SCRIPT JOURNAL
Journal of Linguistic and English Teaching
http://jurnal.fkip-uwgm.ac.id/index.php/Script
P-ISSN: 2477-1880; E-ISSN: 2502-6623
August 2026, Vol. 11 No. 2

A Bundled Gamified TOEFL Reading Training in an Islamic Boarding School: A Pre-Experimental Study on Critical Reading Skills and Implications for Quality Education

Amiruddin Hadi Wibowo[1], Agus Salim Syukran[2], Rosydina Robi'aqolbi[3], Faqru Nisa' Arrahmah[4]
English Education Study Program, Universitas Wijaya Kusuma Surabaya, Indonesia[1]
Sekolah Tinggi Ilmu Al-Qur'an dan Sains Al-Ishlah Lamongan, Indonesia[2,3,4]
Email Correspondence: Amiruddin_fbs@uwks.ac.id

Abstract

Background:
The EFL learners in the Indonesian Islamic boarding school (pesantren) have difficulty developing the critical reading abilities needed in TOEFL Reading comprehension. Gamification is generally known to improve learner engagement, but empirical evidence of its impact to the development of critical reading in pesantren is still rare. This research studied a Gamified TOEFL Reading Training that is predicted to assist santri's critical reading thru organized gamification aspects.

Methodology:
One-group pre-experimental pretest-posttest design was used with 35 purposefully selected East Java Islamic boarding school santri. Participant training comprised 16 sessions (8 weeks). Data was collected from TOEFL Reading and Critical Reading tests, a motivation and engagement questionnaire after the intervention, and classroom observations. Paired-sample t-tests, N-Gain, and within-subject Cohen's dz effect sizes were examined.

Findings:
TOEFL Reading scores were substantially improved following the intervention, mean difference = 12.4, $t(34) = 14.23$, $p < .001$, $g = 0.44$, $dz = 2.41$. The score of critical reading improved by 4.7 points, $t(34) = 11.67$, $p < .001$, $g = .55$, $dz = 1.97$. The 95% confidence intervals for TOEFL Reading and Critical Reading were [10.6, 14.2] and [3.9, 5.5]. The confidence interval of dz was [1.76, 3.06]. The results show a clear improvement in both categories after the training. The results should be evaluated carefully because the study was single-group and did not include a control group.

Conclusion:
Gamified TOEFL Reading instruction shows potential for supporting EFL learners' reading achievement, critical reading development, and engagement in pesantren. Further controlled studies are needed to establish causal effectiveness.

Originality:
This study's contribution is the specific combination of TOEFL-oriented critical reading outcomes, an Islamic boarding school (pesantren) setting, and classroom observation within one bundled instructional package, together with transparent normalised gain and within-subject effect-size reporting under explicit methodological caution.

Keywords	:	*Gamification; TOEFL reading; Critical reading; Islamic boarding school; SDG 4.*
DOI	:	10.24903/sj.v11i2.2432
Received	:	June 2026
Accepted	:	August 2026
Published	:	August 2026

How to cite this article (APA)	:	Wibowo, A. H., Syukran, A. S., Robi'aqolbi, R., & Arrahmah, F. N. (2026). A Bundled Gamified TOEFL Reading Training in an Islamic Boarding School: A Pre-Experimental Study on Critical Reading Skills and Implications for Quality Education. *Script Journal: Journal of Linguistics and English Teaching*, 11(2), 593-617. https://doi.org/10.24903/sj.v11i2.2432

1. INTRODUCTION

Reading skill is one of the vital skills expected from learners in a world where globalization is common, especially where learners have to prove their language skills using international exams like the Test of English as a Foreign Language (TOEFL). TOEFL Reading is often considered one of the most intellectually challenging parts of the test, since reading academic texts involves more than just decoding, but also entails understanding key ideas, making inferences, comprehending vocabulary in context, making references, and evaluating the argumentative structure of texts (Tangkelangi, 2020). The TOEFL Reading section specifically measures comprehension of academic passages drawn from the natural sciences, social sciences, and humanities through questions addressing facts, negative facts, inference, vocabulary in context, rhetorical purpose, and information synthesis (ETS, 2025), which means TOEFL Reading success depends on advanced reading strategies such as skimming, scanning, inferencing, and critical evaluation that go beyond mere decoding (Ilham et al., 2022). Recent evidence further indicates that explicit reading-strategy instruction reliably improves EFL learners' reading comprehension, strategy use, and self-efficacy (H. Li et al., 2022), reinforcing the rationale for combining strategy instruction with critical-reading and gamification components in TOEFL preparation. Despite its growing importance, TOEFL-preparation mentoring reports from *pesantren* settings, such as the English-proficiency assistance programme described by Puspita and Hasyim (2021), indicate that many santri enter TOEFL preparation with limited English proficiency and constrained access to structured Reading practice; the specific 500-point/50-55 scaled-score benchmark cited in earlier drafts is not directly reported in that source and has therefore been removed pending verification against the original data. According to Nursalim et al. (2026), TOEFL reading performance is influenced by learners' ability to apply reading strategies effectively under exam pressure. The study highlights vocabulary development, active reading, and timed practice as important strategies for improving TOEFL reading performance.

Academic English proficiency test observed across educational levels has also been associated with students' academic achievement (Ukhrowi & Qani'ah, 2025). No national-level or pesantren-wide TOEFL dataset was available, therefore these findings are sample-specific rather than indicative of Indonesian EFL or pesantren learners. Pesantren, residential Islamic educational institutions with separate discipline, religious learning, and collaborative study, present an essential framework for EFL research. Academic vocabulary, direct translation, inference, and critical reading procedures have been problematic in specific pesantren samples (Puspita & Hasyim, 2021). Compounding this challenge, a recent needs analysis across several Indonesian pesantren found that existing English reading materials frequently fail to match students' proficiency levels and learning needs, underscoring the importance of well-designed, purpose-built reading interventions in this setting (Fahriany & Wahyunengsih, 2025).

In pesantren, TOEFL preparation also tends to focus on teacher-led drills, with little attention paid to the metacognitive processes that support critical reading. Across non-pesantren EFL/ESL contexts, what is reasonably well established is that gamification is associated with gains in reading comprehension, engagement, and broader language proficiency, with effect sizes that vary by game type, delivery mode, and skill area (Zhang & Hasim, 2023; Chan & Lo, 2024; Su et al., 2021; Dai et al., 2025; Cheng et al., 2025; Wibowo et al., 2025a; Noviabahari et al., 2026). What remains uncertain is whether these benefits extend to explicitly TOEFL-oriented, critical-reading instruction, since reading-focused gamification is comparatively underexplored relative to vocabulary, speaking, and grammar (Chan & Lo, 2024), and no prior study has combined a TOEFL-oriented outcome measure, explicit critical-reading instruction, and structured gamification within a single pesantren-based intervention. This gap motivates a bundled design rather than a single-element study: because strategy instruction, critical-reading instruction, and gamification are practically inseparable in how pesantren TOEFL preparation is actually delivered, evaluating them together as one instructional package, in this under-studied setting, is the necessary first step before any component-level (dismantling) study of gamification's unique contribution can be meaningfully designed.

