ETDC
Educational Technology in Developing Countries

Vol. 1, No.1 (June 2026), 23-42
P-ISSN: xxxx-xxxx | E-ISSN: xxxx-xxxx
DOI: https://doi.org/10.xxxxx/lms.vxxx.xx
https://ejournal.tuah.or.id/index.php/etdc/index

Gamified Literacy Applications and Reading Proficiency Among Struggling Primary Learners in Indonesia: A Cluster Randomized Controlled Trial

Arief Ardiansyah[1]*; Irhamuddin[2]; Odi Tondi Nasution[3]

[1] Universitas Islam Malang, Indonesia.
[2] Al-Azhar University, Cairo, Egypt.
[3] Universitas Islam Negeri Maulana Malik Ibrahim, Malang, Indonesia.

*Correspondence: arief.ardiansyah@unisma.ac.id

Submitted: 21-03-2026	Accepted: 10-05-2026	Published: 17-05-2026

Abstract. Foundational literacy among Indonesian primary learners remains critically low, with PISA 2022 reading scores far below OECD averages, yet rigorous evidence on gamified digital interventions in this context is sparse. This cluster-randomized controlled trial evaluated whether a structured, gamified literacy application (LiterasiKu) improved reading proficiency and for whom the effects were largest. We randomly assigned 24 schools (950 Grade 2–3 students) across three West Java districts to either 16 weeks of daily 30-minute LiterasiKu sessions integrated into regular instruction or standard Kurikulum Merdeka teaching. Three-wave Early Grade Reading Assessment data (baseline, midline, endline) were analyzed using hierarchical linear models with school random effects. The intervention produced medium-to-moderate effects on oral reading fluency (d = 0.58), phonemic awareness (d = 0.52), and reading comprehension (d = 0.47). Growth trajectory analyses showed sustained gains without attenuation across 16 weeks, countering novelty-driven explanations. Exploratory subgroup analyses revealed larger effects for lowest-quartile learners (d = 0.74 for fluency), suggesting equity potential, though regression-to-the-mean and floor effects remain alternative accounts. Implementation fidelity averaged 87.3% with modest teacher training. The findings demonstrate that culturally adapted gamified packages can deliver meaningful literacy gains within existing LMIC primary school timetables. However, without an active digital comparator, attribution to gamification mechanics remains unresolved. Future dismantling trials should isolate gamification from adaptive personalization components, followed by cost-effectiveness analyses before scaling decisions are made.

Keywords: gamification, foundational literacy, reading proficiency cluster randomized controlled trial, primary education.

INTRODUCTION

Foundational literacy, defined as the capacity to read with comprehension by the end of primary schooling, occupies a central position in international education policy as the prerequisite for all subsequent learning (Antoninis et al, 2023; World Bank, 2022). However, across low- and middle-income countries (LMICs), approximately 7 in 10 ten-year-olds cannot read and understand a simple, age-appropriate text, a condition the World Bank designates as learning poverty (Azevedo et al., 2021; World Bank, 2022). Indonesia illustrates this

challenge with particular sharpness. The Programme for International Student Assessment (PISA) 2022 reported a mean reading score of 359 for Indonesian 15-year-olds, extending a downward trajectory from 371 in 2018 and 397 in 2015, and falling well below the Organisation for Economic Co-operation and Development (OECD) averages of 482, 487, and 493 across the same cycles (OECD, 2023). Although Indonesia did not participate in the most recent Progress in International Reading Literacy Study (PIRLS) 2021, the 2011 cycle ranked the country 45th among 48 participants with a mean score of 428, substantially below the international centerpoint (Mullis et al., 2012). National assessment data from the Indonesian Ministry of Education place the literacy baseline at 68.13 percent against a target of 76.62 percent under the National Medium-Term Development Plan 2025–2029 (UNICEF Indonesia, 2025).

These deficits cluster among populations at risk, including children in rural and peri-urban areas, students from low socioeconomic backgrounds, and learners in under-resourced schools where pupil-to-teacher ratios exceed national norms (Hunaepi & Suharta, 2024; Beatty et al., 2021). Recent national survey data indicate that only approximately one-third of Indonesian elementary students engage in independent reading outside school hours (Anggraini et al., 2025), further constraining limited instructional time. Within this policy environment, Kurikulum Merdeka explicitly encourages the integration of technology and innovative pedagogy as responses to the literacy crisis (Kemendikbudristek, 2022; Dirgantoro et al, 2023). Educational technology has consequently received expanding policy and research attention, with national and provincial actors piloting digital literacy materials in Indonesian primary schools (Gildore et al., 2025; Fitri et al., 2025).

Gamification denotes the integration of game design elements such as points, badges, leaderboards, narratives, and progress tracking into educational contexts that are not themselves games (Deterding et al., 2011; Hamari et al., 2014; Kapp, 2012). It has accumulated considerable empirical attention as a strategy for raising motivation, engagement, and achievement, although the meta-analytic record reveals pronounced heterogeneity rather than convergent agreement. Sailer and Homner (2020) synthesized 19 controlled studies on cognitive learning outcomes and reported a g of 0.49 (95% CI [0.30, 0.69]), with smaller effects on motivational outcomes (g = 0.36, k = 16) and behavioral outcomes (g = 0.25, k = 9). More recent and broader syntheses report pooled estimates ranging from approximately g = 0.50 to g = 0.82, depending on inclusion criteria and outcome construct (Bai et al., 2020; Zeng et al., 2024). The mechanism literature has progressed beyond mechanic-based definitions toward psychological constructs through which gamification may operate, including autonomy, competence, and relatedness within self-determination theory (Sailer et al., 2017), challenge skill balance within flow theory (Csikszentmihalyi & Csikszentmihalyi, 2018; Hamari & Koivisto, 2015), and verbal visual integration within the cognitive theory of multimedia learning (Mayer, 2024; Sweller et al., 2019). Within reading instruction specifically, gamified programs have been linked to gains in engagement, time on task, and comprehension via immediate feedback, adaptive difficulty, and social competition (Wang et al., 2023; Qiao et al., 2023; Mason & Rich et al., 2020). Nevertheless, Dichev and Dicheva (2017) observed that core questions about what works, for whom, and under what conditions remain unresolved.

Three empirical gaps motivate the present study the first concerns population and design. The K-12 systematic review by Dehghanzadeh et al. (2023) located 54 empirical studies of gamification across 907 screened records published between 2008 and 2021, while Sailer and Homner (2020) noted that the majority of synthesized studies were conducted in higher education settings. Within the available primary-grade evidence, LMIC trials are sparse

relative to those from high-income contexts, and rigorous cluster-randomized designs in foundational literacy are rarer still (Wang et al., 2024). The applicability of meta-analytic effect sizes to LMIC primary literacy, therefore, depends on a transferability assumption that has not been tested empirically at scale, a gap with direct policy implications for the more than 25 million students enrolled in Indonesia's primary system.

