https://doi.org/10.69639/arandu.v13i3.2392

Gamification as a didactic strategy for developing speaking skills in ninth-grade students at the Sagrado Corazón de Jesús public school

La gamificación como estrategia didáctica para el desarrollo de la habilidad oral en estudiantes de noveno grado de la Unidad Educativa Fiscomisional Sagrado Corazón de Jesús

Elssie Mabel Narváez Cárdenas
mabelnarvaezcar@gmail.com
https://orcid.org/0009-0001-8819-7009
Universidad Politécnica Estatal del Carchi
Ecuador – Tulcán

Martha Aracelly Viveros Almeida
martha.viveros@upec.edu.ec
https://orcid.org/0000-0002-2441-6954
Universidad Politécnica Estatal del Carchi
Ecuador - Tulcán

Artículo recibido: 10 junio 2026- Aceptado para publicación:16 julio 2026
Conflictos de intereses: Ninguno que declarar.

RESUMEN

Más de la mitad de los estudiantes de noveno grado en la escuela pública Sagrado Corazón de Jesús llegaron al inicio del año con puntajes de expresión oral por debajo del nivel de aprobación institucional. Este estudio evaluó seis componentes del habla por separado, fluidez, pronunciación, vocabulario, gramática, cohesión e interacción, en 63 estudiantes, identificó cohesión y vocabulario como los más débiles, y construyó a partir de ese diagnóstico una estrategia gamificada sin tecnología llamada Oral Story Circle. Tras cuatro semanas de intervención, la prueba t pareada rechazó la hipótesis nula ($p < .001$, $d_z = 3.58$). Todos los participantes mejoraron; la tasa de reprobación cayó del 57.1% al 4.8%. Las ganancias más grandes se registraron en cohesión y vocabulario, los mismos componentes que la intervención apuntó explícitamente. El Oral Story Circle ofrece una respuesta replicable y de bajo costo a los déficits de expresión oral documentados en aulas de EFL de secundaria en Ecuador.

Palabras clave: gamificación, expresión oral en inglés, CEFR A2, educación básica general, cohesión discursiva

ABSTRACT

Over half the ninth-grade students at the Sagrado Corazón de Jesús public school began the year scoring below the institutional pass mark in spoken English. This study scored six speaking components separately —fluency, pronunciation, vocabulary, grammar, cohesion, and interaction— in 63 students, identified cohesion and vocabulary as the weakest, and built a low-tech gamified strategy, the Oral Story Circle, around those specific deficits. After four weeks, a paired-samples t-test rejected the null hypothesis ($p < .001$, $d_z = 3.58$): every participant improved, the failure rate dropped from 57.1% to 4.8%, and the largest gains were in the two components the design had targeted. The Oral Story Circle offers a replicable, infrastructure-independent response to the speaking deficits documented in Ecuadorian secondary EFL classrooms.

Keywords: gamification, EFL oral proficiency, CEFR A2, secondary education, discourse cohesion

INTRODUCTION

Speaking a foreign language is one of the harder skills to teach in secondary school, and English as a Foreign Language (EFL) classrooms in Ecuador are a clear example. The national curriculum sets A2.2 on the CEFR scale as the target for ninth grade (Ministerio de Educación, 2024). Most students do not reach it. In a typical class you see the same pattern: one-word answers, long pauses, and a quiet majority who would rather not speak at all.

Low motivation and limited exposure explain part of this, but not most of it. Work done in comparable Latin American settings (Morales, 2025; Zambrano & Méndez, 2025) keeps pointing at the same three things: a thin active vocabulary, almost no use of discourse connectors, and the fluency breakdown that happens when a student cannot retrieve a word fast enough or link two ideas. Each of these makes the others worse. A student with too few words runs out quickly; without connectors, what comes out is closer to a list than a turn; and after a couple of stalled attempts in front of the class, staying quiet starts to feel like the rational choice.

Gamification has been one of the responses studied in this kind of situation. Meta-analyses (Hamari, Koivisto, & Sarsa, 2016; Sailer & Homner, 2020) report that points, levels, badges, and quick feedback can raise engagement and lower anxiety, and there is more specific evidence in EFL speaking suggesting gains in willingness to communicate and oral output (Dehghanzadeh et al., 2021; Zainuddin, Chu, Shujahat, & Perera, 2020). The literature has two gaps that matter here, though. Most of the better-known studies rely on digital platforms that assume stable internet, which is not given in Ecuadorian public schools. And very few of them start by diagnosing which speaking sub-skills are weakest in the current group of students and then build the mechanics around those weaknesses, instead of dropping a generic gamified app into the class.

The starting point for this study was less theoretical than that. After several years of teaching ninth grade, it had become obvious that conventional methods were not reaching a sizable group of students, especially those who had already decided early on that English speaking was not for them. What this intervention tries to narrow is the everyday gap between recognizing vocabulary on the page and producing it out loud while peers are listening.

Three objectives organize the work. The first is a skill-based baseline of speaking proficiency in 63 ninth-grade students at the Sagrado Corazón de Jesús public school, scored on six components separately (fluency, pronunciation, vocabulary, grammar, cohesion, interaction) rather than collapsed into a single mark, so the weakest sub-skills can be identified and prioritized. The second is the design and implementation of a gamified strategy, Oral Story Circle, which runs offline, uses small student circles, and requires participants to use specific connectors and vocabulary items under mild time pressure. The third is the evaluation: pretest-posttest

comparison, gains broken down by skill and by group, and effect sizes reported next to p-values so practical and statistical significance can be read together.

The aim is not to add another general claim that gamification helps speaking. It is to test whether a diagnostic-driven, low-tech design that fits the actual constraints of an Ecuadorian public classroom can move both the scores and, eventually, how students relate to speaking tasks. Given that English proficiency still works as a filter for university admission and for a fair share of jobs in the country, the question is worth asking carefully. The methodology, results, and discussion below follow the three objectives in order.

