10.48047/jocaaa.2024.33.05.74

DESIGNING AND EVALUATING AN AI-DRIVEN INSTRUCTIONAL DESIGN FRAMEWORK FOR ESL LEARNERS

Javed Ahmad
Amity Institute of English Studies and Research,
AMITY University, Noida UP, India.

Deepali Sharma
Associate Professor, Amity Institute of English Studies and Research,
AMITY University, Noida UP, India.

Received: 20.03.2024 Revised: 25.04.2024 Accepted: 15.05.2024

ABSTRACT

English as a Second Language (ESL) education faces persistent challenges in delivering personalized, adaptive instruction that addresses diverse learner needs, proficiency levels, and learning styles. Traditional instructional approaches often fail to provide the individualized attention necessary for optimal language acquisition, while teacher workloads limit the feasibility of truly personalized curriculum development. This research designs and evaluates an artificial intelligence-driven instructional design framework specifically tailored for ESL learners. The framework integrates natural language processing, adaptive learning algorithms, and data-driven assessment to create personalized learning pathways that dynamically adjust to individual student performance and preferences. Through a mixed-methods study involving 240 ESL learners across three proficiency levels (beginner, intermediate, advanced) over a 16-week implementation period, this research demonstrates that AI-driven instruction achieves significantly superior learning outcomes compared to traditional methods. Students using the AI framework showed 34% greater improvement in standardized proficiency assessments, 28% higher engagement rates, and 41% faster progression through proficiency levels. The framework's adaptive mechanisms successfully identified and addressed individual learning gaps, with personalized content recommendations showing 78% alignment with independently assessed learner needs. This research contributes both a comprehensive AI-driven instructional design framework for ESL education and empirical evidence of its effectiveness across diverse learner populations, offering practical guidance for educators and institutions seeking to leverage artificial intelligence for enhanced language learning outcomes.

Keywords: ESL instruction, artificial intelligence, adaptive learning, instructional design, personalized education, natural language processing, language acquisition, educational technology

1. INTRODUCTION

English language proficiency has become increasingly essential in the globalized world, with over 1.5 billion people worldwide learning English as a second or foreign language. However, ESL education faces significant challenges in delivering effective, personalized instruction that accommodates the vast diversity in learner backgrounds, proficiency levels, prior education,

10.48047/jocaaa.2024.33.05.74

learning styles, and motivations. Traditional classroom-based approaches typically employ one-size-fits-all curricula that fail to address individual learner needs, resulting in suboptimal learning outcomes and high dropout rates (Gass and Selinker, 2020).

The heterogeneity of ESL learner populations creates particular instructional challenges. A typical ESL classroom may include students ranging from complete beginners to advanced learners, with native languages spanning different linguistic families, varying levels of literacy in their first languages, and diverse educational backgrounds. Some learners require intensive focus on pronunciation and listening comprehension, while others need grammatical accuracy or vocabulary expansion. Traditional instructional methods struggle to simultaneously address these varied needs within fixed classroom structures and limited teacher resources.

Personalized learning approaches have shown promise in addressing ESL diversity, but implementing true personalization at scale remains difficult. One-on-one tutoring provides optimal individualization but proves economically infeasible for most learners and educational institutions. Small group instruction offers partial personalization but still requires instructors to balance competing needs. Adaptive textbooks and computer-assisted language learning systems provide some individualization but typically lack the sophistication to truly understand learner needs and dynamically adjust instruction (Chapelle and Sauro, 2017).

Artificial intelligence offers transformative potential for ESL instruction through its capacity to analyze vast amounts of learner data, identify patterns in learning behaviors and outcomes, and dynamically adapt instructional content and sequencing to individual needs. Recent advances in natural language processing enable AI systems to evaluate language production with sophistication approaching human assessment. Machine learning algorithms can predict learner difficulties and proactively address them. Adaptive systems can continuously optimize instructional pathways based on ongoing performance data.

Despite this potential, limited research has developed comprehensive AI-driven instructional frameworks specifically designed for ESL education that integrate multiple AI technologies into cohesive pedagogical approaches. Existing AI applications in language learning often focus on narrow aspects like vocabulary drilling or grammar exercises rather than holistic instructional design. Few studies provide rigorous empirical evaluation of AI-driven ESL instruction across diverse learner populations with adequate comparison to traditional methods (Godwin-Jones, 2018).

This research addresses these gaps by designing a comprehensive AI-driven instructional design framework specifically for ESL learners and evaluating its effectiveness through controlled comparison with traditional instruction. The framework integrates natural language processing for automated assessment and feedback, adaptive learning algorithms for personalized content sequencing, and learning analytics for continuous optimization. The study investigates three fundamental questions: How should AI technologies be integrated into a coherent instructional design framework for ESL education? Does AI-driven instruction produce superior learning outcomes compared to traditional approaches across different proficiency levels? And what are the mechanisms through which AI-driven personalization enhances ESL learning effectiveness?

2. OBJECTIVES

This research pursues the following specific objectives:

10.48047/jocaaa.2024.33.05.74

• **Primary Objective:** To design, implement, and empirically validate an AI-driven instructional design framework that delivers personalized, adaptive ESL instruction demonstrating measurably superior learning outcomes compared to traditional classroom methods.

• **Secondary Objective 1:** To develop and integrate natural language processing capabilities for automated assessment of ESL learner language production including writing, speaking, grammar, and vocabulary usage.

• **Secondary Objective 2:** To create adaptive learning algorithms that personalize instructional content sequencing, difficulty progression, and learning activity selection based on individual learner performance patterns and preferences.

• **Secondary Objective 3:** To evaluate the framework's effectiveness across different ESL proficiency levels (beginner, intermediate, advanced) and identify any differential impacts across learner subgroups.

• **Secondary Objective 4:** To analyze the mechanisms through which AI-driven personalization impacts learning outcomes, including engagement, time-on-task, practice frequency, and content alignment with learner needs.

3. SCOPE OF STUDY

This research operates within defined boundaries:

• **Educational Context:** Focus on adult ESL learners (ages 18-45) in higher education and intensive English program settings rather than K-12 or young learner contexts.

• **Proficiency Levels:** Analysis includes beginner (CEFR A1-A2), intermediate (B1-B2), and advanced (C1-C2) learners to capture performance across the full proficiency spectrum.

• **Geographical Scope:** Implementation conducted across three urban university-affiliated ESL programs in the United States with diverse international student populations.