The second gap concerns trajectory evidence. Sailer and Homner (2020) cautioned that many gamification studies rely on short timeframes and lack longitudinal data, which leaves observed effects vulnerable to novelty inflation. Dichev and Dicheva (2017) reached a parallel conclusion, noting that current evidence remains compatible with novelty-driven engagement spikes that may not translate into sustained learning gains. Multi-wave designs with midline assessments are required to distinguish sustained improvement from initial-engagement artifacts. However, such designs remain uncommon in the gamification literature, especially in LMIC primary settings, where infrastructural and logistical constraints often limit designs to single-posttest formats (Coelho et al., 2025; Wang et al., 2024). In the absence of trajectory data, claims about the durability of gamification effects rest on weaker inferential ground than headline effect-size figures suggest.

The third gap concerns heterogeneity and mechanism, with direct equity implications. The Mindspark randomized trial in India found that adaptive personalization without overt gamification produced the largest learning gains among students who had lagged furthest behind (Muralidharan et al., 2019). This evidence raises an empirical question that the gamification literature has yet to resolve, namely, whether the active ingredient in successful technology-enhanced interventions is the gamification layer (points, badges, narrative, leveled progression) or the adaptive-personalization layer that often accompanies it. Theoretical accounts predict that struggling learners should benefit most from gamification through extrinsic-motivational scaffolding (Liu et al., 2018) and from teaching at the right level, operationalized through adaptive difficulty (Banerjee et al., 2017). However, empirical evidence on moderation by baseline proficiency in foundational reading remains sparse and inconsistent (Schiele et al., 2025). Addressing this gap requires designs that report subgroup effects against a transparent quartile structure with effect-size confidence intervals, and that frame attribution conservatively when an active digital comparator is absent.

These three gaps carry substantive consequences for both research and practice. For research, the scarcity of well-designed LMIC primary trials means that international meta-analyses draw on limited evidence from populations where the foundational-literacy crisis is most acute, leaving global syntheses structurally peripheral to the contexts that most require them. For practice, the absence of trajectory evidence and disaggregated subgroup analyses obliges policymakers advocating for the scaled deployment of gamification to rely on aggregate effect-size figures that may not translate from short-term, high-income contexts into long-term realities in LMIC primary care settings. Addressing the three gaps simultaneously requires an LMIC primary trial that combines cluster randomization, multi-wave assessment, subgroup reporting against a transparent quartile structure, and claims framed in proportion to its design constraints. The present study was designed against these requirements. Building on this rationale, the study contributes a cluster randomized evaluation of LiterasiKu, a structured gamified literacy package developed for Indonesian primary schools, and advances three contributions: (a) empirical evidence from an under-represented LMIC primary context that complements the higher-education and high-income emphasis of prior syntheses; (b) trajectory evidence across baseline, midline, and endline that disaggregates sustained learning from novelty-driven engagement; and (c) prespecified

moderation evidence by baseline proficiency quartile that speaks directly to equity. Three research questions guide the analysis. RQ1 (effectiveness) asks whether students in the LiterasiKu condition gain more than students in standard Kurikulum Merdeka instruction on three reading proficiency outcomes (oral reading fluency, reading comprehension, and phonemic awareness) across sixteen weeks. RQ2 (trajectory) asks how reading proficiency gains evolve across baseline, midline, and endline assessments, with particular attention to whether early gains persist or attenuate. RQ3 (moderation, exploratory) asks whether the intervention effect varies by baseline proficiency quartile, with the prespecified expectation that learners in the lowest quartile benefit most. Because the design lacks an active digital comparator, all attributions refer to the gamified literacy package as a whole rather than to gamification mechanics in isolation, a framing decision revisited in the Discussion.

METHOD

Research Design

The study employed a parallel, two-arm cluster randomized controlled trial, with primary schools serving as the unit of randomization to minimize the risk of between-condition contamination at the student level. The design, conduct, and reporting of the trial adhered to the 2010 extension of the Consolidated Standards of Reporting Trials (CONSORT) for cluster-randomized trials (Campbell et al., 2012). Figure 1 presents the CONSORT flow diagram, which summarizes eligibility assessment, randomization, follow-up, and analysis at both the cluster and individual levels.

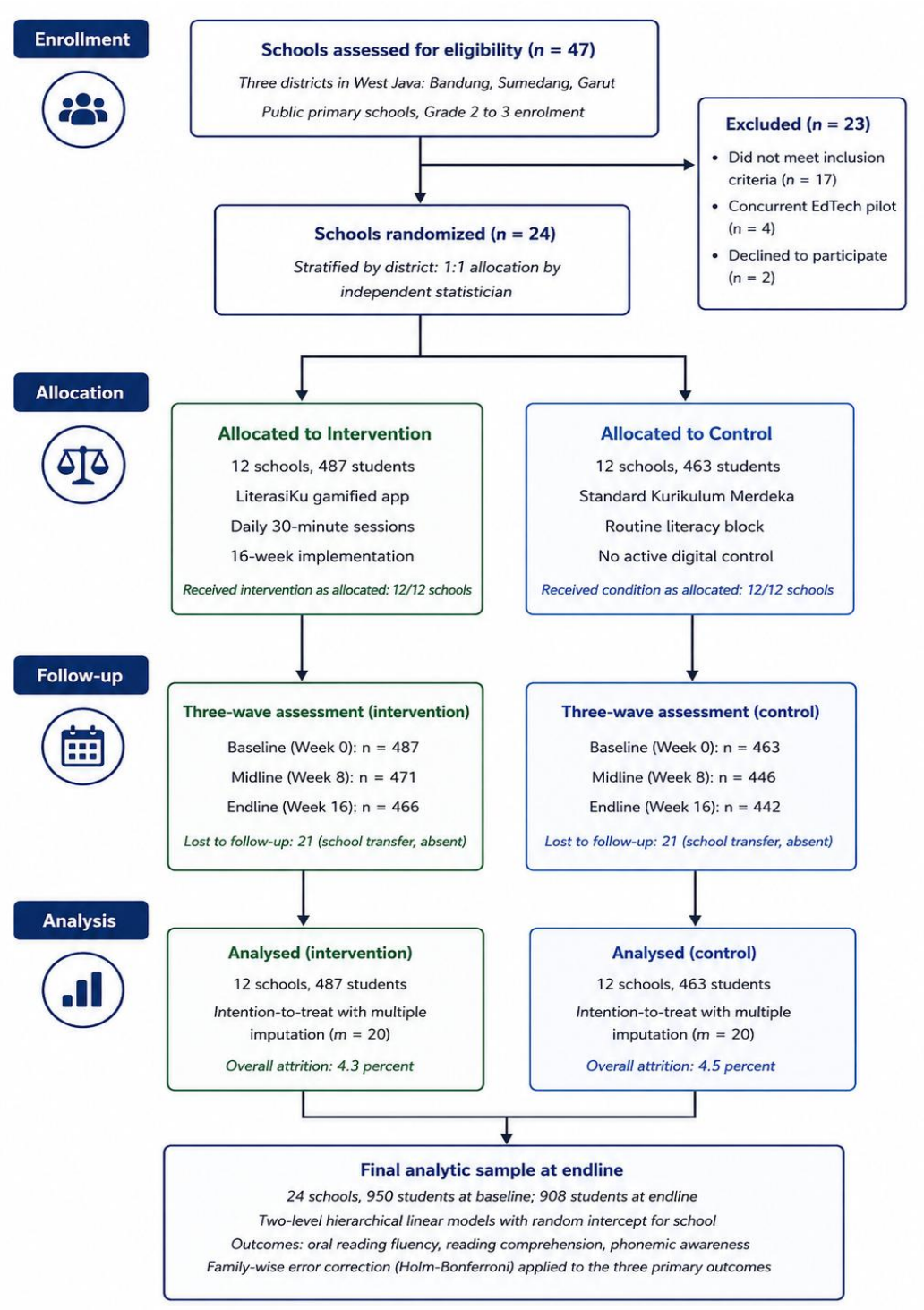

Figure 1. Consolidated Standards of Reporting Trials (CONSORT) 2010 flow diagram.