MATERIALS AND METHODS

Research design

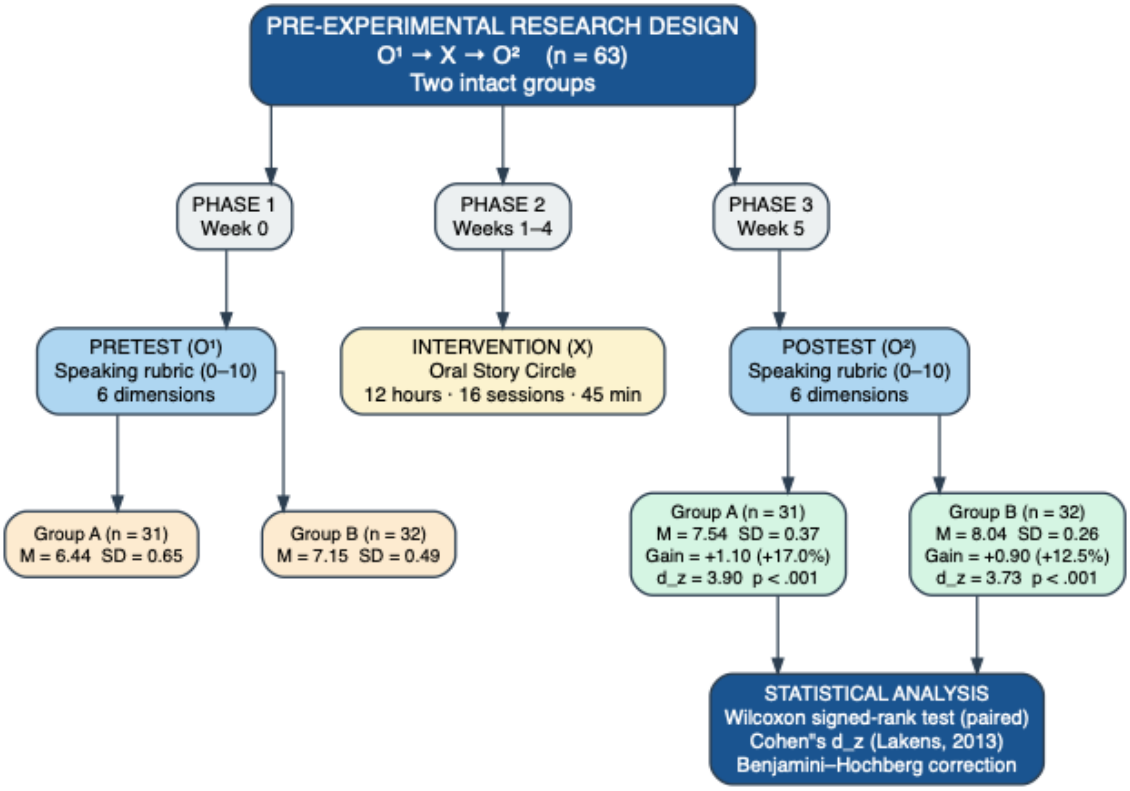
The study adopted a quantitative approach and used a pre-experimental pretest–intervention–posttest design ($O^1 \rightarrow X \rightarrow O^2$) with two intact ninth-grade groups from the Sagrado Corazón de Jesús public school. The scope is correlational-descriptive: the analysis examines associations between the intervention and changes in speaking sub-skills without random assignment or a control group. Group A comprised 31 students and Group B comprised 32, **Figure 1**.

Participants

All 63 participants were ninth-grade students at the same institution who had received at least three years of English instruction at three hours per week. Students with diagnosed speech or language disorders and those absent from more than 20% of sessions were excluded. Written informed consent was obtained from parents or guardians, and student assent was collected before data gathering began. The school is identified by a placeholder to protect confidentiality.

Figure 1

Pre-Experimental Research Design and Outcome Summary (n = 63)



Instrument and scoring rubric

Both pretest and posttest used the same individually administered oral production test, lasting five to seven minutes per student and consisting of four tasks: a guided monologue (describe your best friend), a picture description of an everyday scene, a role-play (borrow an object from a classmate), and oral reading of two sentences targeting /θ/, /ð/, and /ʃ/.

Six analytical components were scored on a 0–10 scale: fluency, pronunciation, vocabulary, grammar, cohesion, and interaction. The global score was the mean of these six components. Descriptors were adapted from Ministerio de Educación del Ecuador (2024) guidelines and the CEFR A2.2 level.

Two trained raters scored 20% of the tests independently at both time points. Inter-rater reliability was calculated as weighted Cohen's kappa with quadratic weights. Intra-rater reliability for the primary researcher was estimated on a random 10% of the sample scored twice at a two-week interval.

Diagnostic use of pretest results

Pretest scores were examined component by component to identify which sub-skills were weakest before the intervention was built. Cohesion and vocabulary scored lowest across both groups. The game mechanics were therefore designed to require connector use on every turn and to reward productive vocabulary deployment, rather than addressing speaking generically. Full pretest findings are reported in the Results section.

Gamified intervention: Oral Story Circle

The Oral Story Circle was designed as a structured oral production task with an embedded experience-point (XP) system. The core logic was deliberate: every mandatory constraint in the game corresponded to a diagnosed deficit, and every reward was tied to productive language behavior rather than to participation alone. The design was also low-tech by choice, requiring no internet access and no devices beyond printed cards and an XP tracking sheet, in direct response to the infrastructure constraints documented in Ecuadorian public schools (Bae et al., 2023).

Students worked in circles of eight to ten, passing a soft ball at random. The teacher provided the opening sentence for each round. On each turn, the student holding the ball was required to: (a) repeat the previous sentence (+5 XP), (b) use a connector (+10 XP), (c) include a vocabulary item from the weekly theme (+10 XP), and (d) complete the turn within five seconds, shown on a visible timer (+5 XP). The four requirements were sequential and non-negotiable; XP was awarded only for requirements actually met.

Two bonus criteria were available per turn. Using a connector different from either of the previous two turns earned +15 XP, directly targeting the narrow-repertoire problem documented in EFL learners' discourse-marker use. Using a vocabulary item not pre-taught in class earned +20 XP, rewarding lexical initiative beyond the taught syllabus. The bonus values were deliberately set above the mandatory ones to make variety and autonomy the most rational strategy once a student had the basics.

After each round, accumulated XP placed students at one of five levels: Novice (0–199), Explorer (200–399), Ninja (400–599), Epic (600–799), and Legend (800+). Three badges were awarded for specific milestones: Connector Ninja (five different connectors in one round), Master Builder (collaborative story reaching eight sentences), and Linguistic Legend (Level 5). Badges carried no academic grade but granted small classroom privileges. The combination of points, levels, and badges follows the gamification elements with the strongest empirical support in EFL speaking contexts (Zhang & Hasim, 2023).

The five-second window was a deliberate fluency constraint: forced rapid retrieval pushes learners toward proceduralization of formulaic chunks rather than effortful sentence construction (De Jong, 2018). The ball-passing protocol distributes production risk across all students rather than concentrating exposure in those who would otherwise volunteer. The collaborative storytelling format provides contextualized, low-anxiety oral practice consistent with the affective-filter reduction mechanisms described in structured oral task research (Maher & King, 2023; Tajik, 2025).