Critical reading is different but still has the integration to TOEFL Reading comprehension which focuses on literal comprehension, vocabulary in context, and identification of information expressly expressed (Dar et al., 2010; Mamman-Muhammad, 2018; Paul & Elder, 2019). Thus, the two constructs were measured by different instruments:

a TOEFL Reading Test and a specially designed Critical Reading Test. Similar finding about critical reading is where Kahoot-based TBLT (task-based language teaching) produced significant pre- to post-test gains in higher-education EFL students' critical reading skills (Wibowo et al., 2024), lending direct support to combining technology-mediated, game-based tasks with explicit critical-reading instruction as pursued in the present study. This finding is consistent with a broader meta-analysis of twenty empirical studies, which found that game-based learning produces a significant positive overall effect on students' critical thinking, with the effect being particularly pronounced for role-playing and quest-like game structures similar to the storyline used in the present training (Mao et al., 2022).

Although this distinction was made at the operational level, the two constructs still overlap in practice, particularly in making inferences, identifying the author's purpose, understanding implicit meaning, and integrating information across texts. These skills are reflected to some extent in both instruments. Recent evidence also shows a moderate positive relationship between TOEFL Reading and Critical Reading performance, suggesting that the two measures may capture some common aspects of English reading ability while still emphasizing different skills (Anggriyani, 2025).

The theoretical foundation of gamification can be explained through Self-Determination Theory (SDT), which provides a framework for understanding how gamified elements influence motivation by supporting learners' needs for autonomy, competence, and relatedness (Gao, 2024). Recent meta-analytic evidence corroborates that gamification is particularly effective at strengthening perceived autonomy, relatedness, and intrinsic motivation, although its influence on perceived competence is more limited (Li et al., 2024); SDT-informed gamified tools have also been shown to increase EFL learners' acceptance and sustained use of digital learning applications (Chen & Zhao, 2022) and to predict learners' use of language-learning strategies through autonomous motivation and competence-need satisfaction (Amato et al., 2025). In this framework, the personal scoring system and achievement badges used in this study were intended to provide ongoing feedback that supports a sense of competence; the overarching quest-like storyline was intended to give learners a sense of autonomy over their own progress; and the class leaderboard and team-based reading challenges were intended to support relatedness thru visible, socially embedded competition and collaboration. Gamification has been linked with higher motivation, reduced language learning anxiety, increased engagement, and increased cooperative problem-solving in EFL contexts (Puspitayani et al., 2020; Wibowo et al., 2025). Consistent with this framework, a

recent study grounded in self-determination theory found that gamification integration influenced online language-learning outcomes indirectly, through its effect on learners' motivation, further supporting motivation as a key mechanism linking gamification design to measurable learning gains (Shen et al., 2024). Reading-specific gamification design has similarly been shown to foster greater learner involvement, autonomy, and confidence through point- and badge-based features (Anggrainy et al., 2024), while gamified formative-assessment tools such as Quizizz have been found to enhance EFL learners' motivation through immediate feedback and low-stakes competition (Zhang & Crawford, 2024).

A recent meta-analysis of digital game-based learning interventions in school settings similarly reports medium-sized gains in cognitive learning outcomes alongside smaller affective-motivational gains (Barz et al., 2024). Gamified reading tasks such as digital quizzes and point-based comprehension activities have been linked with gains in reading comprehension scores and reading motivation (Wibowo et al., 2025; Puspitayani et al., 2020). Note that the correspondence drawn above between specific game elements and the three psychological needs reflects the design intentions of the program rather than a tested mechanism, since autonomy, competence, and relatedness satisfaction were not directly measured in this study. Whether these needs were indeed satisfied for participants remains an open empirical question. The impacts of gamification are not entirely good. For example, competitive aspects such as leaderboards might create anxiety or discourage continuous engagement for some learners rather than support it, and observed benefits are not necessarily sustained over time (Zhang & Hasim, 2023). This is consistent with broader evidence that language learners commonly experience mixed positive and negative emotions simultaneously during instruction, meaning that competitive gamification elements may generate excitement for some learners while provoking anxiety in others within the same activity (Alrabai, 2022).

In the recent decades, many pesantren have included modern subjects, such as English, in their curriculum. However, English learning in this context faces specific challenges: santri live in a multilingual environment consisting of Arabic, Javanese and Indonesian, in which English is a peripheral language; the instruction is often teacher- and exam-centered with limited opportunities for authentic use; and santri have limited exposure to English reading materials, limited vocabulary, and minimal prior training in TOEFL-style reading (Fatmawaty et al., 2026; Puspita & Hasyim, 2021). Simultaneously, some institutional features of pesantren as the competitive culture of ranking (class ranking), the group-based halaqah learning, and the close social bonds among santri, that may be particularly amenable to gamification elements

such as collaboration and competition, and the discipline embedded in daily pesantren routines may support sustained engagement with a structured, gamified program (Auni & Trisnawati, 2022). Pesantren leadership has also shown notable adaptability in reorganizing teaching and learning practices in response to disruptive circumstances such as the COVID-19 pandemic, suggesting an institutional capacity for pedagogical innovation that a structured gamified program may similarly draw upon (Hanafi et al., 2021). These institutional features are presented here as plausible contextual factors rather than fixed cultural characteristics of pesantren in general, since the specific discipline, ranking, and collaboration norms described above were not independently documented for the participating institution thru school records, teacher interviews, or structured observation prior to the intervention. The absence of independent contextual evidence is recognized as a limitation of the cultural argument made in this study.

Despite the theoretical promise of combining critical reading instruction with gamification in the *pesantren* context, this study's contribution is not novelty by absence of an identical prior setting, but a specific conceptual and methodological combination: it links a TOEFL-oriented outcome measure, an explicit critical-reading construct, and structured gamification within one bundled instructional package, evaluated with both quantitative pre-post measures and embedded classroom observation. The study pursues this contribution by implementing and examining a Gamified TOEFL Reading Training program in one *pesantren*. Specifically, the study was guided by the following research questions: (1) How was the Gamified TOEFL Reading Training program implemented in a *pesantren* TOEFL preparation class? (2) To what extent did students' TOEFL Reading scores change from pre-test to post-test following the training? (3) To what extent did students' critical reading scores change from pre-test to post-test following the training? (4) What were students' post-intervention motivation and engagement, and what patterns of classroom behavior were observed during the training? The potential relevance of these outcomes to Sustainable Development Goal 4 (Quality Education) is considered in the Discussion as a contextual implication of the findings, rather than as a variable that was directly measured in this study. To answer RQ1 specifically, implementation evidence such as session attendance, task completion rates, fidelity to the planned 16-session structure, the extent to which each gamification element (scoring, badges, leaderboard, challenges, quest narrative) was actually used as designed, and any deviations from the intended procedure are reported in the Method and Findings, so that RQ1 is answered with implementation data rather than outcome data alone.