Participants and Setting

Schools were drawn from three districts in West Java Province: Bandung Regency, Sumedang Regency, and Garut Regency. Districts were selected based on below-average scores on the national Asesmen Nasional literacy component and on their classification as peri-urban or rural. Eligible schools were selected from a pool of 47 schools that met predefined inclusion criteria. Schools were stratified by district and randomly assigned in a 1:1 ratio to intervention or control using a computer-generated random sequence by an

independent statistician. School principals were blinded to allocation until informed consent had been obtained from teachers and from parents or guardians of participating students. Table 1 summarizes participant and setting characteristics, and Table 2 presents the inclusion criteria and eligibility outcomes.

Table 1. Participant and Setting Characteristics at Baseline

Characteristic	Intervention (12 schools)	Control (12 schools)	Total
Number of students	487	463	950
Mean age in years (SD)	8.2 (0.74)	8.2 (0.72)	8.2 (0.73)
Female (percent)	51.5%	51.0%	51.3%
Household income below Rp 3,000,000 per month	63.0%	60.9%	62.0%
Grade 2 (percent)	52.0%	51.4%	51.7%
Grade 3 (percent)	48.0%	48.6%	48.3%
Bottom-quartile baseline proficiency (within sample)	24.0%	25.4%	24.7%

Table 2. School Inclusion Criteria and Eligibility Outcomes

Criterion	Description	Schools Excluded
Enrolment threshold	Minimum of 30 students in Grades 2 and 3	12 schools
Digital infrastructure	Access to electricity and at least five functioning tablets	5 schools
No competing pilot	No concurrent participation in other educational technology pilots	4 schools
Voluntary participation	School consent to be randomized	2 schools
Total assessed	Schools assessed for eligibility	47 schools
Total randomized	Schools meeting all criteria and consenting	24 schools

An a priori power analysis was conducted under conservative assumptions about cluster size and intraclass correlation. Assuming an average cluster size of 40 students, an intraclass correlation coefficient of 0.14 for reading outcomes (Hedges & Hedberg, 2007, for primary reading), a design effect of approximately 6.5, and an alpha of .05 with two-sided testing, 24 schools provided greater than 80 percent power to detect a minimum detectable effect (MDE) of approximately d = 0.30 at endline, well within the range of recent meta-analytic estimates for educational technology interventions in LMICs (Evans & Popova, 2016; McEwan, 2015).

The LiterasiKu Application

LiterasiKu is a tablet-based, gamified literacy application developed collaboratively by the research team, primary school teachers, and a local educational technology developer. Following the Template for Intervention Description and Replication (TIDieR) reporting framework (Hoffmann et al., 2014), Table 3 documents the intervention's rationale, content, delivery, intensity, and tailoring alongside the control condition for direct comparison.

Table 3. Intervention and Control Conditions

Element	Intervention (LiterasiKu)	Control
Rationale	Adaptive difficulty drawing on teaching at the right level (TaRL); points and badges as extrinsic scaffolding for at-risk learners; culturally familiar narratives drawing on multimedia learning principles	Standard Kurikulum Merdeka instruction without an active digital component
Materials	Tablet-based gamified application with phonemic awareness, syllable construction, vocabulary, sentence reading, and passage comprehension modules	Existing Kurikulum Merdeka literacy textbooks and teacher-prepared activities
Delivery mode	Individual tablet sessions with teacher facilitation; offline functionality after initial download	Whole-class teacher-led literacy block at standard timetable
Gamification elements	Points, badges, leveled progression, narrative storytelling, adaptive difficulty (rule-based)	None
Cultural adaptation	Indonesian folklore and West Java daily-life contexts; story selection by the local curriculum review panel	Standard curriculum content
Dose and intensity	Daily 30-minute sessions for 16 weeks (planned 64 sessions, approximately 32 hours)	Routine literacy block at the same instructional time
Tailoring	Adaptive difficulty advances or pulls back learners based on item-specific accuracy thresholds	No within-program tailoring
Teacher preparation	Two-day training plus biweekly technical-support visits with on-site coaching	No additional preparation beyond regular teaching
Fidelity monitoring	Teacher daily logs plus biweekly observation of 30 percent of sessions	Biweekly observation visits documenting time, materials, and engagement structure

Measures and Procedures

Reading proficiency was measured using an adapted version of the Early Grade Reading Assessment (EGRA), validated for the Indonesian context (Dubect, 2015). Three subtasks were administered at each wave, summarized in Table 4. The Indonesian adaptation had previously been piloted with 120 Grade 2 students in two non-study districts, and items that showed low discrimination or differential functioning across regional dialects were revised before field testing. All assessments were administered individually by 40 trained enumerators (10 per district plus reserves), blinded to group assignment in quiet school settings during regular school hours. Enumerator training comprised 32 hours of content and procedural instruction, double-coded scoring practice, and a certification cutoff of Cohen's kappa greater than .85 against expert reference scoring on five practice cases. Application engagement was automatically captured from usage logs, including session counts, durations, badges earned, and level progression.

Table 4. Outcome Measures and Reliability Indicators

Subtask	Operationalization	Items	Reliability
Oral Reading Fluency (ORF)	Correct words per minute on the Grade 2 level passage	Single timed passage	Test-retest r = .91
Reading Comprehension (RC)	Multiple-choice questions on two short passages	10 items	Cronbach's alpha = .84
Phonemic Awareness (PA)	Initial-sound identification, blending, segmentation	20 items	Cronbach's alpha = .87

Data Analysis

Given the nested data structure of students within schools, two-level hierarchical linear models were estimated using restricted maximum likelihood in R (lme4 package). The primary model specified post-intervention scores as the outcome, with fixed effects for treatment condition, baseline score (grand-mean centered), and their interaction, and a random intercept for school. Effect sizes were computed as standardized mean differences (Cohen's d) using the pooled within-school standard deviation. Ninety-five percent confidence intervals for treatment effects were obtained from the model-based standard errors and reported alongside point estimates and Holm-Bonferroni-corrected p-values across the three primary outcomes. Growth curve models with three time points were used to examine learning trajectories for the second research question, with linear and time-by-treatment interaction terms. Moderation by baseline proficiency quartile was tested through interaction terms following the registered protocol, and we treat subgroup effect sizes as exploratory pending replication.