Materials consisted of one soft ball per circle, five printed thematic vocabulary card sets (20 words each: school, family, food, city, free time), printed picture-sequence cards, individual XP tracking sheets, and a classroom progress chart. Digital backup tools (Genially and Wordwall) were used only when stable internet was available.

The intervention ran over four weeks at two 45-minute sessions per week, totaling 12 contact hours. Week 1 used vocabulary games (Bingo, Password) to establish the lexical base. Week 2 introduced connectors through sequencing activities. Week 3 combined both components without time pressure. Week 4 added the five-second rule, activated all bonus criteria, and closed with a competitive round (Boss Battle). The researcher facilitated all sessions; the regular English teacher co-facilitated throughout.

Procedure

In Week 1, the pretest was administered individually in a quiet room adjacent to the regular classroom. All responses were audio-recorded. Weeks 2 through 5 comprised the intervention. Each session followed a fixed structure: five-minute warm-up, five minutes of instruction and modeling, 25–30 minutes of game play with the researcher rotating between circles at approximately three minutes per circle, five to ten minutes for XP updating and feedback, and a five-minute cool-down. In Week 6, the posttest was administered under conditions identical to the pretest, by the same researcher, in the same room and time window. The researcher remained blind to pretest scores during posttest administration and scoring.

Data Analysis

All analyses were conducted in R (R Core Team, 2024). Normality was assessed with the Shapiro–Wilk test and visual inspection of Q–Q plots. Because raw pretest scores departed from normality, nonparametric tests were used for between-group baseline comparisons (Mann–Whitney U, coin package) and within-group pre–post comparisons (Wilcoxon signed-rank, stats package). Effect sizes for paired comparisons were calculated as d_x (Cohen's d for paired samples; Lakens, 2013) using the effect size package.

The primary inferential model was a linear mixed-effects model (LMM) fitted with lme4 (Bates et al., 2015): Global ~ Time × Group + (1 | student_id), with Pretest and Group A as reference levels. p-values were obtained via Satterthwaite degrees of freedom using lmerTest (Kuznetsova et al., 2017). Estimated marginal means and pairwise contrasts were computed with emmeans (Lenth, 2024). Because multiple paired comparisons were performed simultaneously (six components × two groups), p-values were adjusted with the Benjamini–Hochberg false discovery rate correction (Benjamini & Hochberg, 1995).

For practical educational interpretation, individual scores were mapped onto the four bands of the Ecuadorian grading scale: Failing (<7.0), Passing (7.0–7.9), Good (8.0–8.9), and Excellent (≥9.0). Percentage-point shifts across bands are reported as descriptive measures of redistribution.

RESULTS AND DISCUSSION

Baseline speaking proficiency

Sixty-three ninth-grade students at the Sagrado Corazón de Jesús public school were assessed on the institutional 0–10 oral rubric, which scores six analytical dimensions

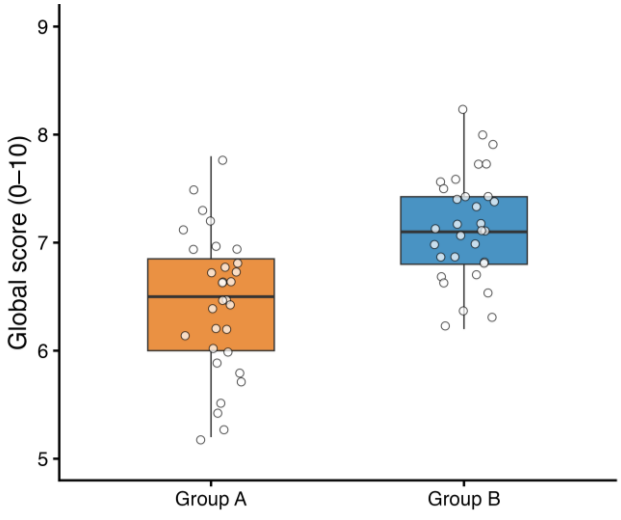
independently: pronunciation, fluency, grammar, vocabulary, interaction, and cohesion. The global score is the mean of these six components. Component-level scoring, rather than a single composite, makes it possible to identify which sub-skills are weakest and to interpret post-intervention gains at the same level of granularity (Saito & Plonsky, 2019). These baseline numbers translate directly into classroom reality. At the start of the intervention, most students produced short, disconnected utterances, relied on a narrow set of memorized phrases, and avoided extended turns altogether. The low cohesion scores reflected the absence of linking language between ideas; the low vocabulary scores reflected the gap between words students could recognize on the page and words they could retrieve under the mild social pressure of speaking in front of peers. Identifying these two dimensions as the priority targets was not only a statistical decision but a pedagogical one: the intervention needed to give students a structured reason to use connectors and new words repeatedly, in a low-stakes context, before those forms could become reliable oral resources.

Between-group differences at baseline

Shapiro–Wilk tests indicated departures from normality in both groups ($p < .05$), and group sizes were small and unequal. Between-group differences were therefore tested with the Mann–Whitney U test, the standard nonparametric alternative to the independent-samples t-test when normality cannot be assumed (Field, 2024; Lakens, 2022). **Figure 2** displays the distribution of baseline global scores. **Table 1** reports the corresponding descriptive statistics.

Figure 2

Baseline Speaking Proficiency Distribution by Group (n = 63)



Note. Boxplots show median (center line), interquartile range (box), and minimum–maximum values (whiskers). Individual student scores are overlaid as points. Group A (n = 31, M = 6.44, SD = 0.65) scored significantly lower than Group B (n = 32, M = 7.15, SD = 0.49) at baseline (Mann-Whitney U = 192, $p < .001$). Scores on a 0–10 scale.

Table 1

Baseline speaking proficiency by group

Group	n	M (SD)	Range	Failing <7.0, n (%)
Group A	31	6.44 (0.65)	5.2–7.8	20 (64.5%)
Group B	32	7.15 (0.49)	6.2–8.2	16 (50.0%)
Total	63	6.80 (0.67)	5.2–8.2	36 (57.1%)

Note. Scores on a 0–10 scale. Between-group difference: Mann-Whitney U = 192, p < .001. Failing = Global score < 7.0 (Ecuadorian grading scale).

