shapes UX and learning outcomes in ELT-oriented VLEs.

Responding to this underexplored area, the present systematic literature review (SLR) examines empirical investigations published between 2015 and 2024 on the joint implementation of gamification and PL in VLEs designed for ELT. It seeks to clarify design features, evaluate their impact on learners, and identify methodological patterns and challenges. The review is guided by four research questions (RQs):

RQ1: What design features and pedagogical strategies are most effective in enhancing UX in ELT-oriented VLEs?

RQ2: How does gamification influence student motivation/engagement and academic performance in these environments?

RQ3: Which PL strategies are commonly employed, and how do they affect educational outcomes and learner satisfaction?

RQ4: What implementation challenges are reported, particularly in low-resource settings or among users with limited digital competence?

By answering these questions, we aim to guide educators, instructional designers, and researchers in designing VLEs that are not merely functional but pedagogically responsive and engaging for diverse language learners.

LITERATURE REVIEW

This section integrates three strands: UX in educational technology (ET), gamification, and PL strategies in ELT, to frame the RQs and establish the theoretical foundations needed to understand how design choices in VLEs influence motivation, engagement, and learning outcomes in ELT.

Conceptual Foundations of UX in ET

UX in ET can be framed by three complementary theoretical lenses. First, the Unified Theory of Acceptance and Use of Technology 2 (UTAUT2) posits that learners' intentions to continue using an online system depend on performance and effort expectancy as well as hedonic motivation, among other factors (Venkatesh et al., 2012). These constructs correspond directly to interface qualities such as clear navigation paths and responsive feedback that reduce friction in LL platforms. Second, Self-Determination Theory emphasises autonomy, competence and relatedness as drivers of intrinsic motivation (Ryan & Deci, 2000). Practically, VLE features that grant students meaningful choices (e.g. selecting task sequences) or provide competence-affirming feedback can satisfy these needs and sustain engagement during repetitive language practice. Third, engagement frameworks rooted in educational psychology describe behavioural, cognitive and emotional components that must be balanced for deep learning (Fredricks et al., 2004). Cognitive-load research demonstrates that poorly organised information architectures induce extraneous load and suppress these engagement dimensions (Sweller, 2011).

Recent analyses of learner forum data confirm that user control and timely system responses are strong affective

predictors of positive UX (Sanchis-Font et al., 2020). These models clarify why intuitive interfaces, motivational design and load-reducing layouts are not solely aesthetic additions but theoretical prerequisites for effective virtual LL.

Gamification in VLEs for ELT

Gamification, defined as "the application of game-design elements in non-game contexts" (Deterding et al., 2011, p. 10), is usually organised around achievement, progression, and social-interaction mechanics (Majuri et al., 2018). A meta-synthesis of 128 studies identified medium-sized advantages in learner motivation and accomplishment but noted that only a small minority (14%) used validated UX instruments, limiting insight into experience quality (Koivisto & Hamari, 2019). The same pattern appears in ELT reviews. Helvich et al. (2023) and Zhang and Hasim (2023) found that most trials lasted fewer than five weeks and relied on ad hoc engagement measures. Nonetheless, controlled classroom evidence is emerging; an eight-week study with Thai EFL undergraduates that combined a branching narrative with points and levels led participants to outperform a non-gamified comparison group on oral-fluency and vocabulary post-tests (Wichadee & Pattanapichet, 2018). Two consistent cautions arise; motivational novelty fades when mechanics are repetitive, and misaligned game tasks can trigger cognitive overload (Helvich et al., 2023; Koivisto & Hamari, 2019). Effective gamification in ELT-oriented VLEs therefore requires varied, curriculum-integrated mechanics and systematic UX monitoring.

PL and AL in VLEs for ELT

PL in virtual ELT platforms is driven by an explicit learner profile, interests, prior knowledge, and self-set goals, against which teachers or students manually select tasks, resources and pacing; common tools include choice boards, goal contracts and self-assessment rubrics that shape weekly paths (Bernacki et al., 2021; Pane, 2017[1]; Walkington & Bernacki, 2020). AL embeds algorithms that adjust difficulty on the fly: spaced-repetition decks resurface forgotten words, speech-recognition engines reintroduce problematic phonemes and rule-based tutors reorder grammar drills when error rates spike (Chaichumpa et al., 2021; Klašnja-Milićević et al., 2011; Nazempour & Darabi, 2023).

Meta-analyses indicate that PL mainly boosts motivation through perceived agency, whereas AL induces small-to-moderate advances in vocabulary and pronunciation when adaptation rules are transparent (du Plooy et al., 2024; Gevorgyan, 2024). Li et al. (2022) found that giving students freedom to choose their own projects, while an algorithm quietly fine-tuned tasks in real time, helped them master the target language forms faster than either strategy on its own. This interplay of meaningful choice and instant adjustment appears to foster a richer UX. Consistent with the conceptual boundaries outlined in the introduction, AL studies are analysed here within the broader PL corpus considering that both emphasise individualised instructional support.

Previous Studies and Gaps in the Literature

Despite the rapid growth of research on digital LL, three persistent gaps remain: conceptual fragmentation, methodological limitations, and contextual blind spots. Conceptual fragmentation is critical; gamification studies focus on motivational benefits (Koivisto & Hamari, 2019; Zhang & Hasim, 2023), while PL/AL reviews emphasise algorithmic tailoring (Ali et al., 2024; Gevorgyan, 2024). Virtually no primary study or synthesis examines how the two logics interact to improve UX.

Methodological weaknesses also compound the problem as interventions are typically brief (five to eight weeks), rely on self-developed or unvalidated instruments, and focus narrowly on receptive vocabulary, leaving productive skills and rigorously validated outcome measures underexplored (Helvich et al., 2023). Ali et al. (2024) further highlight heterogeneous methods, scarce triangulation of mixed data, and sparse documentation of implementation fidelity across PL trials. Contextual blind spots also persist, since evidence remains limited for bandwidth-constrained VLEs or learners with low digital literacy, even though studies consistently flag poor connectivity and limited technical skills as key barriers (Helvich et al., 2023).

This SLR responds directly to these limitations. RQ1 charts effective design features by jointly analysing gamification and PL. RQ2 and RQ3 analyse their separate and combined impacts on motivation/engagement and academic achievement. While UX was not always the explicit focus, studies were included if they explored gamification or PL and provided at least one relevant indicator (whether self-reported, behavioural or performance-based), even if not derived from validated instruments. RQ4 compares implementation challenges, with attention to low-resource contexts and novice users. In doing so, the review offers what may be the first integrated, methodologically focused map of how gamified-personalised VLEs shape the full UX continuum in ELT.

METHOD

This study employed the methodology of a SLR. Marín-Juarros (2022) claims that an SLR refers to an exhaustive examination of the literature using systematic methods that allow for replication and updating, addressing one or more RQs

[1] Pane, J. F. (2017). *How does personalized learning affect student achievement?* RAND Corporation. https://doi.org/10.7249/RB9994

by means of a secondary study that combines the results of primary studies. An SLR reveals gaps, deficiencies, and trends in the existing evidence, providing a foundation for and guiding future research in the field (Munn et al., 2018).

Protocol

This SLR was conducted in accordance with the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) 2020 statement (Page et al., 2021). An a priori protocol was developed to guide the conduct of this review, outlining the objectives, eligibility criteria, search strategy, study selection process, data extraction methods, quality assessment approach, and methods for data synthesis.

Eligibility Criteria

Studies were eligible for inclusion in this review when all the following conditions were met:

1. Empirical studies that examine gamification and/or PL implemented primarily in ELT-oriented VLEs and that report at least one of the following outcomes: UX, student motivation and/or engagement, learner satisfaction, or academic performance.
2. Articles published between January 2015 and July 2024 in English or Spanish.
3. Peer-reviewed journal articles. Grey literature (conference papers, proceedings, dissertations, theses, book chapters, reports) was excluded.
4. Any empirical design (qualitative, quantitative, or mixed methods). Reviews, theoretical papers, opinion essays, and editorials were excluded.
5. Full-text accessible through open access or the University of the Balearic Islands library.

Studies not meeting all the above criteria were excluded.

Information Sources

The literature search was conducted in the following electronic databases: Scopus, ISI Web of Science (WoS), ERIC, and Dialnet. The final database search was completed on 17th July 2024.

Search Strategy

A comprehensive search strategy was developed and employed to identify all relevant studies. It included a set of predefined keywords and synonyms related to UX, gamification, PL and ELT. These terms were combined using Boolean operators to set out the search strings (Table 1) and maximise the retrieval of relevant studies.

Study Selection Process

Initial database searches returned 4673 records. Filters specific to each database, publication year (2015 onwards), language (English or Spanish), and publication type (journal articles) were applied to eliminate results that did not meet the inclusion criteria, together with subject-area filters to retain studies relevant to the field. After this first screening, 90 articles were selected for further review and imported into RefWorks for organisation and deduplication. After removing two duplicates, 88 unique records were imported into Parsif.al to facilitate the screening and eligibility process. The selection process comprised two main stages: a title-and-abstract screening and a full-text assessment.