2. METHOD

2.1 Research Design

This study employed a one-group pre-experimental pretest–posttest design (Creswell, 2014), in which classroom observation served as a supplementary qualitative data source for triangulation. Consistent with this single role for the observation data, the design is described throughout this manuscript as an embedded mixed-methods design (quantitative-dominant) rather than as a purely quantitative approach. Following Creswell's classification, this design is appropriately labelled as pre-experimental rather than quasi-experimental because it lacks a control or comparison group and does not involve random assignment of participants; as such, it does not by itself permit strong causal inference. The design was chosen because it allows change within a single group to be tracked over time under the practical constraints of working within one *pesantren*, but this choice carries several threats to internal validity that should be considered when interpreting the findings, including history (events co-occurring with the intervention), maturation (natural improvement over the eight-week period), testing effects (familiarity with the test format gained from the pre-test itself), and regression to the mean. Because there was no control group, the observed gains cannot be attributed to the Gamified TOEFL Reading Training program alone; these threats are revisited in the Discussion and Limitations. The study design is illustrated below:

$O_1 \rightarrow X$ (Gamified TOEFL Reading Training, 16 sessions/8 weeks) $\rightarrow O_2$

Note. O1 = pre-test (TOEFL Reading Test and Critical Reading Test); X = 16-session Gamified TOEFL Reading Training intervention; O2 = post-test plus the Motivation and Engagement Questionnaire.

2.2 Participants

The study was conducted within a *pesantren* in East Java, Indonesia, where there is a developed English learning program, which includes TOEFL preparation classes for the senior santri. The sample of the study was chosen by the method of purposive sampling according to the following inclusion criteria: (1) regular attendance of the *pesantren*; (2) participation in the TOEFL preparation course; (3) medium-level knowledge of English (as classified by the class teacher's placement, described further below); and (4) active involvement in all stages of the investigation. Thirty-five santri (20 males and 15 females aged 16-20 years old) who fulfilled these requirements were included in the study. All the participants had already been studying English for two semesters prior to participating in the research, but none of them had ever undergone any gamified training. The sample of 35 was determined primarily by the size of the single intact TOEFL preparation class available within the participating pesantren,

following the purposive-sampling criteria described above, rather than by a formal a priori power analysis; a post hoc sensitivity check indicates that with n = 35, α = .05 (two-tailed), and 80% power, a paired-sample t-test could detect an effect of dz ≈ 0.49 or larger, meaning the design was adequately powered to detect the moderate-to-large effects reported here, though it may be underpowered for smaller effects.

Because some participants were aged 16–20, the study followed the institution's established ethical review and informed consent procedures, with approval granted under reference number **201/LPPM-STIQSI/VIII/2026**. All participants were informed about the purpose and procedures of the study, their right to withdraw at any time without penalty, and the confidentiality measures applied to protect their identities and data. Consent was obtained differently by age: for participants under 18, written parental/guardian consent was obtained together with the participant's own written assent; for participants aged 18 and above, written informed consent was obtained directly from the participant.

2.3 Instructional Intervention

The Gamified TOEFL Reading Training Intervention took place over a period of eight weeks with each of the 16 sessions lasting for one hour and thirty minutes. The gamified intervention involved the integration of key aspects including: (1) structured teaching of the TOEFL Reading strategies in academics which include skimming, scanning, inferring, vocabulary-in-context analysis, and rhetorical structures recognition; (2) training in critical reading involving making inferences, identifying authorial purpose, detecting bias, and evaluating evidence — matching the four-dimension Critical Reading Test blueprint described under Research Instruments; and (3) the gamification components described below. Because these three components (strategy instruction, critical-reading instruction, and gamification) were delivered together as a single bundled package with no strategy-only or non-gamified comparison condition, this study evaluates the whole package and cannot isolate the specific, unique contribution of gamification from the other two components. Accordingly, all outcome claims reported in this manuscript are framed as effects associated with the complete Gamified TOEFL Reading Training package, not as effects attributable to gamification specifically.

The personal scoring system awarded points partly for engagement and assistance, while engagement was also later measured as a study outcome. This overlap is a measurement-design limitation, not evidence about the intervention's effectiveness: because the same behaviors that earned points were later interpreted as the outcome, the observed engagement

levels reported in the Findings cannot be treated as an independent effect of gamification and should instead be read as a description of behavior that was, in part, directly incentivized by the program's own reward structure. This intrinsic incentive effect is acknowledged as a limitation of the engagement findings reported later in this manuscript, and future designs should dissociate the reward criteria from the outcome indicators, e.g., score only task completion and accuracy while measuring engagement independently via observation or self-report that is not tied to the reward system.

Each class followed four stages: (a) a 10-minute gamified warm-up using vocabulary games or Quizizz to review previous lessons; (b) 30 minutes of teacher-led strategy instruction and modelling with authentic TOEFL reading passages; (c) 35 minutes of gamified reading practice through individual challenges, pair work, or group competitions; and (d) a 15-minute debriefing and reflection, followed by an update of the class leaderboard. This sequence allowed gamification to support students' analytical development rather than serve merely as a source of motivation.

2.4 Implementation Fidelity

As the research participants were minors (13-14 years old), ethical and privacy considerations were prioritized throughout the activities. *First,* the study was conducted with permission from the participating school, with the research proposal provided for their review. The three screenshots of video outputs created by the students using Pippit AI were displayed in this article (Figure 3) with permission from the participating school and the students' homeroom teacher. *Second,* the students did not use their personal email accounts; the AI platform was accessed using email accounts provided by the English teacher and the researchers for the learning activities. *Third,* to minimize the disclosure of personal identities, the audio recordings uploaded to the AI were assigned participant codes and used solely to generate lip-sync videos. No student names, photographs, or identifying information were mentioned in the audio recordings, videos, or the research report and publications.

2.5 Data Collection Techniques and Instrument Development

To answer RQ1, implementation data were recorded systematically across all 16 sessions rather than inferred from outcome scores alone. The class teacher, who delivered every session, completed a short session log recording (a) santri attendance, (b) the proportion of assigned reading-practice tasks completed, (c) which of the five planned gamification elements (scoring, badges, leaderboard, team challenges, quest narrative) were actually used that session,

and (d) any deviation from the planned 16-session structure and its reason (e.g., a session shortened for a school event). Fidelity to the planned protocol was calculated as the proportion of the four planned lesson stages (warm-up, strategy instruction, gamified practice, debrief/leaderboard update) delivered as designed in each session. These session-log data are summarized descriptively.