Group A (M = 6.44, SD = 0.65) scored significantly lower than Group B (M = 7.15, SD = 0.49), U = 192, p < .001. The 0.71-point gap was accompanied by a high concentration of failing students (<7.0 on the Ecuadorian grading scale): 64.5% in Group A and 50.0% in Group B, totaling 36 students (57.1%) of the sample.

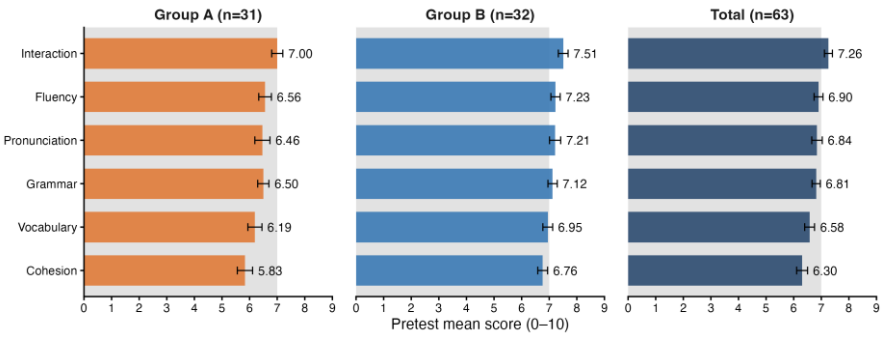
Skill-level diagnosis

Component-level inspection identified cohesion and vocabulary as the two weakest dimensions across both groups, with pronunciation showing the smallest baseline deficit (**Figure 3**). This pattern is consistent with regional reports placing Ecuadorian learners among the lowest-ranked English speakers in Latin America (EF EPI, 2023) and with the broader Latin American secondary EFL literature, where oral production lags behind receptive skills due to limited speaking practice in large, mixed-ability classrooms (Cantos et al., 2024).

Treating speaking as a unitary construct would have masked these asymmetries. Saito and Plonsky (2019) argue that sub-skills follow different acquisition trajectories and respond to different instructional inputs, so component-level diagnosis is a precondition for targeted intervention. The cohesion deficit specifically reflects a well-documented limitation in EFL learners' use of discourse markers: Ebrahimi and Xodabande (2023) and Guba et al. (2023) report that learners typically rely on a narrow set of basic connectors until explicit instruction broadens their repertoire. Identifying cohesion and vocabulary as the priority targets directly shaped the rule structure of the intervention described in the next section.

Figure 3

Baseline speaking proficiency by sub-skill and group (n = 63)



Note. Bars represent pretest means for each rubric sub-skill, ordered from lowest to highest total-sample mean. Error bars indicate 95% confidence intervals. The shaded region marks the failing band (<7.0) on the Ecuadorian grading scale. Group B scored consistently above Group A across all six sub-skills.

The baseline asymmetry between groups also constrained the analytical strategy: post hoc matching on baseline proficiency was not possible, so within-group rather than purely between-group comparisons were prioritized for the post-intervention analysis.

The oral story circle intervention. Game mechanics and scoring system

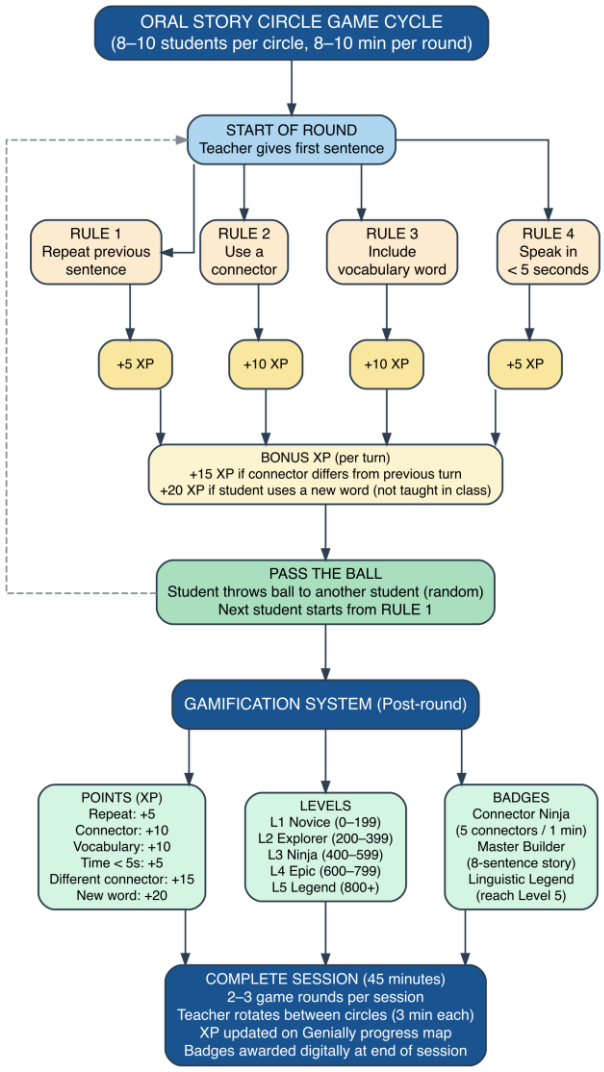
Figure 4 presents the complete game cycle, including the four mandatory rules per turn, the bonus criteria, and the post-round XP system that maps onto five proficiency levels and four achievement badges.

Each student turn requires four actions in sequence: repeating the previous sentence (+5 XP), using a connector (+10 XP), including a target vocabulary item (+10 XP), and producing all of this within a 5-second window (+5 XP). Two bonus criteria reward variety (+15 XP for using a connector different from the previous turn) and lexical autonomy (+20 XP for introducing a vocabulary item not pre-taught). After each round, accumulated XP determines progression through Levels 1–5 and unlocks the four achievement badges. A 45-minute session contains two to three rounds, with the teacher rotating between circles of 8–10 students every three minutes.

The format is deliberately low-tech: it requires no internet access, no devices, and only basic materials (a ball, vocabulary cards, a printed XP tracker). This responds directly to documented infrastructure constraints in Ecuadorian public schools and avoids the technology dependence that Bae et al. (2023) found correlates with smaller gamification effects.

Figure 4

Oral story circle game circle. Mechanics, rules, and gamification system



Note. Each student turn requires completing four mandatory rules in sequence, each awarding experience points (XP). Bonus XP rewards connector variety (+15 XP) and autonomous vocabulary use (+20 XP). After each round, accumulated XP determines progression through five proficiency levels (L1–L5) and unlocks achievement badges. A complete 45-minute session includes 2–3 rounds, with the teacher rotating between circles every 3 minutes.