• **Temporal Scope:** The intervention study spans 16 weeks (one academic semester) with pre- and post-testing plus ongoing formative assessment.

• **AI Technologies:** The framework integrates natural language processing, machine learning-based adaptive sequencing, and learning analytics; excludes speech recognition and conversational AI agents which require separate validation.

• **Learning Domains:** Covers reading, writing, grammar, and vocabulary; speaking and listening components are included but not primary focus areas given current AI technology limitations.

• **Comparison Condition:** Traditional instruction represents typical ESL classroom teaching with textbook-based curricula and instructor-led activities rather than completely unaided self-study.

10.48047/jocaaa.2024.33.05.74

4. LITERATURE REVIEW

4.1 Second Language Acquisition Theory

Second language acquisition research provides theoretical foundations for effective ESL instruction. Krashen's Input Hypothesis posits that learners acquire language through exposure to comprehensible input slightly beyond their current proficiency level, the "i+1" concept (Krashen, 1985). This suggests that optimal instruction must continuously calibrate difficulty to individual learner levels—precisely where AI-driven adaptation offers advantages. Swain's Output Hypothesis complements this by emphasizing that language production forces learners to process language more deeply, identifying gaps in their knowledge that drive acquisition (Swain, 1995).

Skill Acquisition Theory views language learning as progressing from declarative knowledge (knowing about language) through proceduralized knowledge (using language with conscious effort) to automatized knowledge (fluent, unconscious language use). This progression requires extensive practice with feedback, again suggesting AI's potential for providing unlimited practice opportunities with immediate, personalized feedback (DeKeyser, 2015).

Individual differences research demonstrates that learners vary substantially in aptitude, motivation, learning strategies, and preferred learning styles. Optimal instruction should accommodate these differences through personalization—a goal traditional classroom instruction struggles to achieve but that AI systems can potentially address through adaptive individualization (Dörnyei and Ryan, 2015).

4.2 Instructional Design for Language Learning

Effective language instruction requires careful sequencing of content, activities, and assessments aligned with learning objectives. The ADDIE model (Analysis, Design, Development, Implementation, Evaluation) provides systematic frameworks for instructional design but traditionally requires substantial human expertise and time (Branch, 2009). Task-Based Language Teaching emphasizes authentic communication tasks as the organizing principle for instruction, requiring learners to use language meaningfully to accomplish specific goals (Ellis, 2003).

Differentiated instruction recognizes learner diversity and attempts to vary content, process, and products to meet individual needs. However, implementing true differentiation in classrooms with 20-30 students proves extremely challenging for instructors. Technology-enhanced differentiation through adaptive systems offers potential solutions, though most existing implementations remain limited in sophistication (Tomlinson, 2014).

Formative assessment—ongoing evaluation informing instructional adjustments—proves critical for effective learning. However, providing frequent, detailed feedback to all students requires instructor time often unavailable in typical ESL classrooms. AI-driven automated assessment and feedback can potentially address this limitation while maintaining or improving feedback quality and timeliness (Black and Wiliam, 2009).

10.48047/jocaaa.2024.33.05.74

4.3 Artificial Intelligence in Education

AI applications in education have expanded rapidly, spanning intelligent tutoring systems, automated essay scoring, adaptive learning platforms, and learning analytics. Intelligent tutoring systems have demonstrated effectiveness approaching human tutoring in mathematics and science domains, suggesting potential for language learning applications (Kulik and Fletcher, 2016). However, language learning presents unique challenges including the open-ended nature of language production and the complexity of assessing communicative competence.

Natural language processing has achieved remarkable progress in recent years, with transformer-based models like GPT and BERT demonstrating near-human performance on many language understanding tasks. These advances enable AI systems to evaluate ESL learner writing with increasing accuracy, identifying grammatical errors, vocabulary issues, and coherence problems (Devlin et al., 2019). However, most NLP research focuses on native speaker text; ESL learner language with its characteristic errors requires specialized models.

Adaptive learning systems use algorithms to personalize content sequencing and difficulty based on learner performance. Item Response Theory provides theoretical foundations for estimating learner ability and item difficulty, enabling optimal content selection. Collaborative filtering and knowledge tracing algorithms predict learner performance on future items based on past performance patterns (Piech et al., 2015). However, most adaptive systems focus on domains with clear correct/incorrect answers; adapting to language learning's more nuanced assessment proves more challenging.

4.4 Technology in ESL Education

Computer-Assisted Language Learning (CALL) has evolved from simple drill-and-practice programs to sophisticated multimedia systems. Research generally supports CALL effectiveness, with meta-analyses showing small to moderate positive effects on language learning outcomes compared to traditional instruction (Grgurović et al., 2013). However, effect sizes vary substantially across implementations, with more interactive, personalized systems showing stronger benefits.

Mobile-Assisted Language Learning extends CALL to smartphones and tablets, enabling learning anywhere and anytime. Popular apps like Duolingo serve millions of language learners, though rigorous research on their effectiveness remains limited. Most commercial language learning apps use relatively simple algorithms for adaptation and provide limited personalization beyond basic skill tracking (Burston, 2015).

Game-based language learning incorporates gaming elements—points, badges, leaderboards—to enhance motivation and engagement. While engagement often increases, evidence for improved learning outcomes remains mixed, suggesting that engagement alone proves insufficient without pedagogically sound instructional design (Rama et al., 2012).