An intention-to-treat framework was applied throughout, with all randomized students included in the primary analysis according to the school to which they were originally allocated. Sensitivity analyses used treatment-on-the-treated estimates that adjusted for compliance, defined as completion of at least 80 percent of the prescribed sessions. Multiple imputation (m = 20) was used to address missing data under a missing-at-random assumption, with overall attrition rates of 4.3 percent in the intervention arm and 4.5 percent in the control arm. The imputation model included baseline outcome scores, demographic variables, and treatment indicators. Cluster-level baseline equivalence was assessed using standardized mean differences following the CONSORT cluster guidance, with values below 0.10 treated as evidence of adequate balance, and t-tests at the individual level used as a secondary check rather than as the primary equivalence criterion.

RESULTS

This section presents findings on baseline equivalence, primary intervention effects, learning trajectories, moderation analyses, and engagement indicators, organized according to the three research questions stated in the Introduction. Cluster-level standardized mean differences for the three baseline reading-proficiency measures were all below 0.10, indicating adequate balance after randomization, and individual-level t-tests confirmed no statistically significant differences between intervention and control groups at baseline on any outcome measure (all p greater than .20) or demographic characteristic (all p greater than .15). The intraclass correlation coefficient (ICC) was .14 for baseline ORF, .12 for baseline RC, and .11 for baseline PA, indicating meaningful school-level clustering and justifying the

multilevel analytical approach. Table 5 presents descriptive statistics across all three time points.

Table 5. Descriptive Statistics for Reading Proficiency Outcomes by Group and Time Point (N = 950)

Outcome and Group	Baseline M (SD)	Midline M (SD)	Endline M (SD)	ICC
Oral Reading Fluency, Intervention	21.4 (11.8)	30.2 (12.4)	35.7 (13.1)	.14
Oral Reading Fluency, Control	20.8 (11.3)	24.1 (11.9)	27.6 (12.2)	
Reading Comprehension, Intervention	6.3 (3.7)	9.1 (3.9)	11.8 (4.2)	.12
Reading Comprehension, Control	6.1 (3.5)	7.4 (3.6)	8.6 (3.8)	
Phonemic Awareness, Intervention	7.2 (4.1)	10.5 (4.3)	12.4 (4.5)	.11
Phonemic Awareness, Control	7.0 (3.9)	8.4 (4.0)	9.7 (4.1)	

Table 6 presents the hierarchical linear model results for each outcome at endline, controlling for baseline scores and including school-level random intercepts. The intervention produced statistically significant effects on all three reading proficiency outcomes, as assessed by the three primary tests, after Holm-Bonferroni correction. The largest effect was observed for ORF (beta = 6.83 cwpm, 95 percent CI [4.05, 9.61], d = 0.58), followed by PA (beta = 3.94, 95 percent CI [2.16, 5.72], d = 0.52) and RC (beta = 4.21, 95 percent CI [2.09, 6.33], d = 0.47). All effect sizes fall within the medium range according to conventional benchmarks (Cohen, 2013) and exceed the average effect size reported in meta-analyses of educational technology interventions in LMICs (d = 0.15 to 0.30; McEwan, 2015), although they are within the range of well-implemented structured pedagogy and adaptive personalization programs in similar settings (Banerjee et al., 2017; Muralidharan et al., 2019).

Table 6. Hierarchical Linear Model Results: Intervention Effects on Reading Proficiency at Endline

Parameter	ORF (cwpm)	RC (raw)	PA (raw)
Intercept (control mean at baseline-centered = 0)	27.40 (1.04)	8.55 (0.34)	9.62 (0.39)
Baseline coefficient (grand-mean centered)	0.74 (0.04)***	0.61 (0.04)***	0.66 (0.04)***
Treatment effect (beta)	6.83 (1.42)***	4.21 (1.08)***	3.94 (0.91)***
95 percent CI for treatment effect	[4.05, 9.61]	[2.09, 6.33]	[2.16, 5.72]
Holm-Bonferroni adjusted p-value	< .001	< .001	< .001
Cohen’s d (pooled within-school SD)	0.58	0.47	0.52
Between-school variance (random intercept)	5.82	0.51	0.84
Within-school residual variance	88.41	12.94	15.32
ICC at endline	.06	.04	.05

Growth curve analyses revealed that the intervention group exhibited significantly steeper linear growth trajectories than the control group across all three outcomes. Critically, gains did not attenuate between midline and endline. The rate of improvement in ORF in the intervention group accelerated slightly during the second half of the intervention (Weeks 9 to 16). The time-by-treatment interaction was significant for ORF (beta = 0.47 per week, SE = 0.12, p < .001) and for PA (beta = 0.34 per week, SE = 0.09, p < .001), indicating that the gap between groups widened progressively. For RC, the trajectory was more linear, with consistent gains across both halves of the intervention (beta = 0.28 per week, SE = 0.08, p less than .001). These patterns are consistent with sustained engagement rather than purely

novelty-driven gains, paralleling recent findings from a gamification trial (Coelho et al., 2025). However, novelty effects cannot be ruled out solely based on trajectory shape, and we return to this interpretive caution in the Discussion.

The interaction between treatment condition and baseline proficiency quartile was statistically significant for all three outcomes (all p less than .01) in this exploratory subgroup analysis. Table 7 presents the intervention effects disaggregated by baseline quartile. Students in the lowest proficiency quartile (Q1) exhibited the largest intervention effects across all outcomes, with effect sizes ranging from d = 0.63 to d = 0.74. The gradient was monotonically decreasing, with the highest performing quartile (Q4) showing the smallest, though still statistically significant, gains. We report these subgroup effects with appropriate interpretive caution because subgroup analyses are more vulnerable than primary analyses to inflated effect-size estimates and to alternative explanations, as discussed in the Discussion.

Table 7. Intervention Effects (Cohen's d with 95 percent CI) by Baseline Proficiency Quartile

Baseline Quartile	n	ORF d [95% CI]	RC d [95% CI]	PA d [95% CI]
Q1 (lowest)	235	0.74 [0.50, 0.98]	0.65 [0.40, 0.89]	0.71 [0.46, 0.95]
Q2	237	0.61 [0.37, 0.85]	0.55 [0.30, 0.79]	0.58 [0.34, 0.83]
Q3	239	0.48 [0.24, 0.72]	0.42 [0.18, 0.66]	0.43 [0.18, 0.67]
Q4 (highest)	239	0.31 [0.07, 0.55]	0.26 [0.02, 0.50]	0.32 [0.07, 0.56]

Application usage logs indicated a mean of 58.4 sessions per student (SD = 12.7) over the 16-week intervention period, against a target of 64 sessions (91.3 percent of planned dose). Students earned a mean of 34.2 badges (SD = 9.8) and progressed through an average of 7.3 of 10 levels. Within-group regression analyses revealed a positive association between total usage time and endline ORF (beta = 0.18, p less than .01), suggesting a dose-response relationship. We treat this association as correlational rather than as causal evidence of the active ingredient, because students who chose to use the application more may differ on unobserved characteristics from those who used it less. We did not implement complier average causal effect or instrumental-variable estimation.