The titles and abstracts of the 88 unique articles were screened against the predefined eligibility criteria. Articles that clearly did not meet them were excluded at this stage. After the title-and-abstract screening, the full texts of the remaining records were retrieved and examined against eligibility criteria. 39 papers were excluded at this stage; some because the full text was unavailable, others since a closer

Table 1

Search Strings for Database Query

In Spanish	In English
(“experiencia de usuario” OR UX OR “experiencia del estudiante” OR gamificación OR “juegos educativos” OR “juegos serios” OR “personalización del aprendizaje” OR “aprendizaje personalizado”) AND (“entornos virtuales de aprendizaje” OR “e-learning” OR “plataformas educativas virtuales”) AND (“enseñanza del inglés” OR “aprendizaje del inglés” OR EFL OR ESL) AND (“motivación de los estudiantes” OR engagement OR “rendimiento académico” OR “logro académico” OR “satisfacción de los estudiantes”)	(“user experience” OR UX OR “student experience” OR gamification OR “educational games” OR “serious games” OR “personalised learning” OR “learning personalisation”) AND (“virtual learning environments” OR e-learning OR “online learning platforms”) AND (“English teaching” OR “English learning” OR EFL OR ESL) AND (“student motivation” OR engagement OR “academic performance” OR “academic achievement” OR “student satisfaction”)
For Dialnet (160-character limit)	
(“experiencia de usuario” OR UX OR gamificación OR “personalización del aprendizaje”) AND (“entornos virtuales de aprendizaje” OR e-learning) AND (“enseñanza del inglés” OR EFL) AND (motivación OR “rendimiento académico”)	(“user experience” OR UX OR “student experience” OR gamification OR “personalized learning”) AND (“virtual learning environments” OR e-learning) AND (“English teaching” OR ESL) AND (“student motivation” OR engagement OR “academic performance”)

reading showed they did not address the review questions, and finally 13 as they were theoretical, review, or editorial papers rather than empirical studies.

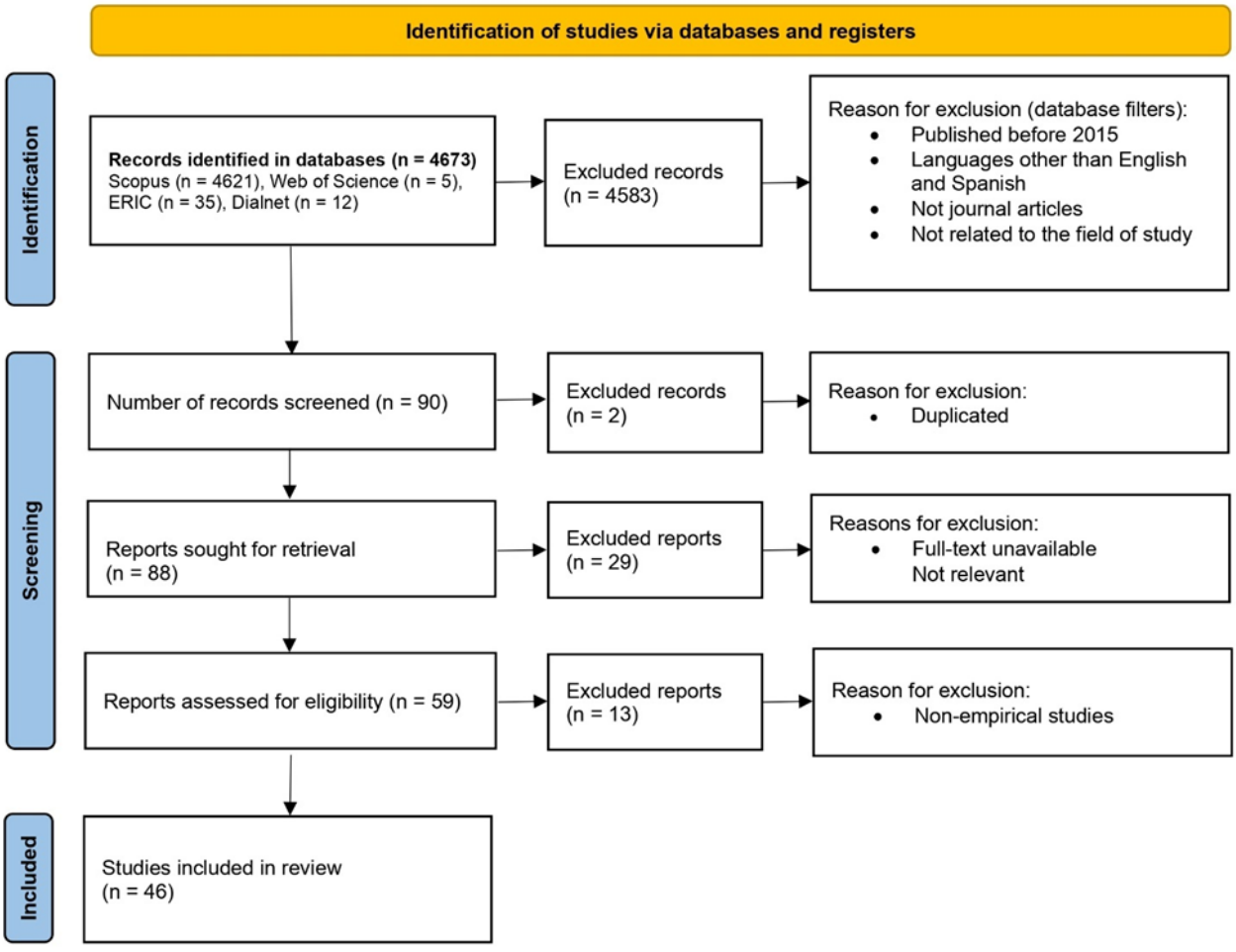
The full-text appraisal therefore resulted in a final sample of 46 empirical investigations. The numbers screened, assessed and excluded at each step, together with the reasons for exclusion, are illustrated in the PRISMA 2020 flow diagram (Figure 1). After the title-and-abstract screening stage, to support internal organisation and data management throughout the review, each article was assigned a unique reference code based on the year of publication (e.g., A for 2015, B for 2016, etc.), and the number of order (e.g., A1, A2, B1, etc.). As 13 studies were excluded during the full-text appraisal, some codes in the final sample appear non-consecutive.

Data Extraction

A structured data-extraction strategy was applied to each of the 46 eligible studies. Two complementary matrices, created in Microsoft Excel, were used to capture both descriptive and analytical information. The descriptive matrix included: Reference code, Title, Authors, Year, Journal, DOI or link, Database, Country/educational level/context, Sample size, Participant profile, Study design, and Study type. The analytical matrix recorded: Reference code, VLE description/platform, Gamification features and pedagogy, Personalisation/adaptive strategies and pedagogy, Duration/intensity of intervention, UX measurement and findings, Motivation, Engagement, Academic performance, Learner satisfaction, Reported links between design strategies and outcomes, Reported implementation challenges, Main conclusions related to the RQs, and quality verdict.

Both matrices were piloted on five studies and adjusted for clarity before full extraction began. All data were extracted from the full-text reports and each field was verified against

Figure 1
PRISMA Flowchart



the original articles. Any discrepancies were resolved through discussion and, where necessary, re-consultation of the source study. No automation tools were employed during data collection. The full descriptive and analytical matrices are presented in Appendix A and Appendix C, respectively.

Quality Assessment and Risk of Bias

The methodological quality and risk of bias for each of the 46 included studies were assessed using the Mixed Methods Appraisal Tool (MMAT), 2018 version (Hong et al., 2018). The MMAT was selected as a validated instrument suitable for appraising diverse study designs (qualitative, quantitative, and mixed methods) present in this review.

Following the MMAT protocol, after confirming suitability via two screening questions, the five relevant design-specific criteria for each study were rated as Yes (met), No (not met), or Can't tell (insufficient information). An overall methodological quality verdict (High, Moderate or Low) was then qualitatively derived for each study with a detailed rationale. This judgement was based on a critical consideration of the pattern of responses, particularly for criteria essential to each study design, rather than a numerical score or arbitrary threshold, aligning with MMAT developer recommendations. These quality assessments were used to understand the general strength of the evidence base, inform the data synthesis, and discuss limitations. The full MMAT assessment for each study is provided in Appendix B.

Data Synthesis

The synthesis was based on the descriptive and analytical data previously extracted, focusing on the relationship between design logics and educational outcomes. This mapping provided a general profile of the evidence base and guided the construction of a comparative matrix that crossed the two focal design logics, gamification mechanics and personalisation strategies, with four core outcome clusters: motivation/engagement, UX, academic performance, and learner satisfaction. Motivation and engagement were analysed jointly due to their frequent conceptual and operational overlap in the included studies. To support the analysis, a heatmap visualisation (generated in R) was produced. Descriptive details (educational level, participant profile, platform, intervention length) were also noted to frame the results, though not analysed separately.

Implementation challenges were collected across studies and sorted into three practical categories: technical, pedagogical, and learner-related, considering local resource levels and digital competence whenever possible. Throughout the analysis, MMAT quality ratings served as a reference point, with lower-quality studies flagged but not removed. In combination, the matrix, heatmap visualisation, and challenge summary underpin the structure of the Results section.

RESULTS

Study Characteristics

46 empirical studies met the final eligibility criteria (2015–2024; median publication year = 2022). Most were published in or after 2021 (74%), indicating a marked recent increase in interest in gamified-personalised ELT-oriented VLEs. Most research was carried out in Asia (26), followed by Europe (12), North/South America (7), and Africa (1). The university sector dominated the corpus (32), with a smaller representation from primary/secondary settings (10) and early-childhood or teacher-training contexts (4).