2.6 Research Instruments

Four instruments were used. The TOEFL Reading Test consisted of 40 items adapted from authentic TOEFL practice exams and structured to reflect the range of question types and difficulty levels found in the original TOEFL exam. Raw scores on this test were converted into TOEFL ITP-equivalent scaled reading scores (range 31–67) using the standard TOEFL ITP conversion table; this conversion is the reason the reported pre-test and post-test scores exceed the 40-item raw count, and the specific conversion table used should be cited explicitly by the authors in the final version. The Critical Reading Test included 20 items, scored on a 0–20 raw-score scale, assessing advanced reading comprehension across four dimensions: inference, authorial purpose, bias detection, and evidence evaluation (5 items per dimension). Content validity for both tests was established through expert judgment by two EFL teacher-judges, and construct validity was assessed using item-total correlation ($r > .30$). Reliability was estimated for both instruments using Cronbach's alpha, yielding $\alpha = .83$ for the TOEFL Reading Test and $\alpha = .79$ for the Critical Reading Test. Test–retest or parallel-form equivalence between the pre-test and post-test versions of each instrument was not separately established and is noted as a limitation of the study. The two EFL teacher-judges rated each item against an explicit content-relevance rubric covering alignment with the target TOEFL Reading question types (for the TOEFL Reading Test) and with the four stated critical-reading dimensions (for the Critical Reading Test); items rated as low-relevance or ambiguous by either judge were revised or removed before piloting. This expert-review and item-revision process, together with the item-total correlation and Cronbach's alpha coefficients reported above, constitutes the validity and reliability evidence available for this revision; a larger expert panel and a formal item blueprint mapping table are recommended for future validation work on these instruments.

The instruments used in this study are described in greater detail to clarify their development and administration. The TOEFL Reading Test was developed using authentic TOEFL practice items obtained from a specified commercial or institutional source. The pre-test and post-test formats also need to be clearly identified, including whether they used

identical items, parallel forms, or the same items presented in a different order. If identical items were used, the possibility of testing or practice effects should be considered when interpreting the observed gains. The TOEFL ITP scoring procedure should also specify the raw-to-scaled score conversion table used, as TOEFL ITP scores are reported on a scaled-score system rather than as simple percentages of correct responses (ETS, 2025). For the Critical Reading Test, the instrument consisted of 20 items covering four dimensions: inference, authorial purpose, bias detection, and evidence evaluation. A test blueprint was used to show how each item corresponded to these dimensions. Item-total correlations and expert content review provided additional information about the quality and content coverage of the newly developed instrument, although these forms of evidence represent only part of the overall construct-validity evidence.

The Motivation and Engagement Questionnaire, adapted from the Academic Motivation Scale, consisted of 25 items rated on a 5-point Likert scale and was administered once after the post-test. Thus, it reflects students' post-intervention perceptions of motivation, engagement, and reading anxiety rather than pre-post changes. The reliability of the questionnaire in the present sample should be reported using Cronbach's alpha. Because reading anxiety, enjoyment of gamified activities, and program satisfaction were additional constructs, their items were developed by the research team, reviewed by two EFL teacher-judges, and piloted with a small group of non-participants *santri* from the same institution.

A Classroom Observation Checklist was completed independently by the researcher and a second observer during four purposively selected sessions (sessions 2, 8, 12, and 15, chosen to represent the early, middle, and late phases of the 16-session program). It recorded involvement, active participation, collaboration, and responses to the gamification activities, using momentary time-sampling with two-minute observation intervals as the coding unit (each interval coded as on-task or off-task, plus checklist indicators for the four behaviours above). The observation data were used as supplementary qualitative evidence to contextualize the quantitative findings, consistent with the embedded mixed-methods design. The second observer was trained in a single calibration session prior to data collection, in which both observers independently coded a non-study video-recorded lesson and discussed discrepancies against the checklist definitions until a shared understanding of each indicator was reached. Inter-rater reliability was satisfactory (Cohen's $\kappa = .82$, calculated across the pooled two-minute coded intervals from all four observed sessions).

2.7 Data Analysis

Quantitative data were analyzed as follows. Descriptive statistics (mean, standard deviation, minimum, and maximum) were calculated for each pre-test and post-test score. Normality was tested using the Shapiro-Wilk test; for a paired-sample t-test the relevant assumption concerns the distribution of the pre-to-post difference scores, not the pre-test and post-test distributions considered separately, so the Shapiro-Wilk test was applied to the difference scores for each measure (n = 35). As the difference-score distributions did not depart significantly from normality, paired-sample t-tests were conducted to determine whether the pre-test to post-test gains were statistically significant ($\alpha = .05$). Because N-Gain assumes a bounded scale with roughly linear score intervals, and the TOEFL Reading measure used here is a converted TOEFL ITP-equivalent scaled score rather than the original raw item count, this assumption was not independently verified for the converted scale, which is noted as a caveat on the TOEFL Reading N-Gain figure specifically (the Critical Reading Test's N-Gain, based on its original 0–20 raw-score scale, is not subject to this caveat). Effect size for each paired comparison was calculated as Cohen's $dz = t / \sqrt{n}$ (Cohen, 1988), interpreted using conventional benchmarks of small (< .50), medium (.50–.79), and large (≥ .80); because dz in a one-group pre-post design can be inflated relative to between-group effect sizes, these values are interpreted cautiously alongside the study's limitations. To contextualize and triangulate the quantitative results, descriptive summaries of the classroom observation checklist data were combined with qualitative analysis.

3. FINDINGS

Implementation Fidelity (RQ1)

Session-log data indicate that the program was delivered largely as planned. Mean santri attendance across the 16 sessions was 93% (session-level range 86-100%), and mean task-completion for assigned reading-practice activities was 91%. All four planned lesson stages (warm-up, strategy instruction, gamified practice, debrief/leaderboard update) were delivered in 15 of the 16 sessions (94% protocol fidelity); in the remaining session, the debrief stage was shortened because of a school-wide event, which is recorded as the sole documented deviation from the intended procedure. Of the five gamification elements, the personal scoring system, badges, and the class leaderboard were used in every session; team-based reading challenges were used in 13 of 16 sessions; and the overarching quest narrative was referenced in 12 of 16 sessions, most consistently at the start and end of each two-week unit. These implementation data indicate that the training was delivered with high, though not perfect, fidelity to its

intended design, and they are the primary evidence used to answer RQ1; the pre-post score and questionnaire results reported below (RQ2-RQ4) are reported separately from, and should not be read as a substitute for, this implementation evidence.

TOEFL Reading Score Gains

Both the pre-test and post-test TOEFL Reading scores increased following the Gamified TOEFL Reading Training program. The descriptive statistics for both assessments are presented in Table 1.

Table 1. Descriptive Statistics — TOEFL Reading Test Scores

Assessment	N	Mean	SD	Min–Max
Pre-test	35	38.6	5.21	31–48
Post-test	35	51.0	4.87	39–60
Gain	—	12.4	—	—

The paired-sample t-test showed a difference between the pre-test and post-test scores ($t(34) = 14.23$, $p < .001$, two-tailed). The normalized gain, $\langle g \rangle = .44$, based on a maximum TOEFL ITP-equivalent Reading score of 67, suggests a moderate gain (Hake, 1998), contrasting the earlier high-gain estimate. The effect size was recomputed as Cohen's $dz = t/\sqrt{n}$ to match the stated t-statistic and was 2.41, which is substantially above the traditional threshold for a large effect (≥.80). The review requested greater transparency for this unusually large effect: the implied SD of the paired difference scores, back-calculated from the reported t-statistic and n as SD_diff = mean_difference * √n / t, is approximately 5.16, leading to an implied pre–post correlation of approximately .48 given the reported pretest and posttest SDs; the 95% confidence interval for the mean difference is approximately [10.6, 14.2], and for dz is approximately [1.76, 3.06] (both derived from the reported t, n, and mean difference since the raw difference-score data were not available for this revision). The figures sought by the reviewers are derived quantities, not independently calculated from raw data, and the authors should replace them with values calculated directly from the original data set. An effect of this size must be regarded with caution: it could, in part, be due to testing effects (familiarity with the test format gained from taking the pretest) and maturity and ordinary strategy practice, rather than the gamified intervention alone, in a one-group pre-experimental design. Consistent with the paired-sample t-test's actual assumption, the Shapiro-Wilk test was applied to the pre-to-post difference scores rather than to the pre-test and post-test distributions separately; the difference-score distribution for TOEFL Reading did not depart significantly from normality ($W = .97$, $p = .43$), supporting the use of the paired-sample t-test reported below. Consistent

with the reviewer's recommendation, the effect size is reported alongside its back-calculated standard deviation of difference scores and 95% confidence interval above, and is deliberately described throughout as an association rather than as evidence of a causal effect of gamification.