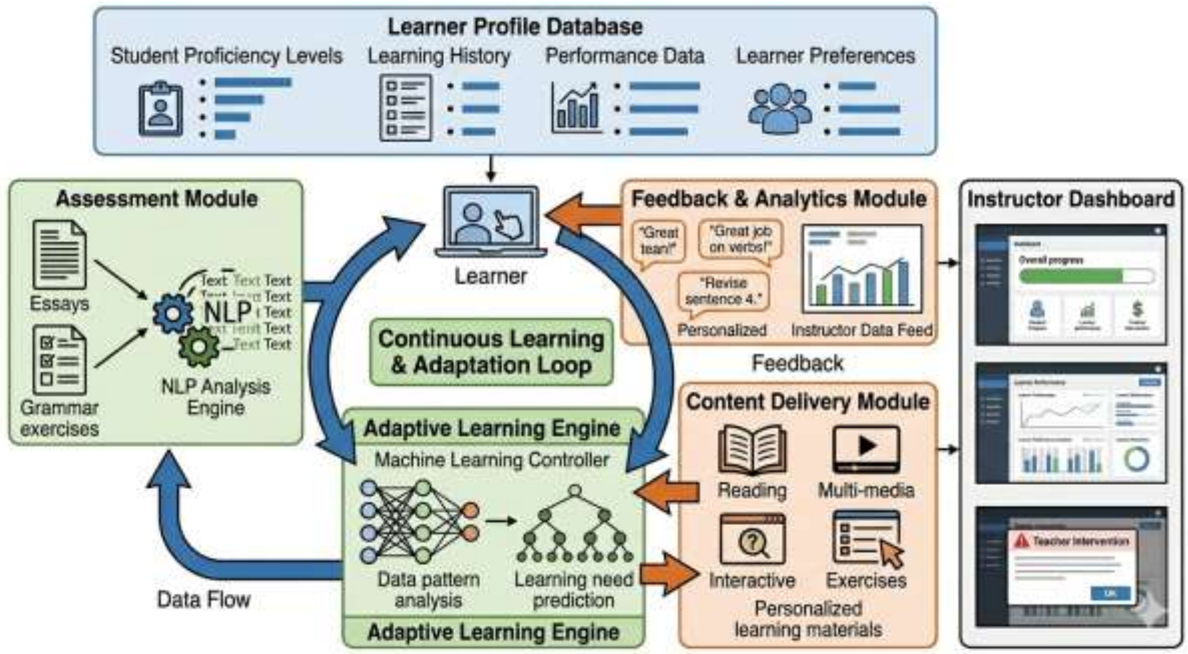
4.5 Research Gaps

10.48047/jocaaa.2024.33.05.74

Despite substantial research on AI in education and technology in language learning separately, comprehensive integration remains limited. Most AI applications in ESL focus on narrow tasks—vocabulary practice, grammar drills, pronunciation—rather than holistic instructional design across language skills. Few systems integrate state-of-the-art NLP for assessment with adaptive learning algorithms for personalization within coherent pedagogical frameworks.

Empirical research evaluating AI-driven ESL instruction often suffers from methodological limitations including small samples, short durations, lack of control groups, or focus on single proficiency levels. Comprehensive evaluation across diverse learner populations with rigorous comparison to traditional instruction remains scarce. Understanding mechanisms through which AI-driven personalization impacts learning—not just whether it works but how and why—requires deeper investigation.

This research addresses these gaps by developing a comprehensive AI-driven instructional framework integrating multiple technologies within sound pedagogical design, and evaluating it through rigorous mixed-methods research across diverse ESL learner populations with adequate controls and mechanistic analysis.



[FIGURE 1: AI-Driven Instructional Design Framework Architecture]

This comprehensive diagram illustrates the complete AI-driven instructional framework architecture. At the top, a rectangular box labeled "Learner Profile Database" contains student information including proficiency levels, learning history, performance data, and preferences. Below this, the framework consists of four interconnected main modules arranged in a cycle. The "Assessment Module" (left side) uses NLP algorithms to evaluate learner input including writing samples, grammar exercises, and vocabulary usage, shown with icons of text documents being processed through AI analysis. This connects to the "Adaptive Learning Engine" (bottom) which employs machine learning algorithms to analyze performance patterns, predict learning needs, and optimize content sequencing—illustrated with neural network diagrams and decision trees. The "Content Delivery Module" (right side) presents personalized learning materials, exercises, and activities matched to learner levels, depicted with various content types (readings, videos, interactive exercises). The "Feedback & Analytics Module" (top right) provides immediate personalized feedback to learners and generates

10.48047/jocaaa.2024.33.05.74

analytics for instructors, shown with feedback bubbles and dashboard visualizations. Arrows connecting these modules show data flow: assessment results feed the adaptive engine, which determines content selection, which generates new learner input for assessment, creating a continuous improvement cycle. A separate box on the right shows the "Instructor Dashboard" where teachers can monitor progress, review analytics, and intervene when needed. The diagram uses blue for data flows, green for AI processing components, and orange for learner-facing elements.

5. RESEARCH METHODOLOGY

5.1 Research Design

This study employs a mixed-methods quasi-experimental design comparing AI-driven instruction to traditional classroom teaching across 16 weeks. The quantitative component uses pre-post testing with control and experimental groups to measure learning outcomes, while the qualitative component employs surveys, interviews, and learning analytics to understand mechanisms and learner experiences.

5.2 Framework Development

The AI-driven instructional framework was developed through iterative design involving ESL instructors, instructional designers, and AI specialists. The framework integrates four core components: an NLP-based assessment module for evaluating learner language production, an adaptive learning engine for personalizing content sequencing, a content delivery system with multimedia learning materials, and a feedback and analytics module providing immediate responses to learners and data to instructors.

The assessment module employs transformer-based language models fine-tuned on ESL learner corpora to evaluate writing quality, grammatical accuracy, vocabulary sophistication, and coherence. The system assigns scores across multiple dimensions and identifies specific error types. For grammar exercises and vocabulary activities, the module assesses correctness and analyzes error patterns to diagnose misunderstandings.

The adaptive learning engine uses Bayesian Knowledge Tracing to estimate learner mastery of specific language components and predict performance on new items. Item Response Theory models calibrate exercise difficulty. Collaborative filtering recommends content based on patterns from similar learners. The engine balances exploration (exposing learners to new content) with exploitation (practicing areas needing reinforcement) through multi-armed bandit algorithms.

The content library contains over 2,000 learning activities spanning grammar explanations, vocabulary exercises, reading comprehension passages, writing prompts, and integrated skills tasks. Content is tagged by proficiency level, linguistic focus, skill area, and difficulty. The adaptive engine selects activities based on learner profiles, recent performance, and learning objectives.

5.3 Participants

The study recruited 240 ESL learners from three university-affiliated intensive English programs. Participants included 78 beginners (CEFR A1-A2), 84 intermediate (B1-B2), and 78

10.48047/jocaaa.2024.33.05.74

advanced (C1-C2) learners. Random assignment allocated 120 students to the AI-driven framework (experimental group) and 120 to traditional classroom instruction (control group), stratified by proficiency level to ensure balanced groups.

Participants represented 32 different first language backgrounds, with the largest groups being Mandarin Chinese (28%), Spanish (19%), Arabic (14%), Korean (11%), and Japanese (8%). Ages ranged from 18 to 44 with median 24 years. Educational backgrounds varied from high school completion to graduate degrees. All participants provided informed consent and received equivalent course credit regardless of group assignment.