23 studies adopted purely quantitative designs but only two were randomised controlled trials (RCTs), 18 used mixed methods, and 5 were qualitative. MMAT appraisal classified 20 studies as high quality, 24 as moderate, and 2 as low. Where available, sample sizes ranged from small classroom cohorts to institution-wide groups, most commonly between 50 and 200 learners. A detailed tabulation of the characteristics of each included study is presented in Appendix A.

Core Synthesis: Comparative Matrix and Co-Occurrence Pattern

A single comparative matrix cross-links all gamification mechanics and personalisation strategies with four outcome clusters (motivation/engagement, UX, academic performance, and learner satisfaction). This comprehensive table (Appendix D) reports the number of studies, the predominant effect (+, +/–, –), and one high-quality exemplar study, where available; note that some design logics were only identified in few studies that did not meet high-quality criteria. For a more accessible overview of the most salient findings, a summarised version is presented in Table 2.

To complement the comparative matrix (Appendix D) and facilitate the visual identification of co-occurrence patterns and the strength of evidence, the data are also represented in a heatmap bubble chart (Figure 2). In this visualisation, the size of each bubble is proportional to the number of studies that investigated the specific combination of a design logic and an outcome cluster, whilst the colour of the bubble indicates the predominant direction of effect, as defined in the accompanying legend.

Design Features and Pedagogical Strategies Enhancing UX (RQ1)

The analysis confirms that no single design logic guarantees a universally positive UX in ELT-oriented VLEs. A small set of design logics emerged as consistently beneficial for UX.

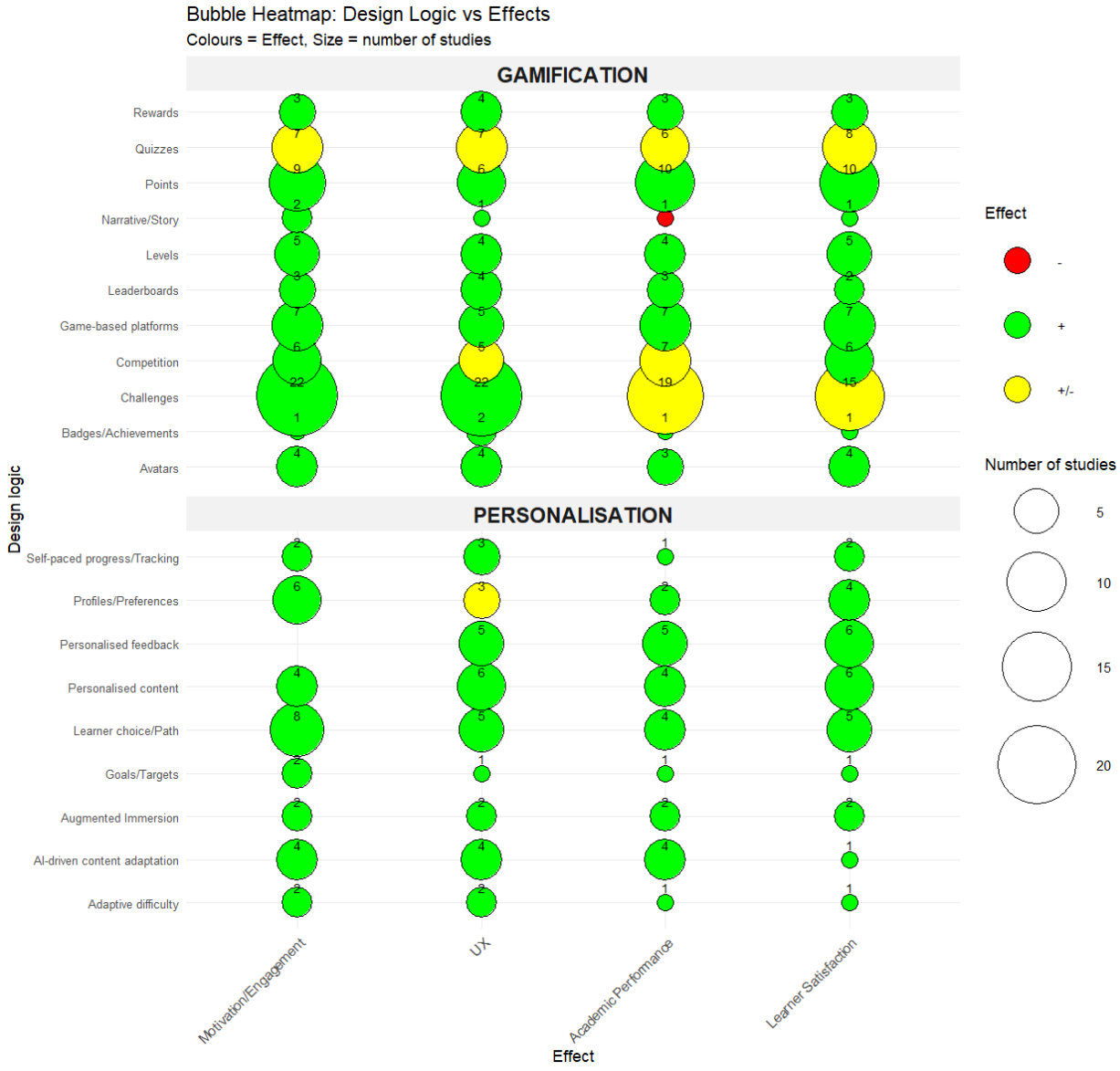
Table 2

Summary of Key Design Logics and their Impact on ELT Outcomes in VLEs.

Design Category	Key Design Logics	Impact on Motivation/ Engagement & UX	Impact on Academic Performance & Learner Satisfaction	Pedagogical Implications
High-impact gamification	Challenges, game-based platforms, points, rewards, levels, avatars, quizzes	Consistently positive	Generally positive, though challenges show mixed results	Effective for engagement; best when integrated with scaffolding
Mixed-result gamification	Competition, leaderboards	Context-dependent; can be demotivating	Mixed results	Requires thoughtful design to prevent exclusion
Consistent personalisation	Learner choice/path, personalised content, personalised feedback	Consistently positive	Consistently positive	Essential for adapting to learner needs
Emerging UX technologies	Artificial intelligence (AI) driven adaptation, augmented immersion	Limited or inconsistent	Limited or inconsistent	Promising; needs further validation

Figure 2

Bubble Heatmap of Design Logics by Educational Outcome and Effect



Challenge-based progression was the clearest pattern (E1, G4, H9, J6, J9) showing that escalating quests calibrated to the learner's level sustained flow and perceived usefulness, especially when difficulty was algorithmically adapted (G4, J6). These studies suggest that precise goal framing and immediate feedback, rather than the game "theme" itself, account for increased engagement.

Points and rewards also improved UX, chiefly by clarifying progress. In D2, F4, H8 and J9 the visibility of score tallies reduced learners' anxiety about their position in the course, while leaderboards (D2, G7, H9) fostered social presence. Nevertheless, H9 indicates that public ranking can undermine enjoyment for low-proficiency groups, implying that competitive displays require opt-out or cohort-based views.

Narrative framing and avatars produced more nuanced gains. When storylines were integrated with learning tasks (E1; avatar-rich H9) participants experienced deeper affective engagement. In opposition, studies where narrative was simply decorative (F4) did not report clear UX effects, emphasising that storytelling must scaffold, not obscure the pedagogy.

Personalisation logics acted as multipliers. Adaptive difficulty (B3, D2, G5, J4), learner-directed pathways (F2), and personalised feedback (H9, I8) collectively enhanced usability and relevance, specifically for novices who risk early abandonment. Remarkably, the highest composite UX scores arose in designs that layered gamified challenges with adaptive support (G5, J6), whereas single-feature deployments, such as quizzes without feedback (H7) or competitive points alone (H9), produced mixed outcomes. Overall, the evidence implies that optimal UX in ELT-oriented VLEs stems from synergistic combinations of challenge, clarity and personalisation rather than from any isolated mechanic.

Effects of Gamification on Motivation and Academic Performance (RQ2)

Gamified elements proved far more reliable in boosting learner motivation than in raising measurable achievement. Most studies (31/46) found at least some motivational benefit, and seven showed a statistically significant increase (e.g. D2, F2, J4, J6, J11, J13). Challenge-based designs were the most powerful driver (22 positive reports), followed by points (e.g. E1, G5, G10, H4, H9, I3), quizzes (e.g. D2, G4, H8), and game-based platforms (e.g. D2, F2, G10, H1). Where challenges were adaptive (G5, J6) or embedded in narrative quests (E1), learners described higher task value and sustained effort. By contrast, quiz-only implementations (H8) and static leaderboards sometimes yielded neutral motivation (G7), especially among low-proficiency cohorts.

Evidence was markedly mixed on academic performance. Only 15 studies detected a clear achievement improvement (e.g. D2, G5, F5, G10, H2, H9), while the modal pattern was "no significant difference" (31 studies). Positive effects clustered around designs that coupled gamified feedback with concept-aligned practice, for instance, adaptive Kahoot! drills (D2) or branching quests with formative checkpoints (J6). However, high-frequency challenge mechanics alone (19; mixed effect) rarely translated into higher test scores, suggesting that motivational lift does not automatically convert to learning outcomes without tight curricular integration and reflection opportunities.