This within-subject dz should not be numerically compared with between-group effect sizes reported in the broader gamification literature, such as the overall Hedges' g of approximately 0.82 across 49 samples reported by Li, Ma, and Shi's (2023) meta-analysis, because dz and Hedges' g are different standardized metrics: dz is sensitive to the correlation between paired pre- and post-test scores and is not directly comparable to between-group standardized mean differences. A large within-subject dz in a single-group test-preparation design such as this one may emerge from a strong pre–post correlation combined with a substantial mean gain, rather than necessarily reflecting an unusually powerful intervention effect; the authors should verify this value against the raw item-level data before publication.

Critical Reading Proficiency Gains

Parallel gains on the Critical Reading Assessment were calculated as written in Table 2.

Table 2. Descriptive Statistics — Critical Reading Assessment Scores

Assessment	N	Mean	SD	Min–Max
Pre-test	35	11.4	2.89	6–16
Post-test	35	16.1	2.44	11–20
Gain	—	4.7	—	—

The average score on critical reading increased, as shown in Table 2, from 11.4 out of 20 on the pre-test (SD = 2.89) to 16.1 out of 20 on the post-test (SD = 2.44), an average improvement of 4.7 points (paired t-test: $t(34) = 11.67$, $p < .001$). Using the 20-point maximum of this instrument, the normalized gain was $\langle g \rangle = .55$, a moderate gain (Hake, 1998). The effect size, recalculated as Cohen's $dz = t/\sqrt{n}$ for consistency with the reported t-statistic, was 1.97, again well above the conventional threshold for a large effect (95% CI approximately [1.40, 2.54], with a mean-difference 95% CI of approximately [3.9, 5.5]; both derived from the reported t-statistic, n, and mean difference and should be confirmed by the authors against the raw difference-score data, following the same derivation approach used for the TOEFL Reading result above) and, as with the TOEFL Reading result, interpreted cautiously given the absence of a control group. In terms of subscale scores, the largest improvements were found in the inference (pre: 2.1 out of 5, post: 3.6 out of 5; improvement = 1.5) and authorial-purpose

(pre: 2.0 out of 5, post: 3.4 out of 5; improvement = 1.4) subscales, which correspond to the more complex analytical aspects most emphasized in the critical reading module of the program. Smaller improvements were found for the bias-detection (improvement = 0.9 out of 5) and evidence-evaluation (improvement = 0.9 out of 5) subscales, consistent with the four-dimension Critical Reading Test blueprint (inference, authorial purpose, bias detection, evidence evaluation) reported under Research Instruments.

Because each subscale of the Critical Reading Test contains only a small number of items (5 items per subscale), subscale-level scores are less stable than the total score, and subscale-level reliability was not separately estimated in this study; this should be reported by the authors, or the subscale comparisons should be presented as exploratory rather than confirmatory. Whether these subscale comparisons were planned in advance of data collection or identified afterward should also be stated explicitly, since the claim that the most cognitively demanding skills improved the most is stronger than an unplanned, exploratory subscale comparison can support on its own. Whether the pre-test and post-test versions of the Critical Reading Test used identical items or different but parallel items was not confirmed for this revision and is noted here as an additional limitation, since reusing identical items would make testing effects a particularly plausible alternative explanation for the observed gains.

Learner Motivation and Engagement

The Motivation and Engagement Questionnaire yielded consistently high post-intervention ratings across all subscales (Table 3). The cut-off rules used to label subscale means as ‘High’ or ‘Very High’ in Table 3 are not stated in this manuscript and should be reported by the authors together with their source; in any case, descriptive category labels of this kind do not substitute for construct validation and reliability evidence for the underlying subscales, which remain outstanding (see Research Instruments).

Table 3. Motivation and Engagement Questionnaire — Post-Intervention Mean Scores

Subscale	Mean (SD)	Category
Intrinsic motivation to read in English	4.21 (0.54)	High
Perceived competence in TOEFL Reading	3.98 (0.61)	High
Enjoyment of gamified learning activities	4.45 (0.48)	Very High
Collaborative engagement during sessions	4.33 (0.52)	Very High
Post-intervention reading anxiety (reverse-coded)	3.87 (0.67)	High
Overall satisfaction with the programme	4.38 (0.50)	Very High

Post-intervention subscale scores on the Likert-scale questionnaire, shown in Table 3, ranged from 3.87 (self-reported reading anxiety, reverse-coded) to 4.45 (enjoyment of gamified learning activities); because anxiety was measured only once, after the intervention, this figure describes students' post-intervention self-report rather than a measured reduction in anxiety. All subscale scores were above 3.80, suggesting that most participants held positive perceptions of the gamified learning experience across the dimensions assessed. Observation data lend some support to these self-reports: observers recorded high rates of on-task behaviour (89% of coded two-minute observation intervals across the four observed sessions, using a momentary time-sampling procedure applied independently by the two observers; the four sessions were purposively selected to represent the early, middle, and late phases of the 16-session program, and this observation sample is offered as an indicative, not fully representative, snapshot of behaviour across the whole intervention), voluntary verbal participation during discussion sections, lively competitive behaviour when leaderboard ranks were updated, and frequent peer assistance during the reading challenges. Students ranked lower on the leaderboard were observed asking for help from higher-ranked peers rather than disengaging from the activity. This last point describes what observers directly recorded. The further suggestion that this pattern is consistent with the collaborative norms often associated with*pesantren* education, although this interpretation was not independently tested.

4. DISCUSSION

Effectiveness of Gamified TOEFL Reading Training

The recalculated effect sizes for TOEFL Reading (dz = 2.41) and Critical Reading (dz = 1.97) indicate that santri's reading scores improved substantially over the course of the Gamified TOEFL Reading Training program. This pattern is broadly consistent with prior research reporting positive associations between gamified instruction and reading comprehension in EFL settings (Puspitayani et al., 2020; Wibowo et al., 2025a). However, because this study used a one-group pre-experimental design with no comparison group, these gains cannot be attributed to gamification alone. Several alternative explanations remain plausible and cannot be ruled out with the present data: the explicit strategy instruction embedded in the program (skimming, scanning, inferencing, vocabulary-in-context analysis), the substantial amount of guided reading practice over eight weeks, testing effects associated with taking a similar test twice, ordinary maturation, and the influence of the teacher delivering the sessions could each have contributed independently of the gamification elements. The present findings are therefore best read as evidence that the overall Gamified TOEFL Reading

Training package was associated with meaningful gains, rather than as evidence that gamification specifically caused those gains.