Implementation fidelity, assessed through classroom observations and teacher logs, averaged 87.3 percent across schools, with no school falling below the a priori 75 percent threshold established in the registered protocol. The fidelity measure aggregated four dimensions specified in advance, namely adherence to the planned dose, quality of teacher facilitation, completeness of the activity sequence, and learner responsiveness during sessions. Adherence to the planned dose averaged 91.3 percent of scheduled sessions; quality of teacher facilitation averaged 86.4 percent of the observation rubric criteria met; completeness averaged 88.9 percent; and learner responsiveness averaged 82.7 percent. The lowest of these four dimensions provides a useful caution against treating the aggregate figure as a single quality index, because acceptable mean fidelity can mask uneven implementation across the constituent dimensions.

DISCUSSION

Main Findings

The trial provides cluster-randomized evidence that a structured, culturally adapted gamified literacy package can produce meaningful improvements in foundational reading skills among Indonesian primary learners in peri-urban West Java. The medium effect sizes observed across the three primary outcomes (d = 0.47 for RC, d = 0.52 for PA, and d = 0.58 for

ORF) exceed the typical effect sizes of educational technology interventions in LMICs reported in earlier meta-analyses (Evans & Popova, 2016; McEwan, 2015), and they approximate the effect sizes of intensive structured pedagogy programs that require substantially greater resource investment per student (Piper et al., 2018). They were achieved within the existing school timetable and infrastructure, suggesting feasibility for broader implementation in comparable settings. The effect-size pattern is also notable in international comparison. The educational technology effect sizes reported by McEwan (2015) range from d = 0.07 for some instructional technology programs to d = 0.30 for the more rigorous trials, while structured pedagogy programs in similar LMIC contexts have produced effect sizes around d = 0.40 to d = 0.60 with substantially greater resource investment per student (Piper et al., 2018). The medium effects observed in the present trial, therefore, sit at the upper end of the educational-technology distribution and approach the structured-pedagogy band, supporting the case that well-implemented digital programs can deliver value comparable to more resource-intensive alternatives in LMIC primary contexts.

The trajectory analysis, showing sustained and modestly accelerating gains over the 16-week intervention period, addresses one critical concern in the gamification literature: that many studies lack longitudinal data and may overestimate effects due to novelty (Sailer & Homner, 2020). The three-wave design demonstrates that gains did not attenuate after the initial weeks. However, trajectory shape alone does not rule out novelty effects, because slow novelty fade rather than rapid fade remains an alternative explanation, and within-app session duration trends would be needed for stronger interpretation. The finding that ORF showed the largest effect is pedagogically meaningful because fluency serves as a bridge between decoding and comprehension, and improvements in fluency have been associated with broader gains in reading proficiency over time (Hasbrouck & Tindal, 2017; Kim et al., 2021). The slightly smaller effect on RC is consistent with the theoretical expectation that complex skills take longer to develop and may benefit from extended intervention periods.

Novelty and Positioning Within the International Literature

The novelty of the present study lies in three connected analytical moves rather than in any single first claim. First, the study contributes to the LMIC primary literacy literature with a cluster randomized design and a three-wave assessment structure, both of which are uncommon in earlier Indonesian gamification work that has more often used quasi-experimental or single-group designs (Gildore et al., 2025; Fitri et al., 2025) and parallel European primary studies of gamification in reading (Ali et al., 2023). The cluster-randomized design provides defensible between-school causal inference, and the three-wave structure enables trajectory analyses that the single-posttest designs of earlier work do not support. The contribution is therefore to extend rigorous trial methods into a content domain (foundational reading in Indonesian primary education) where they have been comparatively rare.

Second, the study foregrounds heterogeneity of treatment effect rather than treating an average effect as the only finding. The exploratory subgroup analysis by baseline quartile generates evidence relevant to current debates about teaching at the right level (Banerjee et al., 2017) and about adaptive personalization as a candidate active ingredient in technology-enhanced learning (Muralidharan et al., 2019). The observed pattern is consistent with a benefit-where-needed-most interpretation, and we treat it as exploratory because subgroup analyses are vulnerable to several alternative explanations, as addressed in the next sub-section. Third, the study connects an Indonesian primary trial to the international gamification

literature in a way that engages with the documented heterogeneity of gamification effects (Dichev & Dicheva, 2017; Sailer & Homner, 2020) rather than selecting a single effect-size point estimate. We frame our results within this heterogeneity rather than as confirmation of any single prior estimate.

A second strand of novelty concerns framing. Because the trial does not include an active digital comparator, attributions are properly made to the LiterasiKu package as a whole rather than to gamification mechanics in isolation. This framing decision matters substantively because the Mindspark trial in India (Muralidharan et al., 2019) demonstrated that adaptive personalization on its own can produce strong gains in LMIC settings, raising the empirical question of whether the active ingredient in LiterasiKu is the gamification layer, the adaptive-difficulty layer, the cultural-relevance layer, or some combination. Our design cannot adjudicate among these candidates, so we present the gamified package as the unit of attribution rather than the mechanics of gamification in the abstract. Reframing the contribution in this way matters for both research and policy. For research, packaging-level claims travel less far than mechanism-level claims, and papers that overclaim mechanisms invite replication failures that damage the broader field. For policy, accurate attribution shapes which intervention components a Ministry of Education would scale and which it would treat as discretionary. The Mindspark literature has shown that adaptive personalization without overt gamification can produce strong effects (Muralidharan et al., 2019), and conservative attribution to the package as a whole leaves space for subsequent dismantling work to identify the cost-effective subset of components for scaled deployment.

Implication and Alternative Explanations for the Moderation Pattern

The moderation analysis, revealing larger effects for the lowest-performing quartile (d = 0.74 for ORF), carries equity-oriented implications. However, those implications must be weighed against four alternative explanations that our design cannot fully discriminate between. The first is regression to the mean. Students assigned to the lowest baseline quartile, by definition, include those whose true scores are higher than their baseline scores due to measurement error, and these students would show systematic gains even in the absence of an intervention (Barnett et al., 2005). The monotonically decreasing gradient across quartiles is consistent with this statistical artifact and cannot be ruled out without designs that explicitly control for it.

The second is floor effects in the EGRA assessment. Students at the very bottom of the baseline distribution may have more headroom to grow than students near the top, who may approach an asymptote on age-appropriate items. The third is differential trajectories in the comparison condition. Matthew effects, in which initially low performers fall further behind without intervention (Stanovich, 1986), would produce a between-group gap that is largest for low performers, even if the intervention worked equally well for all subgroups. The fourth is the operationalization of struggling as the within-sample bottom quartile rather than as a cutoff calibrated against Indonesian Grade 2 oral-reading fluency norms, an operational choice that limits external comparability and that we acknowledge openly.