The corpus indicates that gamification is a robust driver for engagement but a contingent one for achievement; its impact on grades depends on whether the mechanics are pedagogically aligned and supported by adaptive scaffolds rather than used as standalone motivators.

PL Strategies and Their Impact on Learning and Satisfaction (RQ3)

Personalisation was generally a stronger lever for learner satisfaction than for achievement, yet three tactics, adaptive feedback loops, learner-directed pathways, and fine-grained content tailoring, showed promise on both fronts. Adaptive difficulty appeared in two studies with positive motivational effects (B1, F5) and led to small but significant improvement in test scores and satisfaction. Learners expressed that tempo-matched tasks reduced frustration and improved control (B1).

Learner choice/path consistently maintained motivation (8; +; e.g. E2, F2, F4) and satisfaction (5; +; e.g. F2, F5), and improved academic achievement (4; +; e.g. F2, J6). Self-selection of sequence, especially when coupled with progress dashboards, fostered agency and deeper strategy use. Personalised content (6 satisfaction +, e.g. B3, F2; 4 performance +, e.g. F5, G5) and personalised feedback were valued for relevance and timeliness; both strategies correlated with higher post-test means in task-aligned vocabulary and grammar modules (H9, F5).

Profiles/preferences displayed limited evidence of positive effects on motivation (B3, F5) and academic performance (F5, G6), but more consistent effects on satisfaction (F5, G6, J9, J10). Self-paced tracking (G6, G10) and emerging AI-driven content adaptation (J2, J3, J4, J6) registered isolated but clear improvements in retention tests, albeit the small sample (≤ 4 studies each) cautions against over-generalisation.

Within PL studies, satisfaction gains slightly outnumbered those in performance (by a ratio of approximately 1.2:1), suggesting that students tend to value perceived personal relevance even when measurable learning improvements are less pronounced. Positive achievement effects surfaced primarily where personalisation was tightly coupled to formative feedback and curricular alignment; standalone preference toggles or static profiles rarely moved the performance needle. The evidence supports personalisation as

a reliable enhancer of perceived quality and a conditional contributor to learning, contingent on its depth of integration with pedagogical goals.

Challenges, Contradictions, and Methodological Gaps (RQ4)

Although the evidence base covers a wide range of contexts, three persistent fault lines (implementation hurdles, divergent findings, and design–method gaps) undermine the generalisability of results. Table 3 summarises these challenges and their categories, detailing the typical issues described, and the representative studies supporting them.

While the overall trends point to motivational and UX benefits, several contradictions emerged across the corpus. Table 4 illustrates three recurring tensions observed in the findings, each backed by specific studies. These contradictions highlight the nuanced impact of gamification and personalisation and reinforce the need for contextual sensitivity in design and evaluation.

Collectively, these gaps explain why motivational effects seldom translate into robust learning effects and why findings vary across cultural and infrastructural contexts. Future work must couple gamified and personalised designs with curricular alignment and teacher mediation to convert engagement into achievement. It should also adopt longer, mixed-method trials with validated UX instruments to capture sustained learning trajectories, and deliberately sample low-resource and younger cohorts to broaden ecological validity. By addressing these methodological shortfalls, forthcoming studies can produce more definitive guidance for designers and teachers seeking to combine gamification and personalisation in ELT-oriented VLEs.

In summary, the findings directly address the RQs of this SLR by identifying key design features (RQ1), analysing the impact of gamification and PL on learner motivation/engagement and academic outcomes (RQ2–RQ3), and highlighting common implementation challenges in VLEs (RQ4).

DISCUSSION

This section interprets this SLR's main findings, situating them within established theory and prior evidence, and then evaluates the methodological strengths and weaknesses that influence those patterns. It closes by outlining practical implications for ELT stakeholders and priority directions for future research.

Key Findings

This SLR reveals that pedagogical fit and interface quality, not the mere presence of gaming or adaptive features, de-

Table 3

Reported Implementation Challenges by Domain

Challenge Domain	Typical Issues	Representative Studies
Technical/usability	Platform instability, limited bandwidth, device restrictions, teachers' and students' digital training and literacy.	A1, D4, E2, F5, F6, G4, G8, G9, G12, H8, J1
Pedagogical alignment	Weak integration with syllabus, teacher workload, insufficient scaffolding	B2, E2, F4, F6, G1, G3, H1
Cultural/linguistic fit	Competitive mechanics clashing with collectivist norms, language-level mismatches, preference for face-to-face instruction.	B3, D4, F5
Sample & attrition	Small cohorts, voluntary dropout, uneven group sizes, need for family support	G12, H8, I2, J3
Measurement validity	Over-reliance on self-report scales, non-validated instruments	G5, I3, I5, J3
Short intervention span	Durations ≤ 4 weeks too brief to detect achievement gains	I2, J1, J3

Table 4

Contradictions in the Effects of Gamification and Personalisation Strategies

Contradiction	Description	Representative studies
Motivation ↑ vs. Performance ↔	Many studies reported high motivation but no significant improvement in academic performance, especially when gamification lacked pedagogical alignment.	H7 (strong motivation, weak test gains), J6 (both gains, due to strong adaptive scaffolding)
Leaderboards: engagement or anxiety?	Leaderboards boosted engagement through social presence, but sometimes harmed low-performing students' experience, particularly in collectivist contexts.	D2; G7 (positive), H9 (negative for low-level learners)
Narrative coherence vs. superficial appeal	Integrated storylines enhanced UX only when they supported the learning task; decorative narratives had no positive effect.	E1; I5 (positive), F4 (neutral)

termine learning impact on ELT-oriented VLEs. When challenges, points, or adaptive paths were embedded in clear interfaces that minimised extraneous cognitive load and delivered immediate, informative feedback (E2, G4, J7), learners perceived fluent navigation and higher task value. By contrast, the same mechanics led to a poor or neutral UX where apps were pedagogically weak (H4), or where basic usability was hampered by clunky navigation and distracting adverts (A1). This pattern aligns with Cognitive Load Theory (Sweller et al., 2011) and extends Dicheva et al.'s (2015) observation that usability lapses can nullify gamification benefits.

Motivational enhancements proved broad but fragile. Quizzes, points, and escalating challenges reliably triggered short-term behavioural engagement, confirming Self-Determination Theory's claim that clear goals and competence signals energise learners (Ryan & Deci, 2000). However, achievement outcomes emerged only when those mechanics were aligned with the curriculum and paired with formative feedback (D2, G5, J6). The recurring engagement–learning gap mirrors Looyestyn et al. (2017) and Koivisto & Hamari (2018) and suggests that extrinsic rewards alone rarely prompt deep processing.

The mixed appeal of social-comparison mechanics highlights the role of learner disposition. While leaderboards motivated some active users (D2, H9), this effect was not universal, with other learners showing disinterest (H9) or scepticism towards peer feedback (A1). This tension, where social features can both engage and alienate, reflects concerns noted by Antonaci et al. (2019) regarding their potential for negative affective outcomes. By contrast, well-integrated narratives achieved more consistent success, fostering immersive and supportive environments that focused learners on the task (E1, I5). These findings indicate that gamification's efficacy is contingent on an implementation that carefully balances learners' needs for autonomy, competence, and relatedness.

PL strategies proved the strongest dual impact when they combined adaptive feedback, learner choice, and tailored content. Such designs simultaneously satisfied autonomy and competence needs, leading to both higher satisfaction (F2, H9) and modest but significant performance results (B1, J6). Nevertheless, PL success depended on data quality, instructor mediation, and infrastructure; in low-resource contexts, adaptive algorithms often defaulted to "one-size-fits-all," muting their benefits.

Finally, contextual diversity (urban/rural schools, adolescent/adult learners, differing digital literacy) explains the diverse evidence. High-bandwidth and teacher-supported environments converted motivation into achievement, whereas self-paced or low-connectivity settings rarely did. Thus, no single mechanic guarantees success; meaningful impact arises only when interface clarity, adaptive scaffolds, and assessment feedback are consistent with learners' profiles and institutional realities.

Comparison with Prior Studies and Reviews

These results broadly confirm earlier gamification and PL syntheses while adding an ELT-specific perspective. As Dicheva et al. (2015) and Hamari et al. (2014) state, feedback-rich mechanics (quizzes, points, challenges) consistently increase behavioural engagement; nevertheless, echoing Sailer and Homner's (2020) meta-analysis, our data show that such increases in motivation rarely translate into academic improvement unless formative feedback and mastery thresholds are embedded.

Unlike most prior reviews, this SLR treated UX as an independent outcome. This revealed that interface clarity and cognitive-load management, rather than any single mechanic, drive positive UX, supporting the usability emphasis proposed by Seaborn and Fels (2015). For PL strategies, our findings align with recent meta-analytic evidence; adaptive difficulty and personalised content lead to the most reliable academic benefits (Fraulini et al., 2024), though their motivational value weakens when learners sense a loss of autonomy (Fong et al., 2019).

Finally, we corroborate Koivisto and Hamari's (2019) claim that context and implementation fidelity moderate outcomes: social-comparison features benefited high achievers but discouraged novices, a pattern also observed in Antonaci et al.'s (2019) review. Our ELT lens highlights that language-proficiency gaps intensify these divergences, pointing to a need for tiered or anonymous ranking systems in language classrooms.