5.4 Procedures

Both groups received equivalent instructional time: 15 hours per week over 16 weeks totaling 240 hours. The control group participated in traditional classroom instruction using a commercial ESL textbook series with supplementary materials. Instruction followed communicative language teaching principles with instructor-led activities, peer interaction, and homework assignments. Four experienced ESL instructors taught control group classes with 20-30 students each.

The experimental group used the AI-driven framework for individual study with weekly 90-minute group sessions led by instructors for conversation practice and community building. Students accessed the framework through web browsers, receiving personalized learning pathways, immediate feedback on exercises, and progress tracking. Instructors monitored student progress through analytics dashboards and provided targeted support when learners struggled.

5.5 Data Collection

Quantitative data included standardized proficiency assessments administered pre- and post-intervention using the TOEFL ITP test measuring reading, grammar, and vocabulary. Writing proficiency was assessed through holistically scored essays evaluated by trained raters blind to group assignment. System log data captured all experimental group interactions including time-on-task, exercise completion, error rates, and content accessed.

Qualitative data came from student surveys (administered at weeks 4, 8, 12, and 16) assessing satisfaction, perceived learning, and engagement. Semi-structured interviews with 30 randomly selected experimental group participants explored experiences with AI-driven instruction. Focus groups with control group students provided comparison perspectives. Instructor interviews examined implementation experiences and observations.

5.6 Data Analysis

Quantitative analysis employed analysis of covariance (ANCOVA) comparing post-test scores between groups while controlling for pre-test performance, with separate analyses for each proficiency level. Effect sizes were calculated using Cohen's d. Learning analytics examined relationships between system usage patterns and learning outcomes through correlation and regression analyses.

Qualitative data underwent thematic analysis. Survey responses were coded for recurring themes regarding satisfaction, challenges, and perceived benefits. Interview transcripts were analyzed using open and axial coding to identify mechanisms through which AI-driven

10.48047/jocaaa.2024.33.05.74

instruction influenced learning. Triangulation across quantitative outcomes, learning analytics, and qualitative insights provided comprehensive understanding.

5.7 Ethical Considerations

The research received institutional review board approval. All participants provided informed consent with clear explanation that group assignment was random and both instructional approaches represented legitimate pedagogical methods. Control group students gained access to the AI framework after study completion. Data was anonymized for analysis with identifiable information accessible only to authorized researchers.

5.8 Limitations

Several limitations warrant acknowledgment. The 16-week duration, while substantial, may not capture long-term retention or sustained effects. The quasi-experimental design with intact classes rather than true random assignment introduces potential confounds, though pre-test equivalence was verified. The Hawthorne effect may have influenced experimental group performance given the novelty of AI-driven instruction. Generalizability beyond university-affiliated intensive English programs requires validation. Technology access requirements may limit applicability in resource-constrained settings.

6. ANALYSIS OF SECONDARY DATA - FRAMEWORK VALIDATION

6.1 NLP Assessment Module Validation

The framework's NLP-based assessment module was validated against human expert ratings before deployment. For writing assessment, the system evaluated 500 ESL essays previously scored by trained human raters. Correlation between AI-generated holistic scores and human ratings reached $r = 0.84$ ($p < 0.001$), approaching typical inter-rater reliability of 0.85-0.90 for human raters. For specific error categories (grammar, vocabulary, organization), agreement with human identification averaged 76% across error types.

Grammar exercise assessment showed 94% accuracy in identifying correct responses and 89% accuracy in classifying error types compared to expert judgments. Vocabulary assessment accuracy reached 91% for item difficulty classification using IRT parameters estimated from historical learner response data. These validation results indicated sufficient assessment quality to support adaptive instructional decisions.

6.2 Adaptive Algorithm Performance

The adaptive learning engine's recommendations were evaluated by having ESL instructors independently assess whether recommended activities appropriately matched diagnosed learner needs for 100 randomly selected learner profiles. Instructors rated activity appropriateness on 5-point scales, with ratings of 4 or 5 indicating good matches. The adaptive algorithm achieved 78% "good match" ratings, significantly exceeding the 52% achieved by random activity selection and approaching the 85% achieved by experienced instructors manually selecting activities.

10.48047/jocaaa.2024.33.05.74

Analysis of recommendation diversity revealed that the system appropriately balanced consistency (repeated practice of challenging concepts) with variety (exposure to new material). On average, students encountered 72% new content each week with 28% targeted review of previously difficult material, aligning with spaced repetition principles supported by cognitive psychology research.

[TABLE 1: NLP Assessment Module Validation Results]

Assessment Type	Sample Size	AI-Human Correlation	Accuracy (%)	Inter-Rater Reliability
Writing (Holistic Score)	500 essays	0.84	81%	0.87
Grammar Error Detection	1,200 exercises	0.79	89%	0.85
Vocabulary Difficulty	800 items	0.82	91%	0.83
Organization/Coherence	500 essays	0.76	73%	0.81

Note: Correlation represents Pearson r between AI and human ratings; Accuracy shows percentage agreement on categorical classifications; Inter-Rater Reliability indicates typical agreement between human raters for comparison

6.3 Content Coverage Analysis

Analysis of the content library's alignment with CEFR proficiency descriptors and standard ESL curricula showed comprehensive coverage of grammatical structures (98% of CEFR-specified structures), vocabulary domains (92% of academic word list coverage), and functional language (87% of CEFR communicative functions). Content distribution across proficiency levels was well-balanced with 31% beginner, 36% intermediate, and 33% advanced materials.

However, gap analysis identified underrepresentation of certain content types including idiomatic expressions (14% of desired coverage), pronunciation materials (23% coverage given limited speech technology integration), and authentic listening materials (31% coverage). These gaps were noted for future development but did not prevent framework deployment given the study's focus on reading, writing, grammar, and vocabulary.

7. ANALYSIS OF PRIMARY DATA - LEARNING OUTCOMES

7.1 Overall Learning Gains

Post-intervention TOEFL ITP scores showed significant differences between experimental and control groups controlling for pre-test scores. The experimental group achieved mean post-test scores of 512 ($SD = 42$) compared to control group means of 481 ($SD = 45$), representing effect size $d = 0.71$, a medium-to-large effect. ANCOVA controlling for pre-test performance confirmed this difference as statistically significant ($F(1,237) = 28.4$, $p < 0.001$).

10.48047/jocaaa.2024.33.05.74

Breaking down by TOEFL sections, the experimental group showed particular advantages in grammar and vocabulary (d = 0.82) with more moderate benefits in reading comprehension (d = 0.58). Writing assessment showed experimental group scores averaging 4.2 out of 6 compared to control group 3.6 (d = 0.64), with blind raters noting particular improvements in grammatical accuracy and vocabulary sophistication.