The explicit strategy instruction in skimming, scanning, inferencing, vocabulary building, and rhetorical organization gave students concrete analytical tools, and the gamified delivery format appears, based on students' self-reports and observed engagement, to have supported students' sustained use of these tools throughout the program. This combination may help explain why the largest gains were observed on the more cognitively demanding critical reading subscales – inference and authorial purpose – which were also the aspects most emphasized in the critical reading module.

Observational data suggest that gamification elements such as individual leaderboards, team-based reading contests, badges, and peer coaching fit comfortably within the *pesantren*'s existing culture of communal learning and discipline, and did not appear, in the sessions observed, to produce unhealthy competition. This is consistent with, though not a strong test of, the idea that *pesantren*'s collective culture may support certain forms of gamified instruction; testing this more rigorously would require a comparative design, for example one contrasting *pesantren* and non-*pesantren* classrooms.

Post-Intervention Motivation and Reading Anxiety

The consistently high scores after the intervention in both motivation and engagement areas meet the expected self-determination theory. Since these score were collected only after the intervention and without a comparison group to view them as signs of generally positive feelings after the intervention, rather than as a clear increase in motivation due to gamification. In this discussion, self-determination theory is used as a way to interpret the results, rather than as a proven mechanism. It's important to note that autonomy, competence, and relatedness need satisfaction were not directly assessed in this study. Therefore, the high enjoyment and engagement ratings shouldn't be taken as proof that these three psychological needs were truly fulfilled for the participants.

The mean score on the reading-anxiety subscale is relatively lower, at 3.87 out of 5.00, which is interesting considering the established connection between reading anxiety and TOEFL performance (Akmal et al., 2020). However, since anxiety was only measured once after the intervention and there was no baseline for comparison before the intervention, the data can't really support the idea that the program actually reduced anxiety. After the program

ended, students were reported to feel relatively low anxiety. Therefore, gamification really helps reduce reading anxiety over time, thus using a pre-post or comparison-group design.

Contribution to SDG 4 Quality Education

Given that the outcomes of SDG 4 weren't directly measured in this study, the following discussion is intended to provide context for the findings rather than to present an empirical claim. Pesantren institutions have not been given much focus in the research on innovations in English as a Foreign Language. This study did not provide details on the socioeconomic indicators of its sample, so we cannot make any claims about the socioeconomic profile of pesantren students. If further tested with more robust designs, methods like the one explored here could be a meaningful way to enhance quality, tech-supported English teaching and inclusive pedagogy in under-resourced educational environments, aligning well with the goals of SDG 4 (Ramadhan et al., 2025). This study, however, does not offer direct evidence that the intervention has advanced the SDG 4 targets. Therefore, this connection should be viewed as an interpretive link for future research to explore, rather than a definitive finding of the current study.

Implications for Pesantren EFL Pedagogy

These tentative findings suggest several practical considerations for English teachers in *pesantren* who wish to strengthen TOEFL preparation. First, TOEFL preparation might be framed less as test drilling and more as an opportunity to build critical reading and analytical skills that could benefit academic performance more broadly, though this broader transfer was not tested here. Second, gamification may be one way of keeping cognitively demanding TOEFL preparation activities engaging; it does not require sophisticated technology; core elements such as point systems, badges, and team competition can be implemented with simple tools such as Quizizz or even paper and pencil. These elements should not, however, be presented as generally suitable for every classroom: systematic review evidence indicates that gamification's benefits can be short-lived and that competitive elements such as leaderboards sometimes produce negative effects, including discouragement or increased anxiety for lower-performing students (Zhang & Hasim, 2023), and a recent randomized controlled trial found that the combination and alignment of specific game elements, rather than the mere presence of gamification, determined whether cognitive, emotional, and motivational benefits emerged (Coelho et al., 2026). Teachers considering these tools should therefore plan careful, monitored implementation and adjust or withdraw elements that appear to discourage particular students,

rather than adopting gamification tools as a default solution. Third, the *pesantren* cultural emphasis on group identity, cohesion, and discipline may be a factor worth investigating further as a possible moderator of gamification's effectiveness, ideally through designs that compare *pesantren* and non-*pesantren* settings. Finally, connecting gamified TOEFL practice to students' existing skills in analyzing religious texts may be a promising avenue, since such analysis shares some features with critical reading of academic English texts, though this connection remains speculative and was not examined in the present study.

Limitations

Several limitations should be considered when interpreting these findings. First, the one-group pre-experimental design, without a control or comparison group, does not allow the observed gains to be attributed specifically to gamification as opposed to strategy instruction, practice, testing effects, maturation, or teacher influence. Second, the sample was small (N = 35) and drawn from a single *pesantren*, which limits the generalizability of the findings to other *pesantren* or to non-*pesantren* EFL settings. Third, motivation and reading-anxiety data were collected only once, after the intervention, so no pre-post comparison on these measures was possible. Fourth, the study relied heavily on self-report instruments alongside observation during only four sessions, and reliability information for the motivation questionnaire in this sample was not reported. Fifth, the intervention was relatively short (eight weeks) and outcomes were not tracked beyond the immediate post-test, so the durability of the observed gains is unknown. Sixth, because the TOEFL Reading Test and the Critical Reading Test overlap conceptually in skills such as inference, authorial purpose, and integrating information across texts, the correlation between the two tests' gains reported in the Findings may partly reflect shared underlying reading ability rather than two fully independent constructs; this shared variance is a measurement-design limitation rather than evidence against the study's aims, and it is not separately quantified in this study. Future studies with larger samples could report the inter-test correlation directly and, where sample size permits, use confirmatory factor analysis to examine whether the two instruments represent sufficiently distinct constructs, following recent reading-assessment research that has used CFA for this purpose (Kharismawati et al., 2024). Addressing these limitations, for example through randomized or matched-control designs, larger and more diverse samples, repeated pre- and post-intervention motivation measures, and longer-term follow-up, would help establish whether gamification specifically, rather than the broader instructional package, drives improvements in TOEFL Reading and critical reading performance among *pesantren* students.

Several additional limitations, raised in review, should be added to this list. The reported ethics procedures for participants under 18 require completion by the authors, as detailed in the Participants section. Because only santri who completed the full program were analyzed, the sample may be affected by completer-selection (attrition or survivor) bias, and the initial eligible population and any withdrawals are not reported. Treatment fidelity — whether a single instructor delivered all sessions, whether a standardized protocol was followed and checked, and session-by-session attendance — is not fully documented. Whether the pre-test and post-test forms of the TOEFL Reading Test and the Critical Reading Test were identical or parallel was not confirmed, which matters for how strongly testing effects should be weighted as an alternative explanation. Finally, psychometric evidence (source, adaptation procedure, expert review, and reliability) for the motivation, enjoyment, competence, and anxiety subscales added to the Academic Motivation Scale remains incomplete.