Limitations

Despite following PRISMA 2020 guidelines, this review has several limitations. Grey literature was excluded, and searches were restricted to four databases covering 2015–2024, which could bias the evidence by favouring studies that are actually published, privileging those available in English, and missing research that takes longer to reach print.

Moreover, study coverage was uneven. Most investigations centred on higher-education contexts, with scant attention to secondary education or low-resource contexts. Geographically, research clustered in a few parts of Asia and Europe, leaving other areas, especially the Americas and Africa, underrepresented. Consequently, many of the design strategies discussed rest on assumptions such as stable broadband, individual devices, and competitive learning norms that seldom hold in those contexts; what works in well-resourced universities may be impractical or even counter-productive in rural secondary schools where con-

nectivity is deficient, devices are shared, and collaboration is valued over competition. Until these approaches are tested under such constraints, their broader applicability remains uncertain.

Methodological rigour was also limited: only two RCTs appeared among 46 studies; the rest relied on quasi-experimental or single-group designs with small convenience samples, a pattern prone to novelty effects and low external validity. Motivation and UX were often assessed with ad hoc, non-validated instruments. Since UX is inherently multidimensional, this reliance makes the UX findings exploratory and should temper any comparison with motivation or performance outcomes. The absence of validated UX tools thus reduces the reliability of conclusions. Similarly, academic results ranged from isolated vocabulary quizzes to full proficiency tests, hampering cross-study comparability. Finally, most studies gave only limited descriptions of how their gamification or personalisation elements were carried out, making it difficult to judge whether those features were delivered as intended.

Implications

Building on the evidence, the key implications are directed to three stakeholder groups. The emphasis is on how gamified and PL features are designed, integrated, and evaluated, rather than just being present. Educators should treat game mechanics and adaptive tools as scaffolds, intervening with formative feedback when dashboards reveal stagnation and using anonymised or tiered competition to protect novice confidence. Instructional designers must privilege usability over feature counts: mobile-first interfaces, offline caching, and transparent adaptive rules reduce cognitive load and allow teachers to override algorithmic decisions when needed.

Researchers can consolidate the evidence base through long-term, multi-site studies that couple validated UX and motivation scales with curriculum-aligned performance tests, paying particular attention to secondary schools in underrepresented regions such as Latin America and Africa. Reliable instruments, such as the User Engagement Scale, AttrakDiff, or the User Experience Questionnaire for UX, and the Intrinsic Motivation Inventory, MEEGA+, or the Motivated Strategies for Learning Questionnaire for motivation, should be combined with standardised language assessments (e.g., Cambridge Progress Tests, TOEFL Junior modules, or focused grammar cloze tasks). These measures need to be applied both formatively and summatively and clearly linked to specific domains (e.g., grammar accuracy versus communicative competence). Future work should also incorporate grey literature, preregister multi-site randomised controlled trials, and publish complete contextual and implementation details. Sharing open data and detailed implementation notes will shed light on how context, design quality, and pedagogical mediation transform engagement into lasting language gains, thereby enhancing the rigour and generalisability of the findings.

CONCLUSION

This SLR advances the field by showing that design quality and contextual fit, not only the presence of gamified or personalised features, determine learning impact in ELT-oriented VLEs. Treating UX as an independent outcome demonstrates that interface clarity and cognitive-load management support positive UX and mediate the translation of short-term motivation into durable achievement. It also clarifies why the same mechanics can benefit high achievers while alienating novices: cultural norms, language proficiency, and teacher mediation jointly shape learners' responses. The evidence suggests that meaningful gains arise only when adaptive scaffolds, formative feedback, and ethical competition are embedded in clear, low-load interfaces that respect local resource constraints.

Future research should move beyond short, single-site pilots to longitudinal, mixed-method studies that combine validated UX instruments, behavioural analytics, and curriculum-aligned performance tests. Particular attention is needed at the intersection of gamification and personalisation in low-resource settings, where mobile-first, offline-capable designs and teacher-controlled dashboards could foster more equitable benefits. Multi-site RCTs and well-reported quasi-experiments, when randomisation is unfeasible, will be critical for tracing whether initial motivational boosts mature into sustained language proficiency and learner autonomy.

For educational policy and practice, the findings establish three priorities. First, procurement guidelines should privilege usability audits, mobile resilience, and transparent adaptive engines over feature counts. Second, professional development must equip teachers to interpret learning analytics, modulate competitive elements, and integrate adaptive automation with scaffolded dialogue. Third, funding schemes and accreditation frameworks should promote rigorous trials in underrepresented regions and secondary classrooms, ensuring that evidence-based, context-sensitive VLEs become a realistic option for all English language learners.

DISCLAIMER

Some parts of this manuscript were refined for language and style using ChatGPT, under the authors' direct supervision and review. All content was carefully checked and approved by the authors to ensure academic rigour and consistency with the purpose of the SLR.

DECLARATION OF COMPETING INTEREST

None declared.

AUTHORS' CONTRIBUTION

Myriam Tatiana Velarde Orozco: conceptualization; data curation; formal analysis; funding acquisition; methodology; project administration; visualization; writing – original draft; writing – review & editing.

Bárbara Luisa De Benito Crosetti: formal analysis; investigation; methodology; resources; software; supervision; writing – original draft.

REFERENCES

Al-Busaidi, K. A., & Al-Shihi, H. (2012). Key factors to instructors' satisfaction of learning management systems in blended learning. *Journal of Computing in Higher Education*, *24*(1), 18–39. https://doi.org/10.1007/s12528-011-9051-x

Ali, M., Wahab, I. B. A., Huri, H. Z., & Yusoff, M. S. (2024). Personalised learning in higher education for health sciences: a scoping review protocol. *Systematic Reviews*, *13*, Article 99. https://doi.org/10.1186/s13643-024-02478-4

Antonaci, A., Klemke, R., & Specht, M. (2019). The effects of gamification in online learning environments: A systematic literature review. *Informatics*, *6*(3), 32. https://doi.org/10.3390/informatics6030032

Bernacki, M. L., Greene, M. J., & Lobczowski, N. G. (2021). A systematic review of research on personalized learning: Personalized by whom, to what, how, and for what purpose(s)? *Educational Psychology Review*, *33*(4), 1675–1715. https://doi.org/10.1007/s10648-021-09615-8

Chaichumpa, S., Wicha, S., & Temdee, P. (2021). Personalized learning in a virtual learning environment using modification of objective distance. *Wireless Personal Communications*, *118*(3), 2055–2072. https://doi.org/10.1007/s11277-021-08126-7

Cho, I., Yeo, J., Hwang, G., & Yang, H. (2022). Impact of a virtual environment on learning effectiveness, motivation, cognitive load, and group self-efficacy of elementary school students in collaborative learning. *Educational Technology Research and Development*, *70*(6), 2145–2169. https://doi.org/10.1007/s11423-022-10159-z

Deterding, S., Dixon, D., Khaled, R., & Nacke, L. (2011). From game design elements to gamefulness: Defining "gamification". In *Proceedings of the 15th International Academic MindTrek Conference: Envisioning Future Media Environments* (pp. 9–15). Association for Computing Machinery. https://doi.org/10.1145/2181037.2181040

Dicheva, D., Dichev, C., Agre, G., & Angelova, G. (2015). Gamification in education: A systematic mapping study. *Educational Technology & Society*, *18*(3), 75–88.

du Plooy, E., Casteleijn, D., & Franzsen, D. (2024). Personalized adaptive learning in higher education: A scoping review of key characteristics and impact on academic performance and engagement. *Heliyon*, *10*(21), e39630. https://doi.org/10.1016/j.heliyon.2024.e39630

Fong, C. J., Patall, E. A., Vasquez, A. C., & Stautberg, S. (2019). A meta-analysis of negative feedback on intrinsic motivation. *Educational Psychology Review*, *31*(1), 121–162. https://doi.org/10.1007/s10648-018-9446-6

Fraulini, N. W., Marraffino, M. D., Garibaldi, A. E., Johnson, C. I., & Whitmer, D. E. (2024). Adaptive training instructional interventions: A meta-analysis. *Military Psychology*. Advance online publication. https://doi.org/10.1080/08995605.2024.2377884

Fredricks, J. A., Blumenfeld, P. C., & Paris, A. H. (2004). School engagement: Potential of the concept, state of the evidence. *Review of Educational Research*, *74*(1), 59–109. https://doi.org/10.3102/00346543074001059

Gevorgyan, S. (2024). The use of adaptive learning technologies in e-learning for inclusive education: A systematic review. *E-Learning Innovations Journal*, *2*(1), 90–107. https://doi.org/10.57125/ELIJ.2024.03.25.05

Hamari, J., Koivisto, J., & Sarsa, H. (2014). Does gamification work? – A literature review of empirical studies on gamification. In *Proceedings of the 47th Hawaii International Conference on System Sciences* (pp. 3025–3034). IEEE. https://doi.org/10.1109/HICSS.2014.377

Helvich, J., Novak, L., Mikoska, P., & Hubalovsky, S. (2023). A Systematic Review of Gamification and Its Assessment in EFL Teaching. *International Journal of Computer-Assisted Language Learning and Teaching (IJCALLT)*, *13*(1), 1–21. https://doi.org/10.4018/IJCALLT.322394