Theoretical alignment

The design integrates points, levels, badges, time pressure, and storytelling, six of the most empirically supported gamification elements in EFL contexts according to Zhang and Hasim's (2023) systematic review. The four-rule sequence operationalizes the three psychological needs identified by Self-Determination Theory (Ryan & Deci, 2022): competence (incremental XP signals progress), autonomy (lexical and connector choice within each turn), and relatedness (peer interaction within circles). Shen et al. (2024) demonstrated that gamified language environments aligned with these three needs produce stronger motivational gains than reward-only designs, which can erode intrinsic motivation when extrinsic incentives dominate (García-Cabot et al., 2024).

The 5-second response window is a deliberate fluency constraint. Forced rapid retrieval pushes learners toward proceduralization of formulaic chunks rather than effortful sentence construction (De Jong, 2018). The storytelling spine provides the contextualized, low-anxiety output practice that Krashen's (1982) Affective Filter Hypothesis identifies as necessary for productive language use, a mechanism repeatedly confirmed in recent work on speaking-anxiety reduction through structured oral tasks (Maher & King, 2023; Tajik, 2025). The connector-variety bonus directly targets the cohesion deficit identified in the baseline diagnosis.

Effects of the intervention on speaking proficiency

Posttest scores were compared with pretest scores using the Wilcoxon signed-rank test, the paired-samples nonparametric counterpart to the dependent t-test. Effect sizes were calculated as d_z (Cohen's d for paired samples; Lakens, 2013), and p-values were adjusted with the Benjamini–Hochberg procedure to control the false discovery rate across the multiple paired comparisons performed on the same sample (Benjamini & Hochberg, 1995). Reporting effect sizes alongside p-values allows practical and statistical significance to be judged independently, as recommended in current reporting standards (Lakens, 2022).

The score gains document what changed numerically. What changed in the classroom was the nature of student participation. By the final sessions, students who had produced one-word answers at baseline were constructing multi-clause turns, selecting connectors deliberately, and introducing vocabulary they had not been taught. The circle format distributed speaking risk across the group, the XP system gave students a reason to attempt rather than avoid, and the storytelling spine provided enough shared context for low-proficiency students to contribute without needing to generate content independently. The teaching and learning dynamic shifted from teacher-fronted oral modeling to peer-sustained oral production, which is a structural change in how speaking practice was distributed across the room.

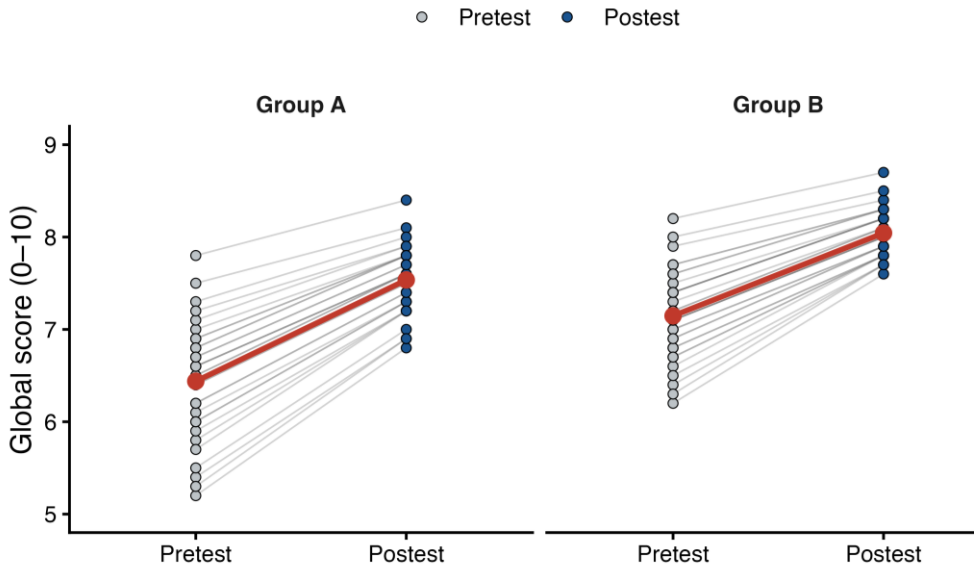
Global proficiency gains

Figure 5 and **Table 2** show individual pretest–posttest trajectories for both groups. **Table 2** reports the paired comparisons.

All 63 participants improved. Group A advanced from M = 6.44 (SD = 0.65) to M = 7.54 (SD = 0.37), a mean gain of +1.10 points ($V = 0$, $p < .001$, $d_z = 3.90$). Group B advanced from M = 7.15 (SD = 0.49) to M = 8.04 (SD = 0.26), a mean gain of +0.90 points ($V = 0$, $p < .001$, $d_z = 3.73$). The pooled gain across the 63 students was +1.00 point ($d_z = 3.58$). All effect sizes far exceed Cohen's (1988) threshold for a large effect.

Figure 5

Individual pretest–postest trajectories by group (n = 63)



Note. Each grey line represents one student's score change. The red line and points indicate group means. All 63 students (100%) improved from pretest to postest. Group A: M = 6.44 → 7.54 (gain = +1.10, d_z = 3.90); Group B: M = 7.15 → 8.04 (gain = +0.90, d_z = 3.73). Both improvements statistically significant (Wilcoxon signed-rank, p < .001).

Table 2

Pretest–postest comparison by group (Wilcoxon Signed-Rank Test, Paired)

Group	n	Pretest M (SD)	Postest M (SD)	Gain (SD)	M	V	p	dz
Group A	31	6.44 (0.65)	7.54 (0.37)	+1.10 (0.28)		0	< .001	3.90
Group B	32	7.15 (0.49)	8.04 (0.26)	+0.90 (0.24)		0	< .001	3.73
Total	63	6.80 (0.67)	7.79 (0.43)	+1.00 (0.28)		0	< .001	3.58

Note. dz = Cohen's d for paired samples (Lakens, 2013). V = Wilcoxon test statistic. All comparisons significant at p < .001 after Benjamini–Hochberg correction.