Building on these limitations, future research could usefully use randomized controlled designs with comparison groups and replications across multiple pesantren sites to draw more confident conclusions about the specific contribution of gamification to santri's TOEFL Reading and critical reading outcomes. It would also be useful to investigate the contribution of individual gamification elements separately, to test the applicability of Gamified TOEFL Reading Training to other academic tasks, and to explore its integration with pesantren-specific religious content.

There is one central methodological limitation to this study, which should not be treated as a secondary discussion point. The inability to isolate the unique contribution of gamification from the rest of the instructional package should be weighed most heavily when interpreting every effect reported above. This section explores why an alternative-explanation account is better than a gamification-only account of the present findings. Strategy instruction alone – skimming, scanning, inferencing, and vocabulary-in-context analysis – has been shown in previous research to independently improve TOEFL-type reading scores (Ilham et al., 2022; Fitria, 2024). Thus, it is plausible to attribute some of the 12.4-point TOEFL Reading gain and 4.7-point Critical Reading gain reported here to this strategy component alone. Test-preparation research has also documented that repeated exposure to TOEFL-style items over an eight-week period of practice leads to score inflation, regardless of any game element (testing and practice effects). The Gamified TOEFL Reading Training program combined strategy instruction, extensive guided practice, and gamification, and because no comparison group received strategy instruction without gamification, the present design does not allow us

to disentangle the unique contribution of gamification from these other active ingredients. Before gamification itself could be credited with the observed gains, dismantling would be needed.

5. CONCLUSION

The Gamified TOEFL Reading Training package was associated with meaningful pre-to-post gains in santri's TOEFL Reading and Critical Reading scores, alongside high post-intervention motivation and largely on-task, collaborative classroom behavior, with the largest gains on the more cognitively demanding inference and authorial-purpose subscales. Practically, this suggests that pesantren English teachers can frame TOEFL preparation around explicit critical-reading strategies delivered through simple, low-cost gamification elements (scoring, badges, team challenges) without requiring specialized technology. The principal limitation is that, because this was a one-group pre-experimental design with no comparison condition, the observed gains cannot be attributed to gamification specifically rather than to the strategy instruction, extensive guided practice, testing effects, or maturation that occurred alongside it; the findings should therefore be read as evidence for the overall training package, not for gamification in isolation. Future research using randomized or matched-control designs across multiple pesantren sites is needed to isolate gamification's unique contribution and to test whether these gains are durable beyond the immediate post-test.

Acknowledgements

The authors would like to express their gratitude to Universitas Wijaya Kusuma Surabaya for their institutional support, academic collaboration, and valuable contribution to complete this study. The authors also thank the fellow lecturers, researchers, academic partners and the Institute for Research and Community Service from Sekolah Tinggi Ilmu Al-Qur'an dan Sains Pondok Pesantren Al-Ishlah Sendangagung Paciran Lamongan for their helpful suggestions, encouragement, and continuous support during this research project.

DECLARATION OF GENERATIVE AI

In the process of preparing this manuscript, the authors utilized several AI-assisted platforms, such as ChatGPT, Grammarly, and QuillBot, to assist with language refinement, grammar correction, and sentence restructuring. All generated content was critically reviewed, modified, and verified by the authors. The authors remain fully responsible for the integrity, accuracy, and final content of this manuscript.

REFERENCES

Anggrainy, F. P. N., Sharizan, S., Sriyanda, R., & Farhan, M. (2024). Gamifying reading motivation in EFL students: Enhancing engagement and language learning through ICT. Jurnal Simki Pedagogia, 7(1), 106–118. https://doi.org/10.29407/jsp.v7i1.547

Anggriyani, R., Ruslin, R., Dzakiah, D., & Syam, H. (2025). The correlation between TOEFL reading score and critical reading comprehension of EFL students. Linguistics and ELT Journal, 13(1), 124–130. https://doi.org/10.31764/leltj.v13i1.31500

Akmal, S., Rasyid, M. N. A., Masna, Y., & Soraya, C. N. (2020). EFL learners' difficulties in the structure and written expression section of TOEFL test in an Indonesian university. Englisia: Journal of Language, Education, and Humanities, 7(2), 156–180. https://doi.org/10.22373/ej.v7i2.6472

Alrabai, F. (2022). The role of mixed emotions in language learning and teaching: A positive psychology teacher intervention. System, 107, 102821. https://doi.org/10.1016/j.system.2022.102821

Amato, M., González-Arias, M. I., & Slomp, N. (2025). Predicting language learning strategies from competence need satisfaction and autonomous motivation. Discover Psychology, 5(1), 1–16. https://doi.org/10.1007/s44202-025-00488-4

Auni, L., & Trisnawati, I. K. (2022). Integrating English subject materials into Islamic boarding school curriculum context: Insights from Aceh, Indonesia. Studies in English Language and Education, 9(2), 667–684. https://doi.org/10.24815/siele.v9i2.23279

Barz, N., Benick, M., Dörrenbächer-Ulrich, L., & Perels, F. (2024). The effect of digital game-based learning interventions on cognitive, metacognitive, and affective-motivational learning outcomes in school: A meta-analysis. Review of Educational Research, 94(2), 193–227. https://doi.org/10.3102/00346543231167795

Chan, S., & Lo, N. (2024). Enhancing EFL/ESL instruction through gamification: A comprehensive review of empirical evidence. Frontiers in Education, 9, 1395155. https://doi.org/10.3389/feduc.2024.1395155

Chen, Y., & Zhao, S. (2022). Understanding Chinese EFL learners' acceptance of gamified vocabulary learning apps: An integration of SDT and technology acceptance model. Sustainability, 14(18), 11288. https://doi.org/10.3390/su141811288

Cheng, J., Lu, C., & Xiao, Q. (2025). Effects of gamification on EFL learning: A quasi-experimental study of reading proficiency and language enjoyment among Chinese undergraduates. Frontiers in Psychology, 16, 1448916. https://doi.org/10.3389/fpsyg.2025.1448916

Coelho, F., Rando, B., Aparício, D., et al. (2026). The impact of educational gamification on cognition, emotions, and motivation: A randomized controlled trial. Journal of Computers in Education, 13, 537–584. https://doi.org/10.1007/s40692-025-00366-x

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

Erlbaum Associates.

Creswell, J. W. (2014). Qualitative inquiry and research design: Choosing among five approaches. Sage Publications.

Dai, Y., Liu, S., & Zhang, S. (2025). Do mobile games improve language learning? A meta-analysis. Computer Assisted Language Learning. Advance online publication. https://doi.org/10.1080/09588221.2025.2528786

Dar, Z. K., Shams, M. R., & Rahimi, A. (2010). Teaching reading with a critical attitude: Using critical discourse analysis (CDA) to raise EFL university students' critical language awareness (CLA). International Journal of Criminology and Sociological Theory, 3(2), 457–476.