Several practical implications nonetheless emerge for Indonesian education policy and for LMIC primary education more broadly. The Kurikulum Merdeka reform (Kemendikbudristek, 2022) explicitly encourages technology integration, and our results suggest that structured gamified applications can be effectively integrated into existing literacy instruction without additional instructional hours. The offline functionality of LiterasiKu addresses connectivity constraints documented as a primary barrier to educational

technology adoption in Indonesian rural schools (Acasus, 2024), and the 87.3 percent implementation fidelity rate, achieved with a modest two-day training and biweekly support, suggests manageable capacity requirements. For LMIC policymakers more broadly, the disproportionate benefits for struggling learners in our exploratory subgroup analysis suggest that gamified literacy applications could serve as an equity-enhancing intervention. However, that suggestion must be replicated under designs that can rule out the alternative explanations described above before it warrants policy action. Cost-effectiveness considerations matter substantively at the scaling stage. Tablet hardware, content licensing, biweekly technical support visits, and teacher training expenses constitute non-trivial recurring costs that have not been quantified in the present analysis. Independent of effect size, an intervention that costs USD 50 per student per year carries different policy implications than one that costs USD 5, and economic evaluation alongside effectiveness evaluation is essential before scaling decisions are taken.

Several limitations of this study should be acknowledged transparently and translated into bounded claims. The 16-week intervention period, although longer than most gamification studies, is insufficient to assess long-term retention and transfer of literacy gains, and follow-up assessments at 6 and 12 months are needed to determine durability. The study was confined to three districts in West Java, and generalizability to other Indonesian provinces with different linguistic contexts (such as Papua, where local languages differ substantially from Bahasa Indonesia) cannot be assumed without replication. Blinding was not feasible, and Hawthorne effects cannot be ruled out, although the sustained trajectory of gains argues against a purely novelty-driven explanation. The most consequential design limitation is the absence of an active digital control condition, which precludes attribution of effects to gamification mechanics rather than to digital delivery, adaptive personalization, or cultural relevance. The intervention's cost-effectiveness was not formally assessed, and an economic evaluation will be needed to inform scaling decisions.

Two further reporting limitations warrant acknowledgement. First, we did not analyze application engagement logs as mediators of treatment effects on reading outcomes. Although session counts, badges earned, and level progression data were collected automatically and are described in the Results, formal mediation analyses linking these engagement variables to reading-proficiency outcomes were not undertaken in the present analysis and constitute a high-priority direction for future work. Second, although the registered protocol specified treatment-by-quartile moderation, additional moderation analyses by gender, household socioeconomic status, grade level, and district were not preregistered and were not undertaken in the present manuscript. Reporting heterogeneous effects across these dimensions is essential for any equity-related claims, and we treat their absence here as a limitation rather than as a feature.

Contribution to Educational Technology Research

Figure 2 synthesizes the contribution of the present study across four research and policy stages, from what was tested through to what remains for subsequent investigation. The framework identifies three connected contributions, namely methodological, theoretical, and policy contributions, each of which carries part of the contribution and each of which is bounded by specific design constraints. The methodological contribution centers on a cluster randomized design with three-wave assessment, CONSORT cluster reporting, family-wise error correction, and subgroup confidence intervals, addressing aspects of design rigor that earlier Indonesian gamification work has often left underdeveloped. This contribution is

bounded by the absence of an active comparator and by the within-sample operational definition of struggling learners.

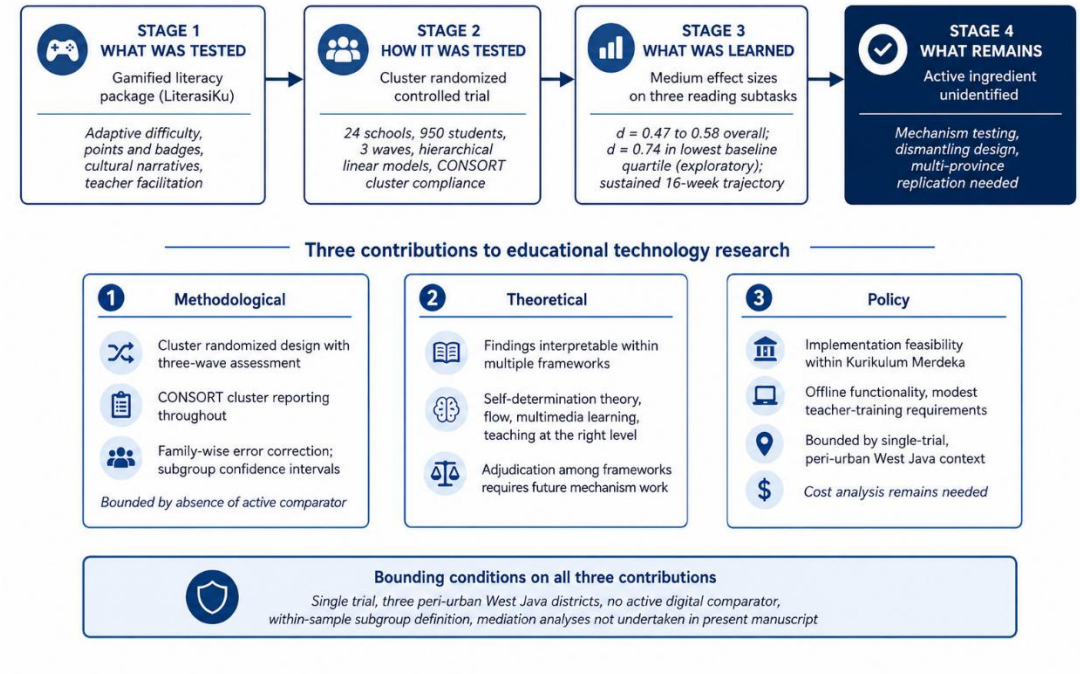

Figure 2. Pathway from trial evidence to scaled educational technology practice.

The theoretical contribution articulates how the present findings can be interpreted within several frameworks, including self-determination theory (Sailer et al., 2017), flow theory (Hamari & Koivisto, 2015), the cognitive theory of multimedia learning (Mayer, 2024), and the teaching at the right level tradition (Banerjee et al., 2017). Each framework offers a coherent reading of the results, but the present design cannot adjudicate among them. We therefore frame the theoretical contribution as interpretive scaffolding for future mechanism work rather than as confirmation of any single framework. Adjudicating among these frameworks will require dismantling designs that manipulate specific theoretical components and measure their effects on hypothesized mechanisms, such as autonomy satisfaction, perceived challenge-skill balance, or cognitive load.

The policy contribution centers on evidence of implementation feasibility for Indonesia's Kurikulum Merdeka reform and comparable LMIC settings. The combination of offline functionality, modest teacher-training requirements, and integration into the existing literacy block supports the case for further development of structured digital literacy materials in this context. However, the policy contribution is bounded by the cost-effectiveness gap noted in the Limitations and by the inferential leap from a single trial in three peri-urban West Java districts to scaled deployment across Indonesia's much larger primary-school network. Articulating the conditions under which the intervention would be expected to scale (stable district capacity, teacher digital readiness, device availability, content review by language and curriculum specialists) is essential before the present evidence base can support any scaling recommendation. Each of these conditions is more variable across Indonesian provinces than the within-study profile of West Java would suggest, and the policy contribution of the present trial is therefore best read as evidence about feasibility under specific district-level conditions rather than as a general endorsement of gamified literacy packages for the Indonesian primary system.