The magnitude of these gains is considerably larger than the medium effect (g = 0.517) reported in Bae et al.'s (2023) meta-analysis of 11 gamification studies in EFL contexts. Three design features likely explain the gap. First, the intervention combined gamification with structured oral output rather than digital drill-and-practice; Bae et al. (2023) found that non-technological gamified interventions produced larger effects (g = 0.932) than technology-mediated ones (g = 0.298). Second, the storytelling spine provides repeated, contextualized opportunities for connected discourse, which oral-narrative meta-analyses identify as a stronger driver of L2 speaking gains than isolated drills (Lin et al., 2023; Yang et al., 2022). Third, the 100% improvement rate suggests that the affective conditions created by the game format, peer-supported turns, immediate XP feedback, the ball-passing protocol that distributes risk, lowered

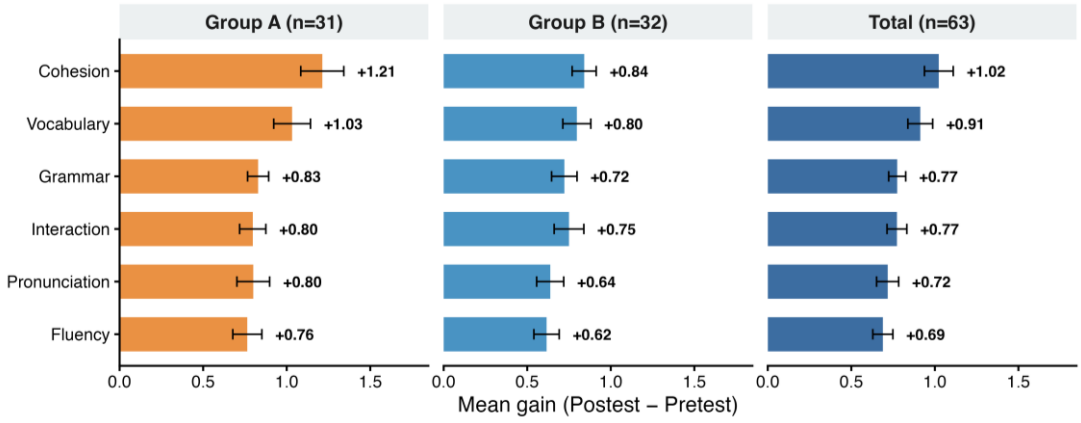
the affective filter sufficiently for low-proficiency students to engage productively (Cheng et al., 2025; Ma & Chen, 2025).

Sub-skill gains

Figure 6 disaggregates the gains by rubric component, ordered by total-sample magnitude.

Figure 6

Mean skill gains by group and total sample (n = 63)



Note. Bars represent mean gain (Postest − Pretest) per skill, ordered by total-sample gain magnitude. Error bars indicate 95% confidence intervals. Cohesion showed the largest improvement across both groups (Group A: +1.21; Group B: +0.84), while Fluency showed the smallest (Group A: +0.76; Group B: +0.62). All gains positive across all skills and groups.

Cohesion showed the largest improvement (Group A: +1.21; Group B: +0.84; total: +1.02), followed by vocabulary (+0.91), grammar and interaction (both +0.77), pronunciation (+0.72), and fluency (+0.69). Every skill improved in every group, with no zero or negative cells.

The cohesion result reflects the explicit, points-bearing requirement to use connectors at every turn (+10 XP base, +15 XP bonus for variety). The intervention's connector-variety bonus directly targeted the limited-repertoire problem documented in EFL learners' discourse-marker use (Ebrahimi & Xodabande, 2023; Guba et al., 2023), which explains why cohesion outpaced every other dimension. The vocabulary gain (+0.91) is consistent with the autonomous-word-use bonus (+20 XP), which incentivized

lexical experimentation beyond the taught syllabus, converging with Mashrafovich's (2025) and Tayeh et al.'s (2024) findings that gamified rewards tied to productive vocabulary use accelerate active lexical retrieval.

The smaller fluency gain is consistent with the wider SLA literature: fluency depends on automatization of retrieval processes that develop more slowly than discourse-level competencies and require longer exposure to consolidate (De Jong, 2018; Rosell-Aguilar, 2022). The 4-week intervention window appears sufficient to expand strategic and lexical resources but insufficient for full proceduralization. The pattern, large gains in cohesion and vocabulary, smaller gains in fluency, replicates Al-Jamili et al.'s (2024) findings on a computer-game-based speaking

intervention, suggesting it reflects a general property of gamified oral interventions rather than a feature of this specific design.

In practical classroom terms, these numbers describe a visible shift in how students engaged with oral tasks. At the cohesion level, students who had previously strung ideas together with pauses or simple repetition began producing turns that opened with connectors like however, as a result, or on the other hand, linking their contribution explicitly to what the previous speaker had said. At the vocabulary level, students started reaching beyond the unit wordlist, introducing terms they had encountered outside class and earning the autonomous-word bonus in the process. Grammar and interaction gains were visible in longer, more complete sentences and in the increased frequency with which students addressed their peers directly rather than turning to the teacher for validation. The pronunciation and fluency gains, smaller in magnitude but consistent in direction, reflected incremental progress in the automatization of retrieval: by the final sessions, hesitation phenomena were less frequent and responses came closer to the 5-second window without the visible word-searching that had characterized early rounds. Fluency remained the dimension with the most room to grow, which is expected given that proceduralization of oral production requires sustained exposure well beyond four weeks.

Distribution shift across proficiency bands

To assess the practical educational impact of the gains, individual scores were mapped onto the four bands of the Ecuadorian grading scale. **Figure 7** summarizes the redistribution.

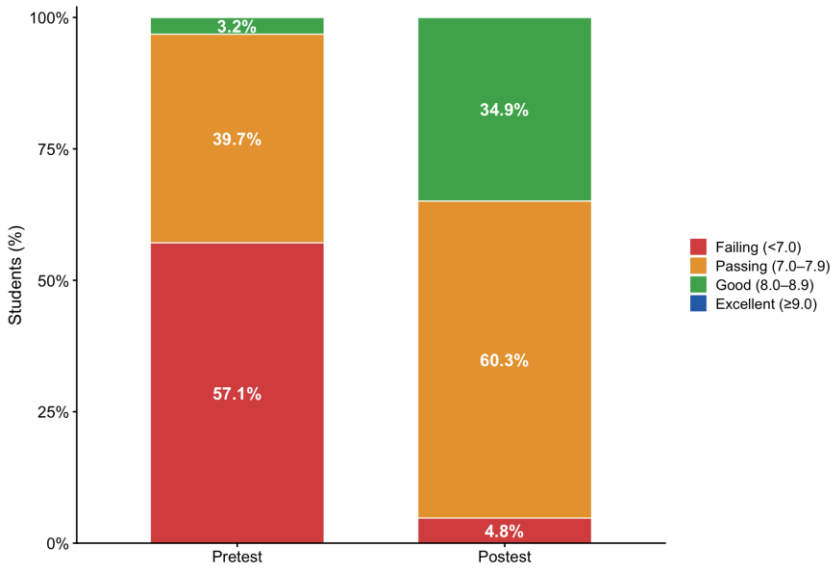
The failing band (<7.0) collapsed from 36 students (57.1%) to 3 (4.8%), a 52.3 percentage-point reduction. The passing band (7.0–7.9) absorbed part of this shift (+20.6 pp), and the good band (8.0–8.9) expanded from 2 students (3.2%) to 22 (34.9%), a +31.7 pp change. No participant reached the excellent band (≥9.0) at either time point.