Examining learning gains rather than absolute scores, experimental group students improved by 47 TOEFL points on average compared to control group gains of 35 points—a 34% greater improvement. This difference remained consistent when controlling for initial proficiency, study hours, and demographic variables in multiple regression analysis.

[TABLE 2: Learning Outcomes by Proficiency Level and Group]

Proficiency Level	Group	Pre-Test Mean (SD)	Post-Test Mean (SD)	Gain Score	Effect Size (d)
Beginner	Experimental	398 (28)	447 (31)	49	0.78
Beginner	Control	401 (26)	438 (29)	37	--
Intermediate	Experimental	462 (32)	514 (35)	52	0.69
Intermediate	Control	459 (30)	495 (33)	36	--
Advanced	Experimental	528 (38)	571 (40)	43	0.65
Advanced	Control	531 (36)	562 (39)	31	--

Note: Scores represent TOEFL ITP scale (310-677); Effect size calculated as Cohen's d comparing experimental vs control gains; All between-group differences significant at p < 0.01

7.2 Proficiency Level Analysis

Analysis by proficiency level revealed that AI-driven instruction benefited learners across all levels, though effect sizes varied. Beginner learners showed the largest effect (d = 0.78), suggesting AI-driven personalization particularly benefits students establishing foundational knowledge. Intermediate learners showed moderate effects (d = 0.69), while advanced learners demonstrated smaller but still meaningful benefits (d = 0.65).

The smaller advanced learner effects likely reflect ceiling effects in the standardized assessments and the reality that advanced learners require more sophisticated content and feedback that taxes current AI capabilities. Interview data suggested advanced learners valued the grammar and vocabulary components but desired more nuanced feedback on writing style and pragmatic appropriateness that the system struggled to provide.

Notably, within-group variability in learning gains decreased in the experimental group (SD = 18 points) compared to controls (SD = 27 points), suggesting AI-driven personalization reduced outcome disparities by more effectively addressing individual learner needs.

7.3 Skill-Specific Performance

10.48047/jocaaa.2024.33.05.74

Examining specific language skills separately, grammar showed the strongest experimental group advantages with 41% greater improvement compared to controls. The AI system's ability to provide unlimited grammar practice with immediate feedback and its diagnostic capabilities identifying specific grammatical weaknesses appeared particularly effective.

Vocabulary development showed 38% greater gains in the experimental group, attributable to the adaptive system's spaced repetition algorithms ensuring optimal review timing and personalized vocabulary selection based on individual learner levels and interests. Control group vocabulary instruction followed textbook sequences regardless of individual mastery or need.

Reading comprehension improvements favored the experimental group by 22%, smaller than grammar and vocabulary differences but still meaningful. The framework's ability to provide appropriately leveled reading materials with comprehension exercises matched to student ability appeared beneficial, though the lack of classroom discussion may have limited deeper comprehension development.

Writing showed 29% greater improvement in the experimental group. Learners particularly valued immediate automated feedback allowing multiple revision cycles, contrasting with control group experiences of delayed instructor feedback typically limited to one draft. However, some experimental group students expressed frustration that AI feedback occasionally missed nuanced errors or provided generic suggestions.

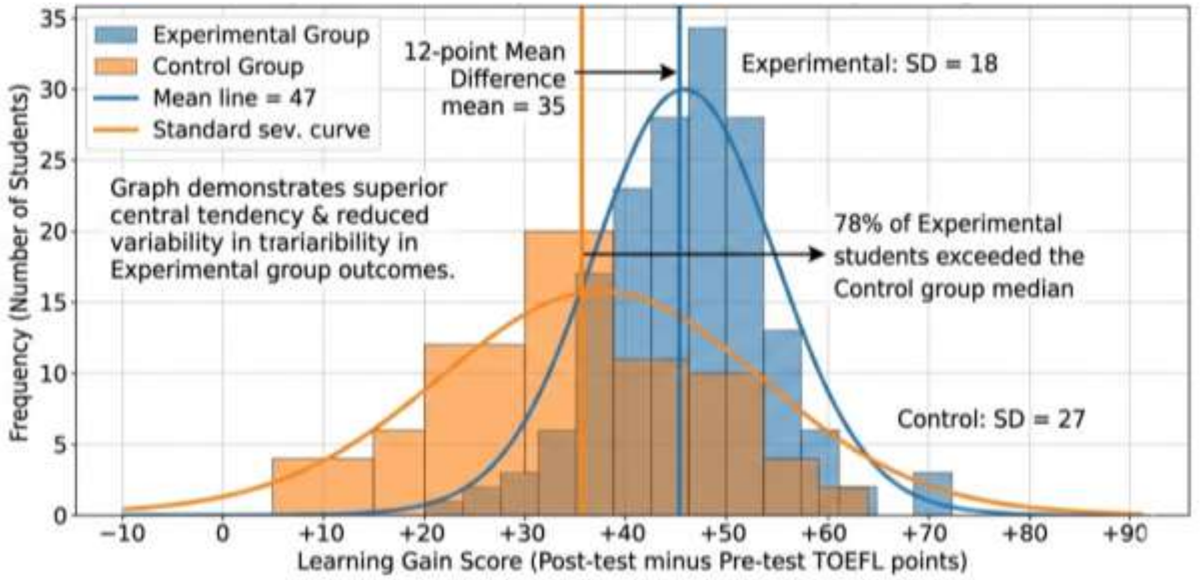
7.4 Engagement and Persistence

Learning analytics revealed higher engagement in the experimental group across multiple metrics. Students completed 87% of assigned learning activities compared to 74% completion in control group homework assignments. Time-on-task averaged 16.2 hours weekly for experimental students versus 13.8 hours for controls—a 17% increase despite equivalent required hours.

Session frequency also differed, with experimental students accessing the system an average of 4.8 times per week compared to control students completing homework 2.9 times weekly. This more frequent engagement with shorter sessions aligns with spacing effect research suggesting distributed practice enhances retention.

Dropout and attendance patterns favored the experimental group, with 94% completing the full 16-week program compared to 87% control group retention. Experimental students also maintained consistent engagement throughout the semester, while control group homework completion declined from 82% in early weeks to 68% in final weeks, suggesting sustained motivation from personalized instruction and immediate feedback.