CONCLUSION

This cluster randomized controlled trial provides rigorous evidence that a structured, culturally adapted gamified literacy package can improve foundational reading proficiency among Indonesian primary learners when delivered within existing instructional time. The medium effect sizes observed across oral reading fluency, phonemic awareness, and reading comprehension, combined with sustained gains across 16 weeks, demonstrate that well-designed digital interventions can move beyond novelty-driven engagement to produce durable learning improvements. Critically, the disproportionate benefits for lowest-quartile learners suggest that gamified applications may serve as equity-enhancing tools in contexts where learning poverty is most concentrated. However, replication under designs that rule out regression-to-the-mean and floor effects remains essential.

The study's most consequential contribution, however, is methodological rather than substantive. By adhering to CONSORT cluster standards, implementing three-wave assessment, reporting subgroup effects with confidence intervals, and engaging transparently with alternative explanations, we demonstrate that rigorous trial methods are feasible in LMIC primary settings despite infrastructural constraints. This methodological transparency matters because the credibility of educational technology research depends as much on disclosing inferential limits as on reporting effect-size magnitudes. The Indonesian primary system serves over 25 million students, and the field advances most productively when studies combine internal validity with honest articulation of what their designs cannot claim.

Future progress requires dismantling trials that isolate gamification mechanics from adaptive personalization, conducting mediation analyses to test whether engagement metrics drive outcomes, and conducting cost-effectiveness evaluations to inform scaling decisions. Until such evidence accumulates, the present findings support cautious optimism: gamified literacy packages can work in under-resourced LMIC classrooms, but attributing effects to specific components and ensuring equitable impact demand continued methodological rigor and contextual attentiveness.

REFERENCES

Acasus. (2024). *Challenges facing EdTech adoption in developing countries*. Acasus Impact Report. https://www.acasus.com/impact/article/challenges-facing-edtech-adoption-in-developing-countries

Ali, M. T., Lykknes, A., & Tiruneh, D. T. (2023). Examining the effects of supervised laboratory instruction on students' motivation and their understanding of chemistry. *Education Sciences*, *13*(8), 798. https://doi.org/10.3390/educsci13080798

Anggraini, D. A., Ekawati, R., Arifin, S., Kuswandi, D., & Ramli, M. (2025). Declining Interest in Reading in Elementary School Students: An Analysis of Inhibiting Factors and Their Impact on Learning Achievement. *Eduvest-Journal of Universal Studies*, *5*(6), 6445-6463. https://doi.org/10.59188/eduvest.v5i6.50197

Antoninis, M., Alcott, B., Al Hadheri, S., April, D., Fouad Barakat, B., Barrios Rivera, M., ... & Weill, E. (2023). Global Education Monitoring Report 2023: Technology in education: A tool on whose terms?.

Azevedo, J. P., Hasan, A., Goldemberg, D., Geven, K., & Iqbal, S. A. (2021). Simulating the potential impacts of COVID-19 school closures on schooling and learning outcomes: A

set of global estimates. *The World Bank Research Observer*, *36*(1), 1-40. https://doi.org/10.1093/wbro/lkab003

Bai, S., Hew, K. F., & Huang, B. (2020). Does gamification improve student learning outcome? Evidence from a meta-analysis and synthesis of qualitative data in educational contexts. *Educational Research Review, 30*, Article 100322. https://doi.org/10.1016/j.edurev.2020.100322

Banerjee, A. V., Banerji, R., Berry, J., Duflo, E., Kannan, H., Mukerji, S., Shotland, M., & Walton, M. (2017). From proof of concept to scalable policies: Challenges and solutions, with an application. *Journal of Economic Perspectives, 31*(4), 73–102. https://doi.org/10.1257/jep.31.4.73

Barnett, A. G., van der Pols, J. C., & Dobson, A. J. (2005). Regression to the mean: What it is and how to deal with it. *International Journal of Epidemiology, 34*(1), 215–220. https://doi.org/10.1093/ije/dyh299

Beatty, A., Berkhout, E., Bima, L., Pradhan, M., & Suryadarma, D. (2021). Schooling progress, learning reversal: Indonesia's learning profiles between 2000 and 2014. *International Journal of Educational Development, 85*, Article 102436. https://doi.org/10.1016/j.ijedudev.2021.102436

Campbell, M. K., Piaggio, G., Elbourne, D. R., & Altman, D. G. (2012). Consort 2010 statement: Extension to cluster randomised trials. *BMJ, 345*, e5661. https://doi.org/10.1136/bmj.e5661

Coelho, F., Rando, B., Aparício, D., Pontífice-Sousa, P., Gonçalves, D., & Abreu, A. M. (2025). The impact of educational gamification on cognition, emotions, and motivation: a randomized controlled trial. *Journal of Computers in Education*, 1-48. https://doi.org/10.1007/s40692-025-00366-x

Cohen, J. (2013). *Statistical power analysis for the behavioral sciences*. routledge.

Csikszentmihalyi, M., & Csikszentmihalyi, I. S. (Eds.). (1992). *Optimal experience: Psychological studies of flow in consciousness*. Cambridge university press. https://doi.org/10.1017/CBO9780511621956

Dehghanzadeh, H., Farrokhnia, M., Dehghanzadeh, H., Taghipour, K., & Noroozi, O. (2023). Using gamification to support learning in K-12 education: A systematic literature review. *British Journal of Educational Technology, 55*(1), 34–70. https://doi.org/10.1111/bjet.13335

Deterding, S., Dixon, D., Khaled, R., & Nacke, L. (2011). From game design elements to gamefulness: Defining gamification. In *Proceedings of the 15th International Academic MindTrek Conference* (pp. 9–15). ACM. https://doi.org/10.1145/2181037.2181040

Dichev, C., & Dicheva, D. (2017). Gamifying education: What is known, what is believed and what remains uncertain. *International Journal of Educational Technology in Higher Education, 14*, Article 9. https://doi.org/10.1186/s41239-017-0042-5

Dirgantoro, K. P. S., & Soesanto, R. H. (2023). Towards a Paradigm Shift: Analysis of Student Teachers' and Teacher Education Institutions' Readiness on Kurikulum Merdeka. *Jurnal Pendidikan Dan Kebudayaan*, *8*(2), 185-201. https://doi.org/10.24832/jpnk.v8i2.4271

Dubeck, M. M., & Gove, A. (2015). The early grade reading assessment (EGRA): Its theoretical foundation, purpose, and limitations. *International Journal of Educational Development*, *40*, 315-322. https://doi.org/10.1016/j.ijedudev.2014.11.004

Evans, D. K., & Popova, A. (2016). What really works to improve learning in developing countries? An analysis of divergent findings in systematic reviews. *The World Bank Research Observer, 31*(2), 242–270. https://doi.org/10.1093/wbro/lkw004