Hollister, B., Nair, P., Hill-Lindsay, S., & Chukoskie, L. (2022). Engagement in online learning: Student attitudes and behaviour during COVID-19. *Frontiers in Education*, *7*, Article 851019. https://doi.org/10.3389/feduc.2022.851019

Hong, Q. N., Pluye, P., Fàbregues, S., Bartlett, G., Boardman, F., Cargo, M., Dagenais, P., Gagnon, M. P., Griffiths, F., Nicolau, B., O'Cathain, A., Rousseau, M. C., & Vedel, I. (2018). *Mixed Methods Appraisal Tool (MMAT), version 2018*. Canadian Intellectual Property Office, Industry Canada. http://mixedmethodsappraisaltoolpublic.pbworks.com/

Hung, H.-T., Yang, J.-C., Hwang, G.-J., Chu, H.-C., & Wang, C.-C. (2018). A scoping review of research on digital game-based language learning. *Computers & Education*, *126*, 89–104. https://doi.org/10.1016/j.compedu.2018.07.001

Klašnja-Milićević, A., Vesin, B., Ivanović, M., & Budimac, Z. (2011). E-learning personalization based on hybrid recommendation strategy and learning-style identification. *Computers & Education*, *56*(3), 885–899. https://doi.org/10.1016/j.compedu.2010.11.001

Koivisto, J., & Hamari, J. (2019). The rise of motivational information systems: A review of gamification research. *International Journal of Information Management*, *45*, 191–210. https://doi.org/10.1016/j.ijinfomgt.2018.10.013

Li, X., Xia, Q., Chu, S. K. W., & Yang, Y. (2022). Using gamification to facilitate students' self-regulation in e-learning: A case study on students' L2 English learning. *Sustainability*, *14*(12), 7008. https://doi.org/10.3390/su14127008

Looyestyn, J., Kernot, J., Boshoff, K., Maher, C., & Vandelanotte, C. (2017). Does gamification increase engagement with online programs? A systematic review. *PLOS ONE*, *12*(3), e0173403. https://doi.org/10.1371/journal.pone.0173403

Majuri, J., Koivisto, J., & Hamari, J. (2018). Gamification of education and learning: A review of empirical literature. In *Proceedings of the 2nd International GamiFIN Conference 2018* (pp. 11–19). CEUR-WS.org.

Marín-Juarros, V. I. (2022). La revisión sistemática en la investigación en tecnología educativa: Observaciones y consejos. *Revista Interuniversitaria De Investigación En Tecnología Educativa*, (12), 62-79. https://doi.org/10.6018/riite.533231

Munn, Z., Peters, M. D. J., Stern, C., Tufanaru, C., McArthur, A., & Aromataris, E. (2018). Systematic review or scoping review? Guidance for authors when choosing between a systematic or scoping review approach. *BMC Medical Research Methodology*, *18*, Article 143. https://doi.org/10.1186/s12874-018-0611-x

Nazempour, R., & Darabi, H. (2023). Personalized learning in virtual learning environments using students' behaviour analysis. *Education Sciences*, *13*(5), 457. https://doi.org/10.3390/educsci13050457

Page, M. J., McKenzie, J. E., Bossuyt, P. M., Boutron, I., Hoffmann, T. C., Mulrow, C. D., & Moher, D. (2021). The PRISMA 2020 statement: An updated guideline for reporting systematic reviews. *BMJ*, *372*, n71. https://doi.org/10.1136/bmj.n71

Ryan, R. M., & Deci, E. L. (2000). Self-determination theory and the facilitation of intrinsic motivation, social development, and well-being. *American Psychologist*, *55*(1), 68–78. https://doi.org/10.1037/0003-066X.55.1.68

Sailer, M., & Homner, L. (2020). The gamification of learning: A meta-analysis. *Educational Psychology Review*, *32*(1), 77–112. https://doi.org/10.1007/s10648-019-09498-w

Sanchis-Font, R., Castro-Bleda, M. J., González, J., Pla, F., & Hurtado, L. (2021). Cross-domain polarity models to evaluate user eXperience in E-learning. *Neural Processing Letters*, *53*(5), 3199–3215. https://doi.org/10.1007/s11063-020-10260-5

Seaborn, K., & Fels, D. I. (2015). Gamification in theory and action: A survey. *International Journal of Human-Computer Studies*, *74*, 14–31. https://doi.org/10.1016/j.ijhcs.2014.09.006

Shamseer, L., Moher, D., Clarke, M., Ghersi, D., Liberati, A., Petticrew, M., Shekelle, P., Stewart, L. A., & PRISMA-P Group. (2015). Preferred reporting items for systematic review and meta-analysis protocols (PRISMA-P) 2015: Elaboration and explanation. *BMJ*, *349*, g7647. https://doi.org/10.1136/bmj.g7647

Sweller, J., Ayres, P., & Kalyuga, S. (2011). *Cognitive load theory*. Springer. https://doi.org/10.1007/978-1-4419-8126-4

Venkatesh, V., Thong, J. Y. L., & Xu, X. (2012). Consumer acceptance and use of information technology: Extending the unified theory of acceptance and use of technology. *MIS Quarterly*, *36*(1), 157–178. https://doi.org/10.2307/41410412

Walkington, C., & Bernacki, M. L. (2020). Appraising research on personalized learning: Definitions, theoretical alignment, advancements, and future directions. *Journal of Research on Technology in Education*, *52*(3), 235–252. https://doi.org/10.1080/15391523.2020.1747757

Wichadee, S., & Pattanapichet, F. (2018). Enhancement of performance and motivation through application of digital games in an English language class. *Teaching English with Technology*, *18*(1), 77–92.

Zainuddin, Z., Chu, S. K. W., Shujahat, M., & Perera, C. J. (2020). The impact of gamification on learning and instruction: A systematic review of empirical evidence. *Educational Research Review*, *30*, Article 100326. https://doi.org/10.1016/j.edurev.2020.100326

Zhang, S., & Hasim, Z. (2023). Gamification in EFL/ESL instruction: A systematic review of empirical research. *Frontiers in Psychology*, *13*, Article 1030790. https://doi.org/10.3389/fpsyg.2022.1030790

APPENDIX A

DESCRIPTIVE matrix of included studies

This appendix offers a structured descriptive profile of the 46 studies reviewed in this SLR. The matrix summarises essential bibliographic and contextual information to support understanding of the research landscape and the diversity of implementations observed. It includes:

- Study identification and metadata (reference code, title, authors, publication year, journal, and source database).
- Geographical and educational context (country, educational level, and setting).
- Participant information (sample size and learner profile, including age, English proficiency, and digital competence when available).
- Methodological characteristics (study design and type).
- VLE description, where described.

Due to its extensive nature, the complete matrix is available in Figshare:

https://doi.org/10.6084/m9.figshare.29316077

APPENDIX B

MMAT ASSESSMENT OF Included Studies

This appendix presents the results of the quality appraisal conducted using the MMAT 2018. Each study was assessed according to its methodological category: qualitative, quantitative (randomised, non-randomised, or descriptive), or mixed methods.

Due to its extensive nature, the complete matrix is available in Figshare:

https://doi.org/10.6084/m9.figshare.29318828

APPENDIX C

Analytical matrix of included studies

This appendix offers a structured narrative summary of each study included in the review. For every article, it compiles key information on the VLE used, specific gamification and personalisation features (when specified), and the educational outcomes assessed such as motivation, engagement, UX, academic achievement, and learner satisfaction. It also captures author-reported findings, links between strategies and results, and implementation challenges.

Due to its extensive nature, the complete matrix is available in Figshare:

https://doi.org/10.6084/m9.figshare.29318840

APPENDIX D

Comparative Matrix of Design Logics and Educational Outcomes

	Design Logic	Motivation/ Engagement	UX	Academic Performance	Learner Satisfaction
Gamification	Avatars	4; +; H9	4; +; H9	3; +; H9	4; +; H9
	Badges/Achievements	1; +; F5	2; +; J9	1; +; F5	1; +; F5
	Challenges	22; +; H7	22; +; H7	19; +/-; H7	15; +/-; H7
	Competition	6; +; F5	5; +/-; F5	7; +/-; F5	6; +; G8
	Game-based platforms	7; +; H8	5; +; H8	7; +; H8	7; +; H8
	Leaderboards	3; +; H9	4; +; H9	3; +; H9	2; +; H9
	Levels	5; +; I2	4; +; I2	4; +; J1	5; +; H9
	Narrative/Story	2; +; E1	1; +; E1	1; -; E1	1; +; E1
	Points	9; +; J3	6; +; H8	10; +; F5	10; +; F5
	Quizzes	7; +/-; H8	7; +/-; H7	6; +/-; J9	8; +/-; H8
	Rewards	3; +; G10	4; +; G10	3; +; G10	3; +; G10
Personalisation	Adaptive difficulty	2; +; F5	2; +; F5	1; +; B1	1; +; B1
	Goals/Targets	2; +; F5	1; +; B1	1; +; B1	1; +; B1
	Learner choice/Path	8; +; F4	5; +; B1	4; +; F5	5; +; F5
	Personalised content	4; +; F5	6; +; F5	4; +; F5	6; +; F5
	Personalised feedback	6; +; H9	5; +; I8	5; +; B1	6; +; H9
	Profiles/Preferences	3; +; F5	3; +/-; F5	2; +; F5	4; +; F5
	Self-paced progress/ Tracking	2; +; G6	3; +; F5	1; +; G10	2; +; J10
	AI-driven content adaptation	4; +; J2	4; +; J3	4; +; J4	1; +; J2
	Augmented immersion	2; +; J1	2; +; J1	2; +; J1	2; +; J1

https://doi.org/10.17323/jle.2025.24203

Reading Images for Knowledge Building: Analyzing Infographics in School Science: A Book Review