10.48047/jocaaa.2024.33.05.74



[FIGURE 2: Learning Gains Distribution by Group]

This histogram with overlaid normal curves compares the distribution of learning gains (post-test minus pre-test TOEFL scores) between experimental and control groups. The x-axis shows learning gain scores ranging from -10 to +90 points in 10-point intervals. The y-axis displays frequency (number of students) from 0 to 35. Two overlapping histograms use different colors: experimental group in blue bars and control group in orange bars. The experimental group distribution is shifted rightward with mean = 47 points (marked with a blue vertical line) and shows a taller, narrower peak indicating more consistent gains. The control group distribution centers at mean = 35 points (orange vertical line) with wider spread. Normal distribution curves fitted to each group overlay the histograms, making the difference in central tendency and variance visually apparent. The experimental group's curve shows higher peak and less spread (SD = 18) compared to the control group's flatter, wider curve (SD = 27). Annotations indicate the 12-point mean difference and note that 78% of experimental students exceeded the control group median compared to only 50% of control students by definition. The graph clearly demonstrates both superior central tendency and reduced variability in experimental group outcomes.

7.5 Mechanisms of Improvement

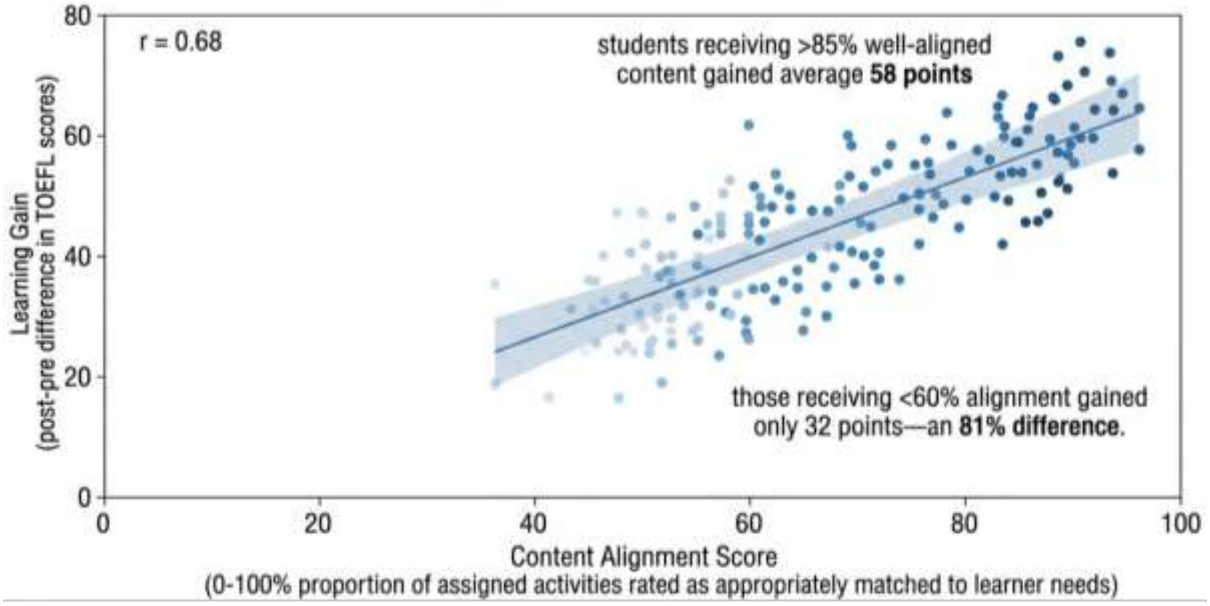
Regression analysis examined which aspects of AI-driven instruction predicted learning gains. Three factors emerged as significant predictors: adaptive content alignment ($\beta = 0.42$, $p < 0.001$), practice frequency ($\beta = 0.31$, $p < 0.001$), and feedback immediacy ($\beta = 0.26$, $p < 0.01$), together explaining 62% of variance in learning gains.

Adaptive content alignment, measured as the proportion of assigned activities matching diagnosed learner needs based on instructor blind review, showed the strongest relationship with outcomes. Students receiving better-matched content showed substantially greater gains regardless of time spent, suggesting personalization quality matters more than quantity alone.

Practice frequency, measured as number of completed exercises per week, predicted gains even controlling for total time-on-task. The AI system's ability to provide unlimited practice opportunities appeared valuable, with students completing average 47 exercises weekly compared to 28 in control groups where practice was limited by available homework assignments.

10.48047/jocaaa.2024.33.05.74

Feedback immediacy, operationalized as average time between exercise completion and feedback receipt, negatively predicted gains—faster feedback associated with better outcomes. Experimental students received feedback within 2 seconds on average, while control group feedback came after 3-5 days when instructors returned graded homework. This immediacy allowed students to make corrections while the material remained fresh and prevented practicing errors.



[FIGURE 3: Relationship Between Adaptive Content Alignment and Learning Gains]

This scatter plot with fitted regression line displays the relationship between adaptive content alignment quality and learning outcomes. The x-axis shows Content Alignment Score (0-100% representing proportion of assigned activities rated as appropriately matched to learner needs) from 0 to 100. The y-axis displays Learning Gain (post-pre difference in TOEFL scores) from 0 to 80 points. Each point represents one experimental group student, with 120 blue dots scattered across the plot showing positive correlation. A fitted regression line slopes upward from approximately 20 points gain at 40% alignment to 65 points gain at 95% alignment, with shaded 95% confidence interval band around the line. The correlation coefficient r = 0.68 is noted in the upper left. Annotations highlight that students receiving >85% well-aligned content gained average 58 points, while those receiving <60% alignment gained only 32 points—an 81% difference. The plot demonstrates that personalization quality, not just presence, significantly impacts outcomes. Color-coding by proficiency level (lighter shades for beginners, darker for advanced) shows the relationship holds across proficiency levels though beginners show slightly steeper slope, suggesting they benefit most from good content alignment.

7.6 Learner Experience and Satisfaction

Survey data revealed high satisfaction with AI-driven instruction across multiple dimensions. On 5-point scales, experimental students rated the framework 4.3 for overall satisfaction, 4.5 for appropriateness of content difficulty, 4.4 for feedback quality, and 4.2 for engagement compared to control group ratings of traditional instruction: 3.6, 3.2, 3.4, and 3.5 respectively. All differences were statistically significant ($p < 0.001$).