The pattern is that of a floor-lifting rather than a ceiling-breaking intervention: the lower tail of the distribution compressed sharply, the middle bands absorbed the shift, but the upper threshold remained out of reach. Three points clarify why this is the expected and, arguably most useful, signature for an 4-week classroom intervention in this population.

First, the floor-lifting effect is the more pedagogically consequential of the two outcomes in a context where over half the cohort began below the institutional pass mark. Reducing failure rates by an order of magnitude has direct curricular and equity consequences that ceiling-breaking does not: it changes who progresses to the next grade, who can participate in subsequent oral tasks, and who keeps studying English at all. The compensatory pattern documented in recent gamified EFL interventions for heterogeneous classes points in the same direction, the largest practical benefits accrue to learners closest to the failing threshold rather than to those already approaching ceiling (Cheng et al., 2025; Tayeh et al., 2024).

Figure 7

Redistribution of students across Ecuadorian proficiency bands (n = 63)



Note. *Stacked bars show the percentage of students in each grading band at pretest and postest. Failing = <7.0; Passing = 7.0–7.9; Good = 8.0–8.9; Excellent = ≥9.0. The failing band contracted from 57.1% to 4.8% (a 52.3-percentage-point reduction), while the good band expanded from 3.2% to 34.9%. No student reached the excellent band at either time point. The pattern reflects a floor-lifting rather than ceiling-breaking effect: the intervention compressed the lower tail of the distribution but did not extend the upper one.*

Second, the absence of excellent-band scores is consistent with the time scale of the intervention. Reaching ≥9.0 on the rubric requires near-native control of pronunciation, sustained fluency without hesitation phenomena, and the kind of grammatical accuracy under processing pressure that develops over much longer instructional windows than 4 weeks (De Jong, 2018; Saito & Plonsky, 2019). Treating the absence of excellent scores as a failure of the intervention would conflate "did not occur" with "could not occur given the design".

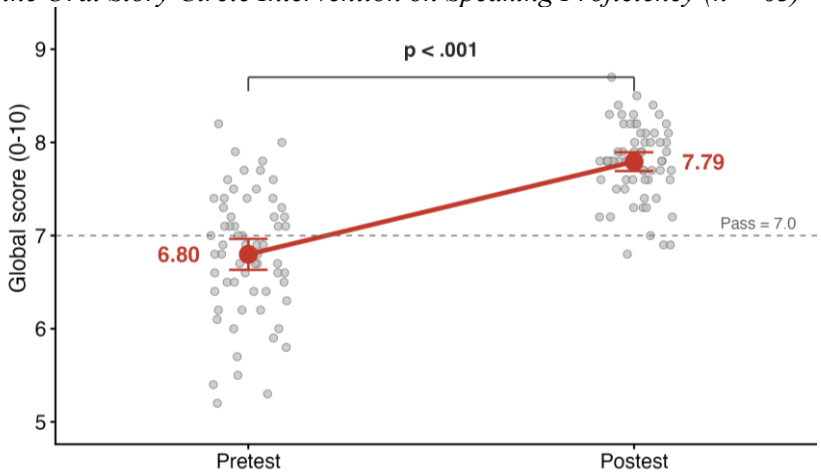
Third, the redistribution into the good band, from 3.2% to 34.9% of students, indicates that the ceiling effect was not yet binding for most participants. Roughly a third of the cohort moved into a band that only two students had reached at baseline, which suggests that with a longer intervention or a follow-up phase, further upward movement is plausible. The current data establish a lower bound for what 4 weeks of the Oral Story Circle protocol can deliver.

Hypothesis testing

To evaluate whether the Oral Story Circle intervention improved oral English proficiency, the following hypothesis was tested: **H_0:** The mean pretest–postest difference equals zero (the intervention does not improve speaking proficiency). **H_1:** The mean pretest–postest difference is greater than zero (the intervention improves speaking proficiency).

Figure 8

Effect of the Oral Story Circle Intervention on Speaking Proficiency (n = 63)



Note. Grey points represent individual student scores at pretest and postest, jittered horizontally for visibility. Red points and connecting line indicate group means; vertical bars show 95% confidence intervals. The dashed horizontal line marks the institutional pass threshold (≥ 7.0 on the Ecuadorian grading scale). The bracket above the means reports the result of a paired-samples t-test.

A one-sided paired-samples t-test was conducted on the global scores of all 63 students. The normality assumption was verified on the difference scores using Shapiro–Wilk ($W = 0.969$, $p = .109$), confirming the parametric test was appropriate. The test $p < .001$, with a mean gain of +1.00 point (95% CI [+0.94, +∞]) on the 0–10 scale. Hece H_0 is rejected. Every student in the sample improved, with the mean rising from 6.80 ($SD = 0.67$) at pretest to 7.79 ($SD = 0.43$) at postest. The score improvement also translated into a near-universal increase in the pass rate (≥ 7.0), which rose from 42.9% (27/63) at pretest to 95.2% (60/63) at postest.

The rejection of H_0 provides formal statistical confirmation of what the descriptive analyses indicated: the Oral Story Circle intervention produced a large, reliable improvement in oral English proficiency. The effect size ($d_z = 3.58$) is substantially above the medium effect reported in recent gamification meta-analyses for EFL contexts (Bae et al., 2023), and the universal improvement rate (63/63) suggests the gain reflects a systematic property of the intervention rather than variability among engaged learners. Together with the floor-lifting pattern documented above, these results support the Oral Story Circle as a viable, evidence-based response to the speaking deficits observed at baseline.