Fitri, S. (2025). The (Technology) Implementation Gap Among Indonesian Teachers: Understanding the Disconnect Between Personal and Professional Technology Use. https://cerj.educ.cam.ac.uk/archive/v12_2025/CERJ_Vol12_211-225.pdf

Gildore, P. J. E., Aryanto, S., Suharjuddin, Denatara, E. T., & Awiria. (2025). Effective reading intervention strategies for primary grade students in Indonesia: a systematic review. *Cogent Education*, *12*(1), 2482470. https://doi.org/10.1080/2331186X.2025.2482470

Hamari, J., & Koivisto, J. (2015). Why do people use gamification services?. *International journal of information management*, *35*(4), 419-431. https://doi.org/10.1016/j.ijinfomgt.2015.04.006

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

Hasbrouck, J., & Tindal, G. (2017). An Update to Compiled ORF Norms. Technical Report# 1702. *Behavioral Research and Teaching*. https://eric.ed.gov/?id=ED605146

Hedges, L. V., & Hedberg, E. C. (2007). Intraclass correlation values for planning group-randomized trials in education. *Educational Evaluation and Policy Analysis, 29*(1), 60–87. https://doi.org/10.3102/0162373707299706

Hoffmann, T. C., Glasziou, P. P., Boutron, I., Milne, R., Perera, R., Moher, D., ... & Michie, S. (2014). Better reporting of interventions: template for intervention description and replication (TIDieR) checklist and guide. *Bmj*, *348*. https://doi.org/10.1136/bmj.g1687

Hunaepi, H., & Suharta, I. (2024). Transforming education in Indonesia: The impact and challenges of the Merdeka belajar curriculum. *Path of Science*, *10*(6), 5026-5039.10.22178/pos.105-31

Kapp, K. M. (2012). *The gamification of learning and instruction: game-based methods and strategies for training and education*. John Wiley & Sons.

Kemendikbudristek. (2022). *Kurikulum Merdeka: Concept, principles, and implementation guidelines*. Ministry of Education, Culture, Research, and Technology of the Republic of Indonesia.

Kim, Y. S. G., Petscher, Y., Wanzek, J., & Al Otaiba, S. (2018). Relations between reading and writing: A longitudinal examination from grades 3 to 6. *Reading and writing*, *31*(7), 1591. https://link.springer.com/article/10.1007/s11145-018-9855-4

Liu, M., Huang, Y., & Zhang, D. (2018). Gamification's impact on manufacturing: Enhancing job motivation, satisfaction and operational performance with smartphone-based gamified job design. *Human Factors and Ergonomics in Manufacturing & Service Industries*, *28*(1), 38-51.

Mason, S. L., & Rich, P. J. (2020). Development and analysis of the elementary student coding attitudes survey. *Computers & Education*, *153*, 103898. https://doi.org/10.1016/j.compedu.2020.103898

Mayer, R. E. (2024). The past, present, and future of the cognitive theory of multimedia learning. *Educational Psychology Review, 36*(1), Article 8. https://doi.org/10.1007/s10648-023-09842-1

McEwan, P. J. (2015). Improving learning in primary schools of developing countries: A meta-analysis of randomized experiments. *Review of Educational Research, 85*(3), 353–394. https://doi.org/10.3102/0034654314553127

Mullis, I. V., Martin, M. O., Foy, P., & Drucker, K. T. (2012). *PIRLS 2011 international results in reading*. International Association for the Evaluation of Educational Achievement. Herengracht 487, Amsterdam, 1017 BT, The Netherlands.

Muralidharan, K., Singh, A., & Ganimian, A. J. (2019). Disrupting education? Experimental evidence on technology-aided instruction in India. *American Economic Review, 109*(4), 1426–1460. https://doi.org/10.1257/aer.20171112

OECD. (2023). *PISA 2022 results: What students know and can do (Volume I)*. OECD Publishing. https://doi.org/10.1787/53f23881-en

Piper, B., Zuilkowski, S. S., Dubeck, M., Jepkemei, E., & King, S. J. (2018). Identifying the essential ingredients to literacy and numeracy improvement: Teacher professional development and coaching, student textbooks, and structured teachers' guides. *World Development, 106*, 324–336. https://doi.org/10.1016/j.worlddev.2018.01.018

Qiao, S., Shen, S., & Huang, R. (2023). The effects of gamification on students' reading engagement: A systematic review. *Educational Technology Research and Development, 71*, 2079–2103. https://doi.org/10.1007/s11423-023-10225-0

Sailer, M., & Homner, L. (2020). The gamification of learning: A meta-analysis. *Educational Psychology Review, 32*, 77–112. https://doi.org/10.1007/s10648-019-09498-w

Sailer, M., Hense, J. U., Mayr, S. K., & Mandl, H. (2017). How gamification motivates: An experimental study of the effects of specific game design elements on psychological need satisfaction. *Computers in Human Behavior, 69*, 371–380. https://doi.org/10.1016/j.chb.2016.12.033

Schiele, T., Edelsbrunner, P., Mues, A., Birtwistle, E., Wirth, A., & Niklas, F. (2025). The effectiveness of game-based literacy app learning in preschool children from diverse backgrounds. *Learning and Individual Differences*, *117*, 102579. https://doi.org/10.1016/j.lindif.2024.102579

Stanovich, K. E. (2009). Matthew effects in reading: Some consequences of individual differences in the acquisition of literacy. *Journal of education*, *189*(1-2), 23-55. https://doi.org/10.1598/RRQ.21.4.1

Sweller, J., Van Merriënboer, J. J., & Paas, F. (2019). Cognitive architecture and instructional design: 20 years later. *Educational psychology review*, *31*(2), 261-292. https://doi.org/10.1007/s10648-019-09465-5

UNICEF Indonesia. (2025). *Country programme document: Indonesia 2026 to 2030* (E/ICEF/2025/P/L.10). United Nations Children's Fund. https://digitallibrary.un.org/record/4085540

Wang, Z., Harun, J., & Yuan, Y. (2024). Enhancing Reading Instruction Through Gamification: A Systematic Review of Theoretical Models, Implementation Strategies, and Measurable Outcomes (2020-2024). *Journal of Information Technology Education: Research*, *23*. https://doi.org/10.28945/5394

Wang, Z., Harun, J., & Yuan, Y. (2024). Enhancing reading instruction through gamification: A systematic review of theoretical models, implementation strategies, and measurable outcomes (2020 to 2024). *Journal of Information Technology Education: Research, 23*, Article 28. https://doi.org/10.28945/5394

World Bank. (2022). *The state of global learning poverty: 2022 update*. World Bank Group. https://www.worldbank.org/en/topic/education/publication/state-of-global-learning-poverty

Zeng, J., Parks, S., & Shang, J. (2024). Exploring the impact of gamification on students' academic performance: A comprehensive meta-analysis of studies from 2008 to 2023. *British Journal of Educational Technology, 55*(5), 1969–1994. https://doi.org/10.1111/bjet.13471.