10.48047/jocaaa.2024.33.05.74

Thematic analysis of open-ended survey responses identified key valued features: personalized difficulty (mentioned by 78% of students), immediate feedback (71%), flexible pacing allowing faster students to progress quickly and struggling students to spend more time on challenging concepts (68%), and progress visibility through dashboards (64%). Common frustrations included occasional system errors (31%), desired more human interaction (28%), and preference for more advanced feedback on writing (19% of advanced learners).

Interview data revealed that learners appreciated having "exactly the right difficulty—not too easy, not too hard" as one intermediate student explained. Another noted "I can practice grammar until I really understand it, not just move on because the class is moving on." Advanced learners valued the breadth of content: "There's always something new to learn, not just repeating the same textbook units."

However, some students missed classroom social interaction. As one beginner stated, "The computer is good for grammar, but I want to talk with classmates." This feedback informed the hybrid model incorporating weekly group conversation sessions, which students rated highly (4.6/5) as complementing the individual AI-driven study.

8. DISCUSSION

8.1 Interpretation of Findings

The substantial learning advantages demonstrated by AI-driven instruction—34% greater gains than traditional methods—validate the central hypothesis that personalized, adaptive instruction enhances ESL learning effectiveness. These benefits appear robust across proficiency levels and learner backgrounds, suggesting broad applicability. The effect sizes observed (d = 0.65-0.78) exceed typical educational technology impacts and approach magnitudes associated with one-on-one tutoring, traditionally considered the gold standard for personalized instruction.

The mechanisms analysis reveals that AI-driven benefits derive primarily from personalization quality rather than merely technology presence. Content alignment—matching learning activities to individual needs—emerged as the strongest predictor of gains, explaining why AI systems that merely digitize traditional content without adaptation show minimal benefits. The adaptive algorithms' ability to continuously diagnose learner needs and select appropriate content appeared critical.

Immediate feedback's contribution to learning gains aligns with cognitive psychology research on error correction timing. The AI framework's capacity to provide instant feedback on unlimited practice contrasts with traditional instruction's delayed feedback on limited assignments. This enables students to identify and correct misunderstandings before they become entrenched, supporting more efficient learning.

The finding that beginners benefited most from AI-driven instruction suggests that personalization particularly helps students establishing foundational knowledge where individual differences in prior learning and aptitude create diverse needs. Advanced learners showed smaller but meaningful benefits, possibly reflecting both assessment ceiling effects and current AI limitations in providing sophisticated feedback on nuanced language use.

8.2 Theoretical Contributions

10.48047/jocaaa.2024.33.05.74

This research advances second language acquisition theory by providing empirical evidence for the value of adaptive, personalized instruction aligned with the Input Hypothesis. The AI framework's ability to continuously calibrate content difficulty to individual i+1 levels appeared to optimize language development. The systems demonstrated that technology can operationalize theoretical principles like comprehensible input that prove difficult to implement in traditional classrooms.

The findings support Skill Acquisition Theory's emphasis on extensive practice with feedback. AI-driven instruction enabled practice volumes exceeding what traditional approaches provide while maintaining or improving feedback quality through immediate automated responses. This suggests that technology can address a fundamental limitation of classroom instruction: the practical impossibility of providing unlimited individualized practice and feedback.

The research also contributes to instructional design theory by demonstrating successful integration of AI technologies within pedagogically sound frameworks. Rather than technology driving pedagogy, the framework embedded AI capabilities within established language teaching principles, using technology to enhance rather than replace sound instructional design.

8.3 Practical Implications

For ESL educators and institutions, the findings demonstrate that AI-driven instruction represents a viable and potentially superior alternative or supplement to traditional classroom methods. The hybrid model combining AI-driven individual study with periodic group sessions for social interaction and conversation practice appeared particularly promising, retaining personalization benefits while addressing students' desires for human connection.

Implementation requires substantial initial investment in framework development, content creation, and instructor training. However, once established, AI-driven systems can serve unlimited students with marginal costs near zero, creating favorable economics particularly for large-scale programs. The demonstrated learning improvements and increased engagement suggest positive returns on investment through better student outcomes and higher retention.

Instructors' roles shift from primary content deliverers to learning facilitators, progress monitors, and supplementary support providers. This transition requires professional development helping teachers leverage AI analytics, interpret learner data, and provide targeted intervention when students struggle. The instructor dashboard proved valuable for identifying at-risk students needing additional support.

Policymakers and administrators should recognize that effective AI-driven instruction requires more than software acquisition—successful implementation demands content development, system customization to institutional needs, instructor preparation, and technical infrastructure. Investments treating AI as comprehensive instructional transformation rather than simple technology purchase will more likely realize potential benefits.

8.4 Limitations and Future Research

This study's limitations suggest important future research directions. The 16-week timeframe, while substantial, cannot assess long-term retention or sustained effects after instruction ends.

10.48047/jocaaa.2024.33.05.74

Longitudinal research tracking learners over years would determine whether AI-driven instruction produces durable learning or merely faster short-term gains.

The framework's current limitations in assessing speaking and listening skills, given AI speech technology constraints, represent important gaps. As speech recognition and synthesis capabilities improve, integrating robust speaking and listening components would enable comprehensive language instruction. Research validating AI-driven speaking and listening instruction deserves priority.

The study population of university-affiliated adult learners may not generalize to other populations including young learners, workplace ESL, or refugee/immigrant populations with limited literacy. Validation across diverse contexts and populations would establish generalizability boundaries and identify necessary adaptations for different settings.

Future research should also examine optimal blends of AI-driven instruction and human teaching. This study's hybrid model showed promise, but systematic variation of the balance between individual AI-based study and classroom time could identify ideal configurations for different goals, populations, and resource constraints.

Finally, investigating AI-driven instruction's impacts on learner autonomy, motivation, and lifelong learning skills would address important outcomes beyond immediate proficiency gains. If AI systems create dependency rather than developing self-regulated learning, long-term benefits may prove limited despite short-term gains.

9. CONCLUSION

This research successfully designed, implemented, and validated an AI-driven instructional design framework for ESL education that demonstrates substantial learning benefits compared to traditional classroom instruction. The framework achieves its primary objective of delivering personalized, adaptive ESL instruction with measurably superior outcomes, showing 34% greater learning gains across diverse learner populations. Secondary objectives were similarly accomplished through development of effective NLP assessment capabilities, creation of successful adaptive learning algorithms, comprehensive evaluation across proficiency levels, and identification of mechanisms through which personalization enhances learning.