Chau Hoang Vo, Tin Thanh Nguyen

Quy Nhon University, Vietnam

Reading images for knowledge building: Analyzing infographics in school science, Written by J.R. Martin and Len Unsworth, New York: ROUTLEDGE, 2024, 304 pp., (Print Book), ISBN: 9780367759216

ABSTRACT

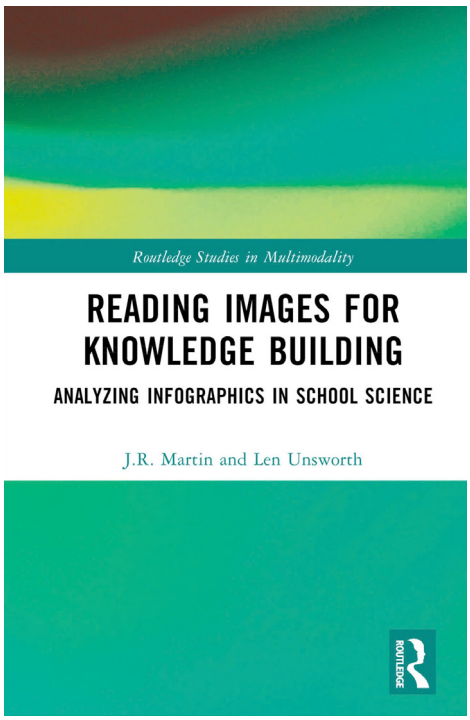
The book "Reading Images for Knowledge Building: Analyzing Infographics in School Science" by J. R. Martin and Len Unsworth, published by Routledge in 2024, explores new perspectives on visual grammar derived from Systemic Functional Linguistics and their application in the multimodal discourse of school science. It examines key components of infographics, specifically image mass (technicality, iconization, aggregation) and image presence (explicitness, affiliation, congruence). The book enriches our theoretical understanding of visual grammar and provides strategies for the effective use of infographics in classrooms.

KEYWORDS

infographics, school science, image mass, image presence, visual grammar, book review

INTRODUCTION

"Reading Images for Knowledge Building: Analyzing Infographics in School Science" is divided into four main parts. Part I revisits Kress and van Leeuwen's foundational concepts, emphasising the need for framework refinement to address issues concerning technical images, and highlighting the importance of captions, annotations and text blocks. Part II examines technicality, iconization, and aggregation - components of image mass. Part III discusses image presence, including visual explicitness, affiliation and congruence. Part IV combines mass and presence for infographic selection and structuring to enhance knowledge-building in junior and senior high school science education.



Part I: Disciplinary Discourse for Knowledge Building

In the introductory chapter and Part I, the authors revisit the foundational knowledge related to visual grammar (VG) developed by Kress and van Leeuwen in Reading Images: The Grammar of Visual Design across four editions (1990, 1996, 2006, 2021), particularly pointing out the difficulties in interpretation challenges students may face with scientific images in terms of technicality and abstraction using the framework of Kress and van

Citation: Vo, C. H., & Nguyen, T. T. (2025). [Review of the book Reading images for knowledge building: Analyzing infographics in school science, by J.R. Martin & Len Unsworth]. *Journal of Language and Education, 11*(2), 175-179. https://doi.org/10.17323/jle.2025.24203

Correspondence:
Chau Hoang Vo,
vohoangchau@qnu.edu.vn

Received: December 11, 2024
Accepted: June 10, 2025
Published: June 30, 2025

Leeuwen (2021). They also emphasise the hitherto neglected but crucial role of captions, annotations, and text blocks that are incorporated within science images in complementing the visual portrayals to represent science concepts (Martin, 2020; Unsworth et al., 2022). As well as dealing with systemic functional descriptions of the meaning-making resources of scientific images, Martin and Unsworth draw on studies that have applied and adapted those descriptions in science pedagogy (Tang, 2020; Tang et al., 2019). They include sections illustrating how the frameworks they propose can be applied in science teaching. This book provides a disciplinary approach across science subjects, revealing the limitations of VG in previous editions.

Part II: Image Complexity – Mass

Building on the foundational concepts, Part II of the book introduces one of the two major aspects of image, *mass*, or image complexity, explored through various dimensions, including *technicality*, *iconization* and *aggregation*. These aspects of mass deal respectively with the ways that ideational, interpersonal and textual meanings are distilled in the semantically dense portrayal of information in infographics.

Part II consists of three chapters, beginning with the exploration of *technicality*, employed to describe the meaning-making resources (image and language) that represent field – a collection of activities that shape our life experiences consisting of the activities themselves and their relevant entities (Doran & Martin, 2021; Martin, 1992, 2020). The book examines the construal of field in terms of: (i) activity (whether it is unfolding over time or not unfolding (i.e. represented as a single action); (ii) classification (types and subtypes) of entities; (iii) composition (whole and part relations) and (iv) property (qualities of entities and activities and their spatio-temporal locations). With the construal of activity and composition, the authors first present its imagic construal, and then continue with annotation, or how language is presented in infographics. However, the structure of the sections on classification and property is different, i.e. images and language are not explored separately; instead, they are intertwined. It is also notable that each aspect of field construal (i.e. activity, classification, composition and property) has its own framework and each framework is summarised as a single figure, making it easier for readers to navigate the available options as they progress through the chapter.

Although the frameworks are conceptually useful, the text stops short of discussing which options are most commonly used, in what contexts and suitable for which levels of learners, or what the advantages and limitations of each option are. In addition, while this section offers a rich typology of imagic construals, it leans heavily on descriptive exemplification. Its explanatory power would be strengthened by a more sustained reflection on the significance of these distinctions for teaching and learning, and their practical implications for educators and textbook designers.

Chapter 4 introduces the term *iconization*, characterised by Martin (2020) as a process that highlights the social values of an event or entity, or its axiological meaning. The framework for analysing the social-semiotic construction of bonding icons (bondicons) is based on the works of Tann (2010, 2013) and Zappavinga and Martin (2018). As stated by the authors, although the original model developed by Tann comprises Gemeinschaft, Doxa and Oracle, only Oracle, a concept referring to famous people and things as bondicons, is relevant in school science discourse. In addition, while the adapted framework features both icons and texts, the discussion of bondicons in this book focuses solely on icons, with subcategories of guru (for famous scientists) and objects (images of animals, plants, artefacts or abstract symbols). Examples of bondicons include the nuclear radiation warning symbol, the DNA double helix or dead animals due to plastic bags. The authors show how students experience bondicons through ways in which such images are represented in textbooks and are established as bondicons. There are different kinds of bondicons aligning groups of students in various ways, namely those oriented to the celebrations of science figures and discoveries, calls for action, and aiming to influence public opinion. The chapter concludes with suggestions for the application of the concept of iconization in science teaching, helping students critically engage with the values embedded in scientific imagery and understand how these visuals shape their alignment with particular scientific perspectives.

However, the strength of this semiotic approach also raises questions about how readily students identify with or critically engage with these bondicons, particularly when their prior experiences or cultural affiliations diverge from those assumed by textbook authors. While the taxonomy of bondicons is conceptually robust, its pedagogical uptake would benefit from more explicit guidance on how teachers might navigate the varying interpretations or resistances that these images could evoke.

Chapter 5, deals with the ways two or more of the aspects of field representation (activity, composition, classification and property) are combined in infographics in a process referred to as *aggregation*. The chapter begins with the introduction of macro- and micro-groupings, with macro-groupings focusing on the layout of the global elements of the infographic (which may be accompanied by co-text), while micro-groupings characterise the relations between component images and verbiage. The chapter then moves on to connect these types of groupings to the frameworks developed in Chapter 3, leading to the formation of a system of aggregation involving two distinct approaches *[accumulation]* across macro-groups and *[integration]* within one macro-group. The authors consequently use this system to observe four textbook representations of global warming. This chapter also concludes with an outline of suggested pedagogical applications, emphasising the need for explicit

attention to infographics as an integral part of science pedagogy.

What stands out in this chapter is its reconceptualisation of aggregation, not as a mere technical layout, but as a semiotic and pedagogical issue with real implications for meaning-making in science education. The authors offer a systematic model for analysing infographics and prompt us to rethink how knowledge is hierarchized and accessed. The repeated inconsistencies and omissions in the textbook examples highlight the need for science educators to act as both interpreters and designers of visual meaning. This calls for a pedagogical shift in which infographics should be seen not just as illustrations, but as complex meaning-making tools that students must learn to analyse and construct.