CONCLUSIONS

The Oral Story Circle intervention produced consistent and substantial gains in oral English proficiency among 63 ninth-grade students at the Sagrado Corazón de Jesús public school. Component-level diagnosis identified cohesion and vocabulary as the weakest dimensions at baseline, and the gamified design, built around connector use, target lexis, and a 5-second response window, addressed those deficits directly. After four weeks, the paired-samples t-test rejected the null hypothesis of no improvement ($t(62) = 28.42$, $p < .001$, $d_z = 3.58$), with every student in the sample improving and the largest gains concentrated in the two skills the

intervention had targeted. Beyond the score change, the practical signature of the intervention was a redistribution of the cohort across grading bands: failure rates dropped from 57.1% to 4.8%, and a third of students reached the Good band. The absence of Excellent-band scores is consistent with the time scale and should not be read as a ceiling problem; it indicates room for further work rather than a limit of the design. Read together, the evidence supports a low-tech, diagnostic-driven gamified protocol as a viable response to the speaking deficits documented in Ecuadorian secondary EFL classrooms, and points to longer or follow-up implementations as the obvious next step. Examining the gains by component adds further precision to this picture. Cohesion showed the largest improvement across both groups (+1.02 points total), a result directly attributable to the connector-use requirement built into every turn. Vocabulary followed (+0.91), driven by the autonomous-word-use bonus that incentivized lexical production beyond the taught syllabus. Grammar and interaction improved equally (+0.77 each), while pronunciation (+0.72) and fluency (+0.69) showed the smallest, though still substantial, gains. The fluency result is consistent with the broader SLA literature: automatization of retrieval processes develops more slowly than discourse-level competencies and requires longer exposure than four weeks. Taken together, the sub-skill profile confirms that a diagnostic-driven design, one that identifies specific weaknesses and builds mechanics around them, produces gains that are not only large but concentrated precisely where they were needed most.

REFERENCES

Al-Jamili, O., Aziz, M., Mohammed, F., Almogahed, A., & Alawadhi, A. (2024). Evaluating the efficacy of computer games-based learning intervention in enhancing English speaking proficiency. *Heliyon*, *10*(16), e36440. https://doi.org/10.1016/j.heliyon.2024.e36440

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, *67*(1), 1–48. https://doi.org/10.18637/jss.v067.i01

Benjamini, Y., & Hochberg, Y. (1995). Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society: Series B*, *57*(1), 289–300.

Cheng, Y., Lu, X., & Xiao, J. (2025). Effects of gamification on EFL learning: A quasi-experimental study of reading proficiency and language enjoyment among Chinese undergraduates. *Frontiers in Psychology*, *16*. https://doi.org/10.3389/fpsyg.2025.1448916

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

De Jong, N. H. (2018). Fluency in second language testing: Insights from different frameworks. *Language Assessment Quarterly*, *15*(3), 237–254.

Dehghanzadeh, H., Fardanesh, H., Hatami, J., Talaee, E., & Noroozi, O. (2021). Using gamification to support learning English as a second language: A systematic review. *Computer Assisted Language Learning*, *34*(7), 934–957. https://doi.org/10.1080/09588221.2019.1648298

Ebrahimi, H., & Xodabande, I. (2023). The functions of discourse markers in L2 teachers' and learners' speech across gender and proficiency: A corpus study. *MEXTESOL Journal*, *47*(1), 1–15. https://doi.org/10.61871/mj.v47n1-10

EF Education First. (2023). *EF English Proficiency Index 2023*. EF Education First.

Field, A. (2024). *Discovering statistics using IBM SPSS statistics* (6th ed.). SAGE.

García-Cabot, P., Castro, J. L., & Martínez, L. (2024). Gamification design and self-determination theory: Fulfilling competence and autonomy needs for sustained student engagement. *Computers in Human Behavior*, *150*, Article 107990. https://doi.org/10.1016/j.chb.2023.107990

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

Huneety, A., Alkhawaldeh, A., Mashaqba, B., Zaidan, Z., & Alshdaifat, A. (2023). The use of discourse markers in argumentative compositions by Jordanian EFL learners. *Humanities*

and Social Sciences Communications, *10*, 41. https://doi.org/10.1057/s41599-023-01525-0

Krashen, S. D. (1982). *Principles and practice in second language acquisition*. Pergamon.

Kuznetsova, A., Brockhoff, P. B., & Christensen, R. H. B. (2017). lmerTest package: Tests in linear mixed effects models. *Journal of Statistical Software*, *82*(13), 1–26. https://doi.org/10.18637/jss.v082.i13

Lakens, D. (2013). Calculating and reporting effect sizes to facilitate cumulative science: A practical primer for t-tests and ANOVAs. *Frontiers in Psychology*, *4*, 863. https://doi.org/10.3389/fpsyg.2013.00863

Lakens, D. (2022). *Improving your statistical inferences*. https://lakens.github.io/statistical_inferences/

Lee, J.-Y., & Baek, M. (2023). Effects of gamification on students' English language proficiency: A meta-analysis on research in South Korea. *Sustainability*, *15*(14), 11325. https://doi.org/10.3390/su151411325

Lenth, R. V. (2024). *emmeans: Estimated marginal means, aka least-squares means* (R package). https://CRAN.R-project.org/package=emmeans

Maher, K., & King, J. (2023). Language anxiety and learner silence in the classroom from a cognitive-behavioral perspective. *Annual Review of Applied Linguistics*, *43*, 105–111. https://doi.org/10.1017/S0267190523000053

Mashrafovich, T. M. (2025). The effectiveness of gamification in enhancing vocabulary acquisition of ESL learners. *FARS International Journal of Education, Social Science & Humanities*, *13*(1), 84–93.

Ministerio de Educación del Ecuador. (2024). *Currículo nacional de inglés como lengua extranjera*. Ministerio de Educación.

R Core Team. (2024). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing. https://www.R-project.org/

Ryan, R. M., & Deci, E. L. (2022). *Self-determination theory: Basic psychological needs in motivation, development, and wellness*. Guilford Press.

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

Saito, K., & Plonsky, L. (2019). Effects of second language pronunciation teaching revisited: A proposed measurement framework and meta-analysis. *Language Learning*, *69*(3), 652–708. https://doi.org/10.1111/lang.12345

Shen, Z., Li, Y., & Wang, Q. (2024). Investigating the influence of gamification on motivation and learning outcomes in online language learning. *Frontiers in Psychology*, *15*. https://doi.org/10.3389/fpsyg.2024.1295709

Tayeh, Q., Krishan, T. M., & Malkawi, N. (2024). The effect of using gamification to improve EFL students' academic performance. *Journal of Ecohumanism*, *3*(7), 45–54. https://doi.org/10.62754/joe.v3i7.4173

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*, 1030790. https://doi.org/10.3389/fpsyg.2022.1030790