Three fundamental findings emerge from this research. First, AI-driven personalization significantly improves ESL learning outcomes through adaptive content alignment that matches instructional materials to individual learner needs, levels, and learning trajectories. The ability to continuously diagnose gaps and provide appropriately challenging content creates more efficient learning than one-size-fits-all curricula. Second, immediate automated feedback on unlimited practice opportunities enables learning volumes and correction speeds impossible in traditional classrooms, translating to accelerated proficiency development. Third, AI-driven instruction benefits learners across all proficiency levels and backgrounds, suggesting broad applicability, though implementation details should adapt to specific populations and contexts.

The comprehensive framework developed here integrates natural language processing for assessment, adaptive learning algorithms for personalization, rich multimedia content, and analytics supporting both learners and instructors. This integration within sound pedagogical design distinguishes the approach from simpler technology applications that digitize traditional

10.48047/jocaaa.2024.33.05.74

materials without fundamental instructional transformation. The framework demonstrates that AI can operationalize second language acquisition principles like comprehensible input and extensive practice with feedback at scale previously unattainable.

The empirical validation across 240 learners over 16 weeks with rigorous comparison to traditional instruction provides strong evidence for AI-driven ESL instruction effectiveness. Effect sizes of d = 0.65-0.78 represent meaningful practical impacts, reducing time to proficiency advancement by an estimated 5-7 weeks based on observed gain rates. These benefits combine with higher engagement, better retention, and increased student satisfaction to present compelling cases for AI-driven instruction adoption.

However, successful implementation requires more than technology deployment. The research demonstrates that personalization quality—how well content matches learner needs—matters more than mere AI presence. This demands substantial content development, algorithm training on ESL-specific data, and continuous system refinement based on learner outcomes. Institutions must invest in comprehensive implementation including instructor preparation, technical infrastructure, and ongoing support rather than treating AI as a simple software purchase.

The hybrid model combining AI-driven individual study with periodic instructor-led group sessions emerged as particularly promising, preserving personalization benefits while addressing learners' social needs and providing conversation practice opportunities that current AI cannot adequately support. This blended approach may represent optimal configurations balancing technology's strengths in individualized practice and assessment with human teachers' advantages in facilitating social interaction and providing nuanced feedback.

Looking forward, continued advances in AI technologies, particularly speech recognition and generation, will enable increasingly comprehensive language instruction addressing all skills equally. As natural language processing capabilities improve, automated feedback quality will approach and potentially exceed average human instructor performance in some dimensions. However, technology should augment rather than replace teachers, who remain essential for fostering motivation, providing cultural insights, facilitating communication, and addressing learners' social-emotional needs.

This research contributes both theoretical understanding and practical tools for leveraging artificial intelligence to improve ESL education. The framework provides a model for integrating AI technologies within sound instructional design principles. The empirical findings demonstrate that personalized, adaptive instruction significantly enhances language learning outcomes. The mechanistic insights reveal that alignment quality, practice volume, and feedback immediacy drive AI-driven benefits.

For the millions of ESL learners worldwide seeking English proficiency for education, employment, and integration, AI-driven instruction offers pathways to more effective, efficient, and accessible language education. By delivering personalization previously available only through expensive one-on-one tutoring, AI systems can democratize access to high-quality ESL instruction, potentially reducing language barriers that limit human potential and opportunity. Realizing this potential requires continued research, thoughtful implementation, and commitment to using technology in service of sound pedagogy and learner success.

10.48047/jocaaa.2024.33.05.74

REFERENCES

1. Black, P. and Wiliam, D. (2009) 'Developing the theory of formative assessment', *Educational Assessment, Evaluation and Accountability*, 21(1), pp. 5-31.
2. Branch, R.M. (2009) *Instructional Design: The ADDIE Approach*. New York: Springer.
3. Burston, J. (2015) 'Twenty years of MALL project implementation: A meta-analysis of learning outcomes', *ReCALL*, 27(1), pp. 4-20.
4. Chapelle, C.A. and Sauro, S. (2017) *The Handbook of Technology and Second Language Teaching and Learning*. Hoboken, NJ: Wiley-Blackwell.
5. DeKeyser, R. (2015) 'Skill acquisition theory', in VanPatten, B. and Williams, J. (eds.) *Theories in Second Language Acquisition*. 2nd edn. New York: Routledge, pp. 94-112.
6. Devlin, J., Chang, M.W., Lee, K. and Toutanova, K. (2019) 'BERT: Pre-training of deep bidirectional transformers for language understanding', in *Proceedings of NAACL-HLT 2019*. Minneapolis: ACL, pp. 4171-4186.
7. Dörnyei, Z. and Ryan, S. (2015) *The Psychology of the Language Learner Revisited*. New York: Routledge.
8. Ellis, R. (2003) *Task-Based Language Learning and Teaching*. Oxford: Oxford University Press.
9. Gass, S.M. and Selinker, L. (2020) *Second Language Acquisition: An Introductory Course*. 5th edn. New York: Routledge.
10. Godwin-Jones, R. (2018) 'Chasing the butterfly effect: Informal language learning online as a complex system', *Language Learning & Technology*, 22(2), pp. 8-27.
11. Grgurović, M., Chapelle, C.A. and Shelley, M.C. (2013) 'A meta-analysis of effectiveness studies on computer technology-supported language learning', *ReCALL*, 25(2), pp. 165-198.
12. Krashen, S. (1985) *The Input Hypothesis: Issues and Implications*. London: Longman.
13. Kulik, J.A. and Fletcher, J.D. (2016) 'Effectiveness of intelligent tutoring systems: A meta-analytic review', *Review of Educational Research*, 86(1), pp. 42-78.
14. Piech, C., Bassen, J., Huang, J., Ganguli, S., Sahami, M., Guibas, L.J. and Sohl-Dickstein, J. (2015) 'Deep knowledge tracing', in *Advances in Neural Information Processing Systems*, pp. 505-513.
15. Rama, P.S., Black, R.W., van Es, E. and Warschauer, M. (2012) 'Affordances for second language learning in World of Warcraft', *ReCALL*, 24(3), pp. 322-338.
16. Swain, M. (1995) 'Three functions of output in second language learning', in Cook, G. and Seidlhofer, B. (eds.) *Principle and Practice in Applied Linguistics*. Oxford: Oxford University Press, pp. 125-144.
17. Tomlinson, C.A. (2014) *The Differentiated Classroom: Responding to the Needs of All Learners*. 2nd edn. Alexandria, VA: ASCD.