Part III: Image Recognizability – Presence

Shifting from semantic density in Part II to perceptual accessibility – the second key dimension of image analysis – Part III of the book offers a compelling exploration of how images and infographics can be designed and utilised to convey scientific concepts in a manner that is not only accessible but readily understandable for students. The key dimensions of this "recognizability" (or *presence* in the authors' terms) are *explicitness*, *affiliation* and *congruence* – three pillars that underpin the accessibility of visual representations for students (Martin & Unsworth, 2024, p. 140).

Explicitness refers to the nature and extent of clarity and precision in depicting scientific phenomena and their contexts. The main aspects of explicitness are *completeness, environment* and *discernibility*. *Completeness* concerns whether the representation of a phenomenon is truncated either by being cut off, such as an image of just the top of the barrel of a Bunsen burner to show the flame features, or cropped, such as the image of a bicycle wheel showing only the section with the gears. *Environment* concerns whether the background or context of a phenomenon is depicted and to what extent, such as an isolated image of the kidney function without any contextual depiction of the renal system. *Discernability* concerns clarity and precision of depiction. In the context of explicitness, the authors illustrate how explicit imagery significantly aids in deepening student understanding (e.g. Chidrawi et al., 2013; Zedalis et al., 2018). This facilitates a direct and clear transmission of information, where each infographic or image acts as a precise visual explanation that complements and reinforces textual content.

Affiliation underscores the significance of engaging students through relatable and appealing content depiction. The authors explore how the physical and visual characteristics of infographics, such as colour, layout, and iconography, can dramatically influence the learner's engagement and understanding (e.g. Kinnear & Martin, 2021a; Kinnear & Martin, 2021b). They also discuss affiliation in terms of infotainment, provided principally by the inclusion of cartoons in science texts. Some such cartoons prompt or challenge students to extend their understanding of science concepts, while others simply add relevant or sometimes, unrelated whimsical elements to concept portrayal (e.g. William & Gaton, 2013; Kinnear, 2017; Lofts, 2015). By making content more "alive" and present to the viewer, educational materials can foster a deeper connection with the subject matter, thereby facilitating a more intuitive and impactful learning experience. Through practical examples, Martin and Unsworth illustrate how these principles can be effectively applied in the science curriculum, aiming to equip educators and content creators with the knowledge to design infographics that not only convey information accurately but are also easily interpreted by students.

Congruence concerns the extent to which the representation of phenomena aligns with their perceptual reality. Images range from realistic to highly abstract. While naturalistic images are important for relating target concepts to students' lived realities, reduced congruence is frequently required to clearly represent the essential features of the phenomenon being studied. The authors identify and illustrate parameters along which congruence varies. These include colour and two- or three-dimensional perspective as well as factors such as *essentialization*, *vision*, *proportionality*, *reconfiguration* and *genesis*. *Essentialization* refers to strategic deletion from the representation of elements of the phenomenon considered peripheral to the particular pedagogic focus, which for example, results in highly simplified depictions of cells. *Vision* can refer to magnification or other means of depicting internal components that are not normally visible, such as x-ray or cross-sections. Sometimes the proportion dimensions of phenomena are distorted in order to show relationships, such images of DNA showing chromosomes as if they were similar in size to the double helix. *Reconfiguration* can occur in images of the functioning of the ear with cochlea "rolled out" (Martin & Unsworth, 2024, p. 189) as a horizontal tube and *genesis* refers to a single image of a plant at various growth stages from seed to fully grown.

Part IV: Applying Image Analyses for Knowledge Representation

In Part IV of the book, the authors consider both the elements of image complexity (mass) from Part II and recognizability (presence) from Part III when selecting and ordering infographics to enhance knowledge-building in junior and senior high school science. This gives rise to the question: When is it suitable to use a straightforward, visually appealing infographic, and when does the complexity of the subject matter necessitate an infographic with less visually congruent elements to facilitate deeper understanding?

The analysis of selected infographics about the greenhouse effect in junior and senior high school textbooks informs the discussion of how to balance the orientation to mass or presence in using infographics with students at different

stages of science learning. Results of the analyses show that in some cases, senior high school infographics are more accessible than those in junior high school texts and could be more suitable for use in junior classes, while in other cases, there are no clear distinctions between images occurring in textbooks at both levels. This surprising reversal calls attention to the common assumption that complexity increases linearly across grade levels and highlights the need for more intentional design and use of visuals.

The authors have also pointed out that the lack of comprehensive curriculum guidance regarding the suitable level of technical detail for junior and senior high school levels, coupled with the inconsistency in textbook infographics across grade levels, places the responsibility of determining an appropriate learning progression for students squarely on teachers. While the chapter does not fully critique the systemic issues behind these vague curriculum statements, it offers practical strategies for navigating this ambiguity. They advise teachers to balance the two elements when choosing infographics, especially in the initial stages of knowledge building, as well as scaffolding students' reading strategies through "deconstructive interpretive practices" (Martin & Unsworth, 2024, p. 214). This involves explicit teaching of the visual communication strategies used for representing the interaction of the various elements of field in order to convey scientific concepts. The suggestion to map meaning across modalities (image, annotation, caption, and text) and to critically examine what is present, omitted or assumed in infographics is particularly valuable, offering teachers a systematic approach to cultivating students' multimodal literacy.

Part IV also examines the interplay of mass and presence in infographics to show how the techniques of multimodal representation vary across the science sub-disciplines, and hence the distinctive multimodal literacies students need in the different subject areas. This part is particularly useful as it offers teachers a thoughtful set of guiding questions related to mass and presence for their own reflection when selecting infographics and for students' critical engagement when interpreting or designing them.

CONCLUSION

Overall, "Reading Images for Knowledge Building: Analyzing Infographics in School Science" by Martin and Unsworth (2024) provides new insights into how language and image resources are deployed in infographics to communicate science concepts. The authors' exploration into how multimodal literacy is interwoven within science education illuminates the process of both creating and interpreting infographics, fostering an enriched learning environment, and facilitating the nurturing of critical thinking and knowledge exploration. The book stands not just as an invaluable resource but also as a call to action for leveraging the substantial potential of infographics for enhancing teaching and learning in school science.

Unlike existing studies that simply analyse science images using VG (e.g., Christidou et al., 2023; Fernández-Fontecha et al., 2018) or testing the effectiveness of infographics in science classes (e.g., Agley et al., 2021; Basco, 2020; Menendez et al., 2024; Scott & Jenkinson, 2020), Martin and Unsworth's work offers a distinct contribution by addressing the unique semiotic nature of scientific infographics and proposing purpose-built tools for both analysing and designing them within science education. Their approach not only bridges disciplinary literacy and multimodal analysis but also foregrounds students' active meaning-making through both interpretation and production of infographics, which is an area less emphasized in prior research.

Although the proposed frameworks are comprehensive and suggest potential for application beyond science, the book does not explicitly explore how scalable or adaptable these tools might be in other disciplinary contexts. Furthermore, while the authors provide numerous prompts and pedagogical suggestions throughout the book to support student interaction with infographics, the absence of empirical classroom evidence, such as classroom-based trials or implementation studies, leaves open questions about the practical feasibility of incorporating such detailed analytical tools in real instructional settings. The addition of small-scale empirical insights would have strengthened the book's applicability in practice. Nonetheless, the clarity and pedagogical orientation of the frameworks make them well-suited for integration into teacher training modules focused on visual literacy and multimodal analysis, as well as direct application in classroom activities such as infographic analysis, visual scaffolding and student-generated infographics. This book serves as a valuable resource for a wide audience, including educators, curriculum designers and linguists who are interested in the intersections of language and imagery in science pedagogy.

ACKNOWLEDGMENTS

This work was facilitated by academic support (insights into key termonologies mentioned in the book under review and the needed components of a high-quality book review) from Professor Len Unsworth, Research Director of Educational Semiotics in English and Literacy Pedagogy at the Australian Catholic University, and Dr. Thu Ngo, Senior Lecturer at the University of New South Wales Sydney. This work did not receive any financial funding or grants.

DECLARATION OF COMPETING INTEREST

None declared.

AUTHORS' CONTRIBUTION

Chau Hoang Vo: Conceptualization; Data curation; Formal analysis; Funding acquisition; Methodology; Project administration; Visualization; Writing – original draft; Writing – review & editing.

Tin Thanh Nguyen: Formal analysis; Investigation; Methodology; Resources; Software; Supervision; Writing – original draft.

REFERENCES

Kress, G., & van Leeuwen, T. (2021). *Reading images: The grammar of visual design* (3rd ed.). Routledge. https://doi.org/10.4324/9781003099857

Martin, J. R. (2020). Revisiting field: Specialized knowledge in secondary school science and humanities discourse. In J. R. Martin, K. Maton, & Y. J. Doran (Eds.), *Accessing academic discourse: Systemic functional linguistics and legitimation code theory* (pp. 114-148). Routledge.

Martin, J. R., & Unsworth, L. (2024). *Reading images for knowledge building: Analyzing infographics in school science*. Routledge.

Tang, K.-S. (2020). *Discourse strategies for science teaching and learning: Research and practice*. Routledge.

Unsworth, L., Tytler, R., Fenwick, K., Humphrey, S., Chandler, P., Herrington, M., & Pham, L. (2022). *Multimodal literacy in school science: Transdisciplinary perspectives on theory, research and pedagogy*. Routledge. http://doi.org/10.4324/9781003150718

CONTENT

EDITORIAL

RESEARCH PAPERS

REVIEW PAPERS

BOOK REVIEWS