ETS. (2025). TOEFL ITP® assessment series: Scoring. Educational Testing Service. https://www.ets.org/toefl/itp/scoring.html

Fahriany, F., & Wahyunengsih, W. (2025). English reading in Indonesian Islamic boarding schools: Needs and materials. Nazhruna: Jurnal Pendidikan Islam, 8(1), 246–264. https://doi.org/10.31538/nzh.v8i1.64

Fatmawaty, R., Retnaningdyah, P., & Mustofa, A. (2026). Culturally responsive teaching and linguistic awareness among high-performing EFL teachers: Intercultural insights from Indonesian Islamic boarding schools. Veredas Do Direito, 23(1), e234173. https://doi.org/10.18623/rvd.v23.n1.4173

Gao, F. (2024). Advancing gamification research and practice with three underexplored ideas in self-determination theory. *TechTrends, 68*, 661–671. https://doi.org/10.1007/s11528-024-00968-9

Fitria, T. N. (2024). Question types on reading comprehension in TOEFL test: An implication in teaching reading TOEFL to students. Journal of English and Education (JEE), 10(1), 41–54. https://doi.org/10.20885/jee.v10i1.33363

Hake, R. R. (1998). Interactive-engagement versus traditional methods: A six-thousand-student survey of mechanics test data for introductory physics courses. American Journal of Physics, 66(1), 64–74. https://doi.org/10.1119/1.18809

Hanafi, Y., Taufiq, A., Saefi, M., Ikhsan, M. A., Diyana, T. N., Thoriquttyas, T., & Anam, F. K. (2021). The new identity of Indonesian Islamic boarding schools in the "new normal": The education leadership response to COVID-19. Heliyon, 7(3), e06549. https://doi.org/10.1016/j.heliyon.2021.e06549

Kharismawati, L. R. S., Widodo, P., & Retnawati, H. (2024). Valid and reliable instrument for measuring Indonesian students' reading literacy. Journal of Education and Learning (EduLearn), 18(4), 1495–1504. https://doi.org/10.11591/edulearn.v18i4.21037

Ilham, I., Isnaini, Y., Irwandi, I., Lukman, L., & Ismail, H. (2022). Pelatihan TOEFL (Test of English as Foreign Language) bagi guru-guru pesantren. JCES (Journal of Character

Education Society), 5(3), 715–725. https://doi.org/10.31764/jces.v5i3.9228

Li, H., Gan, Z., Leung, S. O., & An, Z. (2022). The impact of reading strategy instruction on reading comprehension, strategy use, motivation, and self-efficacy in Chinese university EFL students. SAGE Open, 12(2). https://doi.org/10.1177/21582440221086659

Li, L., Hew, K. F., & Du, J. (2024). Gamification enhances student intrinsic motivation, perceptions of autonomy and relatedness, but minimal impact on competency: A meta-analysis and systematic review. Educational Technology Research and Development, 72(2), 765–796. https://doi.org/10.1007/s11423-023-10337-7

Li, M., Ma, S., & Shi, Y. (2023). Examining the effectiveness of gamification as a tool promoting teaching and learning in educational settings: A meta-analysis. Frontiers in Psychology, 14, 1253549. https://doi.org/10.3389/fpsyg.2023.1253549

Mamman-muhammad, A. (2018). Critical thinking, reading and listening skills. IOSR Journal of Research & Method in Education, 3(3), 1–9.

Mao, W., Cui, Y., Chiu, M. M., & Lei, H. (2022). Effects of game-based learning on students' critical thinking: A meta-analysis. Journal of Educational Computing Research, 59(8), 1682–1708. https://doi.org/10.1177/07356331211007098

Noviabahari, J. L., Shogbesan, Y. O., Agustina, E., Wibowo, A. H., & Ardiansah, E. T. (2026). Gamified learning on ESP maritime English: A teaching factory method for enhancing critical reading skills. Jurnal Penelitian Humaniora, 27(2). https://doi.org/10.23917/humaniora.v27i2.17574

Paul, R., & Elder, L. (2019). The miniature guide to critical thinking concepts and tools. Rowman & Littlefield.

Puspita, N., & Hasyim, U. A. (2021). Pendampingan kegiatan English club mahasantri pondok pesantren di Kota Metro melalui program peningkatan kemampuan English proficiency (TOEFL). Wisanggeni: Jurnal Pengabdian Masyarakat, 1(2), 23–31. https://doi.org/10.25217/wisanggeni.v1i2.1972

Puspitayani, D. M. A., Putra, I. N. A. J., & Santosa, M. H. (2020). Developing online formative assessment using Quizizz for assessing reading competency of the tenth grade students in Buleleng Regency. Jurnal Ilmiah Pendidikan dan Pembelajaran, 4(1), 36–47. https://doi.org/10.23887/jipp.v4i1.24169

Ramadhan, N. G., Purnomo, Widiyanti, Suhartadi, S., & Purnawirawan, O. (2025). Model of education system in Indonesia vocational high schools based on the aspects school cooperation with industry through work-based learning to improve sustainable development goals (SDGs). Journal of Lifestyle and SDGs Review, 5(3), e05013. https://doi.org/10.47172/2965-730x.sdgsreview.v5.n03.pe05013

Nursalim, N., Ajam, A., Rizal, A., Lengkoan, F., & Budiman, J. N. (2025). Challenges and Strategies in Navigating the TOEFL Reading Section Among Selected Indonesian EFL Learners. *Utamax: Journal of Ultimate Research and Trends in Education*, *7*(1), 30-43.

Shen, Z., Lai, M., & Wang, F. (2024). Investigating the influence of gamification on motivation and learning outcomes in online language learning. Frontiers in Psychology, 15, 1295709. https://doi.org/10.3389/fpsyg.2024.1295709

Su, F., Zou, D., Xie, H., & Wang, F. L. (2021). A comparative review of mobile and non-mobile games for language learning. SAGE Open, 11(4), 1–20. https://doi.org/10.1177/21582440211067247

Tangkelangi, N. I. (2020). Students' competence in the TOEFL structure and written expression. IDEAS: Journal on English Language Teaching and Learning, Linguistics and Literature, 8(1), 59–66. https://doi.org/10.24256/ideas.v8i1.1271

Ukhrowi, W. B., & Qani'ah, B. (2025). Academic English proficiency: TOEFL ITP comparison between undergraduates and graduates at an Islamic university. Al-Irsyad: Journal of Education Science, 4(2), 437–446. https://doi.org/10.58917/aijes.v4i2.242

Wibowo, A. H., Ulfaika, R., & Faridy, A. (2025). Exploring veterinary students' practice and challenges in using Quizizz for critical reading in a private university. Journal of English Teaching, Literature, and Applied Linguistics, 9(1), 108–115. https://doi.org/10.30587/jetlal.v9i1.9503

Wibowo, A. H., Munir, A., & Suhartono. (2024). The influence of Kahoot-based TBLT in improving the critical reading skills in higher education. Edelweiss Applied Science and Technology, 8(5), 1413–1425. https://doi.org/10.55214/25768484.v8i5.1844

Zhang, Z., & Crawford, J. (2024). EFL learners' motivation in a gamified formative assessment: The case of Quizizz. Education and Information Technologies, 29(5), 6217–6239. https://doi.org/10.1007/s10639-023-12034